Supplements of the Anuario de Filología Vasca "Julio de Urquijo", LIII

TOOLS FOR LINGUISTIC VARIATION

Gotzon Aurrekoetxea, Jose Luis Ormaetxea
(eds.)

eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea
Bilbao Bilbo
2010

ASJU-REN GEHIGARRIAK
ANEJOS DEL *ASJU*
SUPPLEMENTS OF ASJU

XI. LUIS MICHELENA-IBON SARASOLA, *Textos arcaicos vascos. Contribución al estudio y edición de textos antiguos vascos,* 1989. 12 €.

XII. HUGO SCHUCHARDT, *Introducción a las obras de Leizarraga. Sobre el modo de disponer la reimpresión, en particular sobre las erratas y variantes en el texto de Leizarraga,* 1989. 8 €.

XIII. MANUEL AGUD - ANTONIO TOVAR, *Diccionario etimológico vasco, I. A-Ardui,* 1989, 1993. Agotado.

XIV. JOSEBA A. LAKARRA (ed.), *Memoriae L. Mitxelena magistri sacrum,* 1991. 36 €.

XV. RICARDO GÓMEZ - JOSEBA A. LAKARRA (arg.), *Euskalaritzaren historiaz I: XVI-XIX. mendeak,* 1992. 18 €.

XVI. BEÑAT OYHARÇABAL, *La pastorale souletine: édition critique de "Charlemagne",* 1990. 18 €.

XVII. RICARDO GÓMEZ - JOSEBA A. LAKARRA (arg.), *Euskalaritzaren historiaz II: XIX-XX. mendeak.* Prestatzen.

XVIII. JOSEBA A. LAKARRA, *Harrieten Gramatikako hiztegiak (1741),* 1994. 10 €.

XIX. MANUEL AGUD - ANTONIO TOVAR, *Diccionario etimológico vasco, II. Ardun-Beuden,* 1990, 1993. Agotado.

XX. LUIS MICHELENA, *Lenguas y protolenguas,* 1990 (1963, 1986). 8 €.

XXI. ARENE GARAMENDI, *El teatro popular vasco. Semiótica de la representación,* 1991. 12 €.

XXII. LASZLÓ K. MARÁCZ, *Asymmetries in Hungarian,* 1991. 15 €.

XXIII. PETER BAKKER, GIDOR BILBAO, NICOLAAS G. H. DEEN, JOSÉ I. HUALDE, *Basque pidgins in Iceland and Canada,* 1991. 10 €.

XXIV. MANUEL AGUD - ANTONIO TOVAR, *Diccionario etimológico vasco, III. Beule-Egileor (Babarraso-Bazur),* 1991. Agotado.

XXV. JOSÉ M.ª SÁNCHEZ CARRIÓN, *Un futuro para nuestro pasado,* 1991. 15 €.

XXVI. MANUEL AGUD - ANTONIO TOVAR, *Diccionario etimológico vasco, IV. Egiluma-Galanga,* 1991. Agotado.

XXVII. JOSEBA A. LAKARRA - JON ORTIZ de URBINA (eds.), *Syntactic theory and Basque syntax,* 1992. 18 €.

XXVIII. RICARDO GÓMEZ - JOSEBA A. LAKARRA (arg.), *Euskal dialektologiako kongresua (Donostia, 1991ko irailaren 2-6),* 1994. 21 €.

XXIX. JOSÉ I. HUALDE - XABIER BILBAO, *A phonological study of the Basque dialect of Getxo,* 1992. 8 €.

XXX. MANUEL AGUD - ANTONIO TOVAR, *Diccionario etimológico vasco, V. Galani-Iloza,* 1992. 8 €.

TOOLS FOR LINGUISTIC VARIATION

Gotzon Aurrekoetxea, Jose Luis Ormaetxea
(eds.)

TOOLS FOR LINGUISTIC VARIATION

Gotzon Aurrekoetxea, Jose Luis Ormaetxea
(eds.)

eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

The publication of this book has been possible thanks to the finantial support of the Department of Education, Universities and Research of the Basque Government (ref. AE-2009-1-8).

ISBN: 978-84-9860-429-0
Depósito legal / Lege gordailua: BI - 2.378-2010

Fotocomposición / Fotokonposizioa: Ipar, S. Coop.
Zurbaran, 2-4 (48007 Bilbao)

Impresión / Inprimatzea: Itxaropena, S.A.
Araba kalea, 45 (20800 Zarautz-Gipuzkoa)

INDEX

INTRODUCTION

Once again we have to say that the use of technology is absolutely essential to make progress in dialectology. Dialectology needs technological tools to develop and to advance in the research of language variation as has already been demonstrated in different occasions, in different countries, and in different ways to analyse linguistic variation.

The preoccupation for the use of technology in Dialectology is not new. There are lots of citations about this subject in the scientific literature. We can quotate i.e. Kretzschmar (1999), amongst others.

When the aim is to speak about the future of Dialectology the use of technology is a common place.

We think that we already are in the future, because in nowadays Dialectology it is essential to take into account that the use of automated tools is a task which cannot be postponed. Most of them use it in their usual researches, but there are many others who think that in Dialectology the use of technology could be important in the future, but no now; that is to say, they do not use it.

We are absolutely convinced that we cannot postpone the use of technology; each project we postpone the use of it is one occasion we have lost. Time runs against us.

Teachers of Dialectology must speak in the class-rooms to university students (in undergraduate studies but also in postgraduate), that as in other kinds of researches in Dialectology also is crucial the use of tools that allow us to research more efficiently and quickly using automated computer programs.

This book gets together several European researchers, some of them well-known in the international field and other younger that manage diverse research projects or are taken part in research teams. The papers were shown in "Tools for Linguistic Variation" symposium, held at the Faculty of Letters of the UPV/EHU (The University of the Basque Country), in October 1 and 2, 2009.

All of them are joined by the conviction that the use of technology is crucial at this moment. And all of them use in their researches automated computer programs.

Hans Goebl (U. Salzburg) shows the maximum advances that he has done in the well-known VDM program, in the dialectometrical team of the University of Salzburg in his "Introducción a los problemas y métodos según los principios de la Escuela Dialectométrica de Salzburgo (con ejemplos sacados del 'Atlante Italo-Svizzero', AIS)" entitled paper. The contribution shows some aspects of the dialectometry: interpunctual dialectometry, dendrograme's dialectometry, and correlative dialectometry. The VDM program is one of the most widely used tools in Dialectology when it is necessary to use the statistics.

John Nerbonne, Jelena Prokić, Martijn Wieling and Charlotte Gooskens (U. Groningen) present "Some further dialectometrical steps": Levenshtein distance (one of the most successful methods to determine sequence distance), using the RuG/L04 package to visualize geographical patterns (world wide known program), Inducing segment distances empirically, Co-clustering to detect linguistic basis of dialect differentiation, Understanding Séguy's curve, etc.

The technology used in different geolinguistical researches has been shown by Roland Bauer, the team Guylaine Brun-Trigaud and Pierre-Aurélien Georges, and the team Gotzon Aurrekoetxea and Aitor Iglesias. Roland Bauer (U. Salzburg) deals with oral dialectology in his "Le projet VIVALDI: présentation d'un atlas linguistique parlant virtuel", which studies Italian dialects in an on-line environment. Guylaine Brun-Trigaud (CNRS) and Pierre-Aurélien Georges (Université Nice Sophia Antipolis) present the THESAURUS OCCITAN in two papers: "Le THESAURUS OCCITAN: une base de données multimedia consacrée aux dialectes occitans" and "the THESAURUS OCCITAN: a multimedia database dedicated to Occitan dialects. Presentation of its morphosyntax module", respectively; that is to say, the characteristics of data processing (the addition of new data, lemmatisation, syntactical tree tagging) and work features. Gotzon Aurrekoetxea (UPV/EHU) and Aitor Iglesias (UPV/EHU) shows the technological tools that EUDIA research team (UPV/EHU) take in their researches from gathering information for the EDAK corpus, in the entitled "Technology for prosodic variation" paper. Pilar García Mouton (CSIC, Madrid) deals with the project to launch on-line the ALPI atlas in her "El procesamiento informático de los materiales del *Atlas Lingüístico de la Península Ibérica* de Tomás Navarro Tomás": The paper takes into account the composition of the new team, the description of the project and its methodological aspects.

Esteve Clua (UPF, Barcelona) shows the need of the linguistic analyse of the data before their dialectometrical analyse in his "Relevancia del análisis lingüístico en el tratamiento cuantitativo de la variación dialectal".

With regard to syntactical aspects the book enrichs with two contributions: Ernestina Carrilho (U. Lisboa) shows the tools that the Centro de Linguística of the University of Lisboa use to analyse syntactic variation in the contribution entitled "Tools for dialect syntax: the case of CORDIAL-SIN (an annotated corpus of Portuguese dialects)". And Inés Fernández-Ordóñez (U. Complutense, Madrid) contributes showing the value of COSER corpus in her "New methods for the study of grammatical variation and the *Audible Corpus of Spoken Rural Spanish*".

Maria-Pilar Perea (U. Barcelona), using *La flexió verbal* of the Catalan gathered in the first middle of the 20th century, shows in "The application of speech synthesis and speech recognition techniques in dialectal studies" one aspect no studied in dialectology until nowadays.

Ekaitz Santazilia (UPV/EHU) discuss the use of the article in the Basque according to the Bourciez Corpus (1895) and using VDM program to show the geographical distribution in "Un retrato del artículo vasco en el año 1895 mediante el programa *VDM*".

Gotzon Aurrekoetxea, Jose Luis Ormaetxea

INTRODUCCIÓN A LOS PROBLEMAS Y MÉTODOS SEGÚN LOS PRINCIPIOS DE LA ESCUELA DIALECTOMÉTRICA DE SALZBURGO (CON EJEMPLOS SACADOS DEL "ATLANTE ITALO-SVIZZERO", AIS)

Hans Goebl
Universidad de Salzburgo

Abstract

This paper documents the many taxometric and cartographic achievements of the Salzburg school of dialectometry. The paper discusses the following topics: 1) problems of measurement ("taxatation") of linguistic atlas data (with particular consideration of Romance linguistic atlases), 2) establishment of the data matrix, 3) choice of the similarity index (here: Relative *and* Weighted Identity Value*), 4) generation of the respective similarity and distance matrices, 5) their subsequent cartographic exploitation, which encompasses the following cartographic tools: similarity maps, parameter maps, interpoint maps, dendrograms (and their spatial projection), and correlation maps. The ultimate purpose of these highly sophisticated cartographic tools (choropleth and isopleth maps) is to increase our knowledge of the complex mechanisms of the "dialectal management of space by man". From a methodological point of view our paper deals with problems related to Romance dialectology and linguistic geography, historical linguistics, numerical classification, statistics and statistical cartography.*

The examples are drawn from the Italian linguistic atlas AIS (Atlante-Italo-Svizzero, *recte:* Sprach- und Sachatlas Italiens und der Südschweiz *published by Karl Jaberg and Jakob Jud, Zofingen: Ringier, 1928-1940, 8 volumes) whose data have been dialectometrically analyzed between 2007 and 2009.*

The taxometric calculations and their respective visualizations are realized by a powerful computer program called "Visual DialectoMetry" (VDM), created by Edgar Haimerl between 1997 and 2000 in Salzburg, which is freely available for research purposes.

Key words: *dialectometry, geolinguistics, linguistic geography, dialectology, cartography, numerical classification.*

1. Prólogo

Es bien sabido que el AIS (= "Atlante italo-svizzero") es el segundo atlas lingüístico, por lo que hace a su importancia, de la filología románica. Fue elaborado por dos romanistas suizos —Karl Jaberg y Jakob Jud— de 1919 al 1927 y publicado en

ocho volúmenes entre los años 1928 y 1940. La realización de las encuestas dialectales estuvo a cargo de tres lingüistas de renombre y con una gran experiencia: de la parte norte y central de la red del AIS se encargó Paul Scheuermeier (1888-1962), quien además era un excelente etnógrafo y un gran fotógrafo; del sur de Italia y Sicilia se encargó Gerhard Rohlfs (1892-1986) —más tarde fue catedrático de filología románica en Tubinga y Munich y llegó a ser uno de los más grandes romanistas— y de Cerdeña se encargó Max Leopold Wagner (1880-1962), en cuyo honor se han nombrado actualmente en ella calles y plazas.

La palabra y el método *dialectometría* fueron, como es bien sabido, acuñados en el año 1973 por el dialectólogo tolosano Jean Séguy (1914-1973). Séguy fue también el autor del "Atlas linguistique et ethnographique de la Gascogne" (ALG): ante la enorme variabilidad interna de los datos del ALG, se sintió retado a buscar una posibilidad que con métodos cuantitativos le permitiera comprender de forma global esa variabilidad, es decir, desistiendo de muchos detalles cualitativos.[1]

Séguy se introdujo más de forma intuitiva que premeditada en el camino de la "taxonomía numérica", con cuya utilización ya se habían obtenido a principio de los años setenta del siglo pasado grandes éxitos en el análisis cuantitativo de masas de datos biológicos, económicos y psicológicos. Debido a su temprana muerte (1973), Séguy lamentablemente no pudo llegar a madurar su teoría.

Mis primeros intentos dialectométricos empezaron paralelamente a los de Jean Séguy a principios de los años setenta[2] y desde un principio estuvieron explícitamente marcados por la clasificación numérica[3] y la visualización sistemática de los resultados obtenidos a través de la utilización consecuente de la entonces todavía muy joven informática.

Junto a la dialectometría existen actualmente una multiplicidad de otras *metrías*, las cuales comparten todas las siguientes específicas características metodológicas:

a) El tratamiento de datos masivos.
b) El esfuerzo en recoger y descubrir de forma cuantitativa, es decir, con la ayuda de las matemáticas y la estadística, las estructuras ocultas y las regularidades subyacentes en el conjunto de los datos analizados.
c) El afán en medir primero de forma adecuada la masa de datos analizada, lo cual está vinculado a una transferencia de la información que, a veces, puede resultar muy difícil.
d) La dependencia de los métodos cuantitativos aplicados a los principios y planteamientos de cada una de las respectivas disciplinas. Así como la psicometría depende y viene determinada por la psico-logía, y la socio-metría por la socio-logía y la econo-metría por la econo-mía, del mismo modo depende la dialecto-metría de la dialecto-logía y sigue sus pautas.

---

[1] Xavier Ravier da cuenta de la enunciación siguiente de Jean Séguy hecha poco antes de su muerte (en 1973): "Désormais je peux crever tranquille. L'idée fixe qui me hantait depuis trente ans est réalisée: à partir de mille millions de chiures de mouche scrupuleusement intégrées, arriver, par une série d'abstractions à la fois mathématiques et réalistes, à faire tenir le gascon dans une formule ou un schéma." (Ravier 1976: 390).

[2] Véanse mis contribuciones de 1971, 1975 y 1976.

[3] Cfr. las obras todavía fundamentales de Sneath-Sokal (1973), Bock (1974) y Chandon-Pinson (1981).

La dialectometría no es en ningún caso —así lo veo yo por lo menos desde el punto de vista de Salzburgo— una subdisciplina de la estadística o de la geografía cuantitativa, sino de la dialectología o (dicho de forma más exacta) de la dialectología o la geografía lingüística románica. Subrayo de forma especial la palabra *románica*, porque entre la dialectología o la geografía lingüística de los romanistas, los germanistas y los anglicistas, etc., existen ciertas diferencias que no deben ser pasadas por alto ni ignoradas.

La dialectometría que se lleva a cabo en Salzburgo se basa en un fundamento teórico específico y una no menos específica superestructura metodológica.[4] Desde el punto de vista teórico parto de la idea de que la multiplicidad dialectal de nuestros países es el resultado de una actitud lingüística especial de sus habitantes respecto al espacio por ellos habitado. Parece ser que las personas no sólo gestionan de forma especial el espacio por ellas habitado con el trabajo de sus manos, sino que además también lo hacen a través de su "faculté langagière". Por eso hablo yo desde hace algunos años de una "gestión basilectal del espacio por parte del *Homo loquens*".

Aunque los principios teóricos y metodológicos de la dialectometría practicada en Salzburgo fueron publicados ya en el año 1984 en mi trabajo de habilitación con el título de "Dialektometrische Studien", fue a partir del año 1999 cuando los métodos definidos llegaron, desde el punto de vista informático, a "alzar el vuelo". Fue entonces cuando fueron implementados por mi amigo Edgard Haimerl en un genial programa informático llamado "Visual DialectoMetry" (VDM). Todos los gráficos dialectométricos de este artículo han sido generados mediante el uso del VDM.

Ya he mencionado que la dialectometría de Salzburgo se basa en la geografía lingüística románica[5] e intenta enriquecerla y desarrollarla con los conocimientos de la dialectometría. Una de las características principales de la geografía lingüística románica es el uso intensivo de mapas. Este uso tiene sus inicios en la utilización de "mapas mudos" con los que desde un principio se evaluó el atlas lingüístico más importante de los estudios románicos, es decir, el "Atlas Linguistique de la France" (ALF),[6] y termina con la utilización de los mapas de similitud, los mapas de parámetros, los mapas interpuntuales, los mapas de parámetros y los mapas correlativos que van a ser presentados y comentados a continuación. En realidad, en la romanística, desde hace más de cien años, cada geolingüista debería ser un buen cartógrafo.

Al final de esta presentación expondré la quintaesencia de mi exposición resumiéndola en siete tesis.

---

[4] Para una buena comprensión del desarrollo y la situación actual de la dialectometría de Salzburgo pueden verse —con preferencia a las contribuciones escritas en lenguas románicas— mis siguientes libros o artículos: (1981), (1983), (1984) —libro capital desde el punto de vista metodológico—, 1987, 1992, 2003, 2005 y 2008. Véanse tambien los trabajos dialectométricos de Gotzon Aurrekoetxea (1992), Roland Bauer (2009), Lluís Polanco Roig (1984), Xulio Fernández Sousa (2006) y Paul Videsott (2009).

[5] Para una visón global de esta véase el primero volumen de Pop (1950).

[6] Cfr. la presentación de mapas mudos para la explotación geolingüística de los mapas originales del ALF, hecha por Karl Jaberg ya en el año 1906. Véase también el libro introductorio al ALF del mismo autor publicado dos años más tarde (Jaberg 1908) del cual existe igualmente una versión española de 1959.

2. De los mapas-AIS originales a los mapas de trabajo A: algunos principios cartográficos

Los ocho volúmenes del AIS contienen 1705 mapas de gran formato, cada uno de los cuales nos posibilita un fascinante vistazo a la dinámica y biología de los dialectos de la Italo-, Sardo- y Retorromania. Pero yo no puedo, ni quiero conformarme, como dialectómetra, con el estudio de la variación de mapas aislados del atlas, sino que mi intención es la de ir a la búsqueda —tal como hizo Séguy con el ALG— de una variación global más amplia, que incluya la variación interna total del conjunto de los 1.705 mapas del AIS.

Cuando se tiene entre manos un objetivo tan amplio, se tiene que llevar a cabo con métodos escogidos de manera precisa y exacta, lo cual en algunos casos sobrepasa de forma amplia los límites de la lingüística. Estos métodos son, sobre todo, de naturaleza estadística y cartográfica. Voy a empezar con la exposición de los métodos cartográficos.

Contemplando la red del AIS como un espacio, es decir, como una parte de la superficie terrestre habitada por los humanos, se puede imaginar que este espacio fue gestionado (y lo sigue siendo) no sólo desde el punto de vista agrícola, urbanístico o económico, sino además desde el punto de vista dialectal. En el marco del AIS se llevó a cabo la investigación de la gestión basilectal de Italia dentro de una red de 407 puntos de encuesta. En estas 407 localidades se utilizó —lamentablemente no en todas— el cuestionario *normal* del AIS,[7] de manera que, debido a las lagunas de los datos correspondientes, sempre muy molestas desde el punto de vista taxométrico, se tuvo que renunciar, durante la dialectometrización integral del AIS, a los datos de 29 localidades.[8] Quedan así 382 localidades, que pueden representarse como los nudos de una red que pueden por una parte *triangularse* y por la otra *poligonizarse*.

En el lenguaje especializado de la cartografía y de la geometría se habla de "triangulación de Delaunay". Existe además el proceso de *poligonización*, que para nuestro objetivo es mucho más importante. El nombre técnico es "poligonización de Voronoi". Entre la *triangulación* y la *poligonización* existen simples relaciones geométricas.[9]

En un primer momento se triangula la red objetivo de nuestro análisis; después se trazan sobre los lados del triángulo las mediatrices. Estas mediatrices se alargan hasta que se encuentren con las restantes mediatrices. De esta manera surgen los puntos de los ángulos de los polígonos. El resultado final es la teselación de Voronoi.

---

[7] Véanse a este propósito la versión original de la introducción al AIS de Karl Jaberg y Jakob Jud de 1928 y también la traducción italiana publicada en 1987.

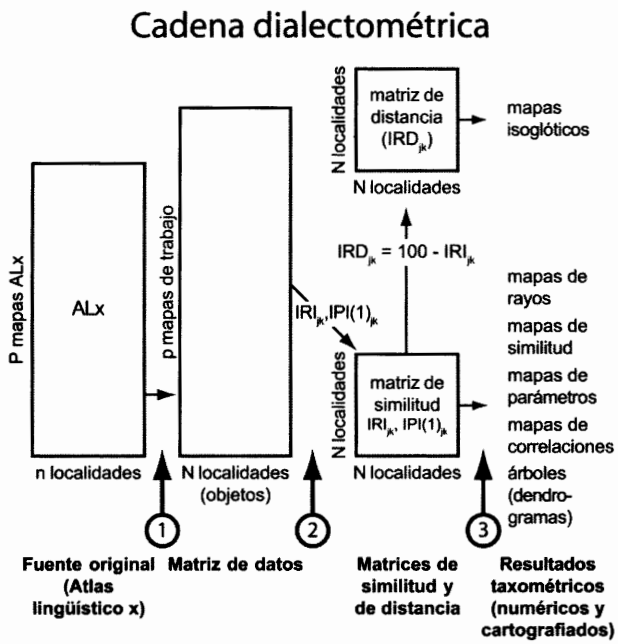
[8] La obtención de la cifra 382 resulta todavía más compleja. Es preciso añadir, a los 407 puntos de encuesta del AIS, cinco puntos suplementarios, correspondientes a las ciudades de Bolonia, Florencia, Milán, Turín y Venecia, donde fueron llevadas a cabo dos encuestas normales, sociolingüísticamente diferenciadas, y dos puntos artificiales equivalentes al italiano y al francés. De la suma así calculada (407 + 5 + 2 = 414) hay que deducir tres encuestas no románicas (griego: PP.-AIS 748 y 792; albanés: P.-AIS 751) y 29 encuestas realizadas por medio de un cuestionario reducido: 414 – 32 = 382.

[9] Para todos los problemas geométricos y computacionales relativos a la triangulación de Delaunay y a la poligonización de Voronoi véase el libro fundamental de Okabe, Boots y Sugihara (1992).

El procedimiento aquí descrito fue empleado por primera vez en el marco de la geografía lingüística general ya en al año 1898 por el germanista Karl Haag, pero fue lamentablemente olvidado posteriormente.

3. De los mapas-AIS originales a los mapas de trabajo B: algunos principios de metrología y de taxación

Paso ahora a presentar los métodos dialectométricos utilizados en Salzburgo, que en la práctica constituyen una forma de cadena: véase la figura 1. A la izquierda el rectángulo representa cada uno de los atlas lingüísticos que va a ser dialectometrizado; aquí se trata del AIS. A partir de una medición basada en principios filológicos tradicionales[10] es posible obtener una matriz de datos de los datos del AIS que está formada por N puntos de encuesta y p mapas de trabajo.[11] Aquí se trata de 382 loca-



Flecha 1: aplicación de la taxación
Flecha 2: selección de la medida de similitud (o de distancia)
Flecha 3: determinación de la explotación taxométrica y visual deseada

Figura 1

Esquema de los métodos dialectométricos utilizados por la Escuela Dialectométrica de Salzburgo

[10] Véase la flecha 1 en la figura 1.

[11] Para una discusión y explicación del concepto de "mapa de trabajo" cfr. Goebl (1984: I, 31-40) y (2008: 33-35).

lidades y de 3.911 mapas de trabajo. Desde el punto de vista de la estadística, a cada mapa de trabajo corresponde un atributo nominal. Además, la matriz de datos con sus 3.911 atributos nominales es bastante grande. Los resultados obtenidos a partir de ella son, por tanto, muy fiables.

Después de elaborar la matriz de datos, ésta tiene que ser evaluada por medio de una medida de similitud adecuada.[12] Para esto la estadística nos ofrece muchas posibilidades, entre las cuales debemos escoger, como dialectómetra, la adecuada a los propios objetivos. Para nuestro objetivo se han mostrado como especialmente adecuadas las medidas IRI ("índice relativo de identidad")[13] e IPI ("índice ponderado de identidad").[14]

Con el IRI o IPI se calcula una matriz de similitud que tiene una dimensión de N por N y por lo tanto es cuadrada. Por medio de una sencilla transformación (s + d = 1) se puede derivar de la matriz de similitud (s) una matriz de distancia (d).

La matriz de similitud pasa a contener la totalidad de las variaciones cualitativas recogidas primero en la matriz de datos, pero en forma cuantitativa compacta. El principio de la compactación cuantitativa de la información cualitativa es muy importante; ahora bien, su comprensión no es fácil. Muchos humanistas no pueden o no quieren comprender ni aceptar este principio. Que esto es así ya lo constaté en los años 70. Hoy en día es todavía así, lamentablemente, en muchos casos. Para comprender la dialectometría se debe, por tanto, comprender el principio de transferencia de la información del nivel de la cualidad al de la cantidad.

A la derecha de la figura 1 se pueden ver diferentes valoraciones cuantitativo-cartográficas que nosotros empleamos en Salzburgo para visualizar nuestros cálculos dialectométricos.[15] La estadística y también la cartografía o la visualística moderna tienen disponible todavía muchas más posibilidades de explotación de las matrices de similitud. Nosotros en Salzburgo, de manera consciente, siempre nos hemos limitado a unas pocas de estas posibilidades especialmente escogidas. Ello se debe a que estas se adaptan bien a los planteamientos de la geografía lingüística románica y permiten así tender un puente claramente reconocible entre el saber tradicional de la geografía lingüística románica y el conocimiento nuevo de la dialectometría. Además, contemplamos el espacio a analizar como invariable y las diferentes formas en que se manifiesta la "gestión basilectal del espacio" como las variables a analizar.

¡Empecemos nuestro paseo en el marco de la cadena dialectométrica a la izquierda, es decir, en el momento de la realización de los llamados "mapas de trabajo"!

Se trata de una actividad que en el marco de la geografía lingüística románica se ha realizado desde hace más de cien años,[16] con gran éxito, miles de veces. No sólo existen en nuestra disciplina muchas publicaciones al respecto sino que además hay un amplio consenso sobre cómo y por qué se realizan estos análisis. Se pueden llevar a cabo taxaciones fonéticas y léxicas: véanse los mapas 1 y 2. En Salzburgo denominamos "taxación" al proceso de realización de los mapas de trabajo nominales. Cada mapa de tra-

---

[12] Véase la flecha 2 en la figura 1.

[13] Para la explicación y definición del IRI cfr. Goebl (1981: 357-361) y (1984: I, 75-79).

[14] Para la explicación y definición del IPI cfr. Goebl (1987: 67-79).

[15] Véase la flecha 3 en la figura 1.

[16] Véanse a este propósito los mapas tipificados (y coloreados) en Brun-Trigaud, Le Berre y Le Dû (2005) y en Veny (2007-2009).

bajo contiene un número exactamente definido de "taxatos" a los cuales corresponden áreas geográficas más o menos grandes y dotadas de configuraciones muy variables.

Una taxación fonética se puede efectuar sólo con aquellos mapas del AIS que se basan en un mismo étimo. El mapa 1 se refiere al mapa 27 del AIS (*il suo cognato*) y al étimo latino COGNÁTU para "cuñado". En este caso se analizó solamente la U latina final. El mismo mapa del AIS permitiría analizar también el desarrollo de la C inicial, de la O pretónica, del grupo -GN-, de la A acentuada o de la -T- intervocálica. Se puede apreciar, pues, que mapas del AIS con una etimología homogénea se pueden taxar de múltiples maneras desde el punto de vista fonético. En nuestro análisis dialectométrico del AIS hemos extraído a partir de 257 mapas originales del AIS, cada uno de ellos con un mismo étimo, un total de 1766 mapas de trabajo fonéticos.

El mapa 2 constituye uno de los 1225 mapas de trabajo léxicos, que hemos realizado en Salzburgo. Se trata de las denominaciones utilizadas para *sobrina,* que se encuentran en el mapa 22 (*la vostra nipote*) del AIS.

Cuando se interpreta la estratificación geográfica de estos taxatos a la luz de la lingüística histórica, se puede reconstruir el surgimiento de esta estratificación. Cada una de las taxaciones de este tipo nos cuenta en cierta manera una pequeña historia. Se puede suponer que de la síntesis de muchas de estas historias parciales se puede llegar a conseguir algo semejante a una historia global del espacio analizado.

Téngase en consideración que lo que se ve en los mapas de trabajo está basado sobre todo en una multitud de diferentes áreas. En el mapa 1 figuran diez áreas de diferentes tamaños y de muy diferentes configuraciones geográficas. Las diferencias de tamaño son fáciles de indicar: el área más grande —aquí con el número uno, de color rosa— comprende 198 puntos de encuesta, es decir 198 polígonos. Las áreas más pequeñas —aquí con los números 10 y 11 y señaladas con tonos azules— están constituidas por un único punto de encuesta o por un solo polígono.

Cada uno de los mapas de trabajo tiene, por lo que se refiere al número de sus taxatos (o áreas), así como a la extensión y la forma de las respectivas áreas en el espacio, un perfil diferente. En el marco de la dialectometría de Salzburgo es muy importante el análisis exacto de las áreas y de sus dimensiones.

El análisis cuantitativo de las áreas taxatorias de las cuales dispone cada uno de los puntos de una red de atlas, es muy clarificador para una mejor comprensión de los mecanismos y regularidades que determinan el funcionamiento comunicativo de la respectiva red. Se puede decir metafóricamente hablando que cada área taxatoria funciona como una "oportunidad comunicativa". Cuanto más grandes son las superficies medias de las áreas taxatorias de una localidad, tanto mejores son sus oportunidades comunicativas. La figura 2 visualiza los patrimonios areales de dos puntos de nuestra red. El punto 1 (Breil/Brigels), ubicado en los Grisones, dispone de oportunidades comunicativas muy inferiores a las del punto 376 (Venecia). Esto se explica por el hecho de que tiene en su patrimonio areal muchas áreas de tamaño muy reducido: véase la primera columna del histograma con 1288 áreas cuya extensión varía entre 1 y 38 (= 10% de 382) puntos, mientras que en el patrimonio areal de Venecia se encuentran sólo 571 áreas de tamaño muy reducido y, por el contrario, muchas áreas de tamaño más grande: véanse al respecto las tres últimas columnas del histograma de Venecia (a la derecha) que señalan la presencia de 411, 610 y 799 áreas cuyas extensiones varían entre 271 (= 71% de 382) y 308, 309 (= 81% de 382) y 347 así

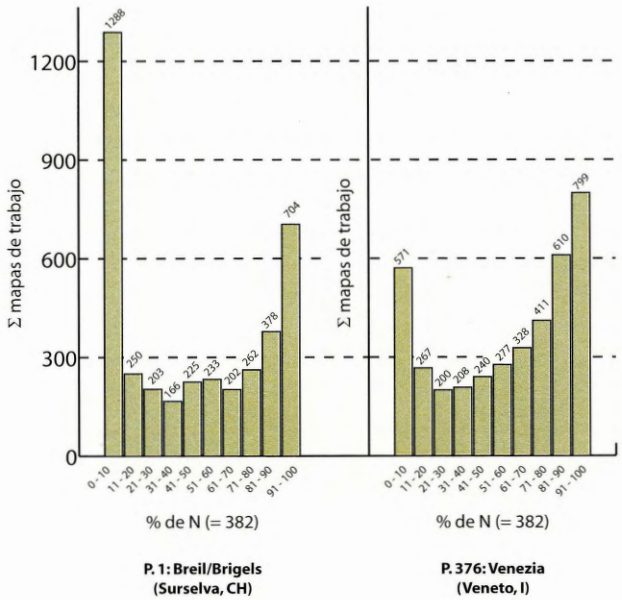


Figura 2

Histogramas relativos a la composición interna de los patrimonios areales de los puntos-AIS 1 (Breil/Brigels) y 376 (Venecia)

como 348 (= 91% de 382) y 382 puntos respectivamente. El rendimiento análogo del punto 1 es, de lejos, inferior.

La determinación de los taxatos y al mismo tiempo de sus áreas sólo puede llevarse a cabo aplicando los conocimientos que provienen del fundo clásico de la lingüística románica. Se puede decir entonces que la dialectometría de la Escuela de Salzburgo, en lo que se refiere a sus fundamentos, se sustenta sobre dos pilares metodológicos:

1) áreas geolingüísticas (de carácter nominal);
2) la participación, en el proceso de constitución de la matriz de datos, de los conocimientos especializados (*expert knowledge*) de la lingüística románica.

Esto es lo que la diferencia, fundamentalmente, de otras escuelas dialectométricas como, por ejemplo, de la de Groninga. Pero pasemos ahora a la medición de similitud.

4. De la matriz de datos a la matriz de similitud

Como índice estándar de la medición de la similitud se puede considerar el IRI ("índice relativo de identidad"),[17] que constituye la mejor medida de similitud de

[17] El nombre originario alemán de este índice es "Relativer Identitätswert", el cual se tradujo inicialmente en francés en la forma "indice relatif d'identité" (IRI). Posteriormente las restantes denominaciones románicas fueron acuñadas según la designación francesa.

nuestra dialectometría. Desde el punto de vista matemático el IRI es muy sencillo. Mide la relación entre el número de parejas de mapas de trabajo idénticos y la cantidad de mapas de trabajo disponibles que existen en los vectores de dos puntos de encuesta a comparar. Cuando en un mapa de trabajo aparece el mismo taxato entre los dos puntos de encuesta a comparar, hablamos de una *co-identidad*. En el caso contrario hablamos de una *co-diferencia*. En el *numerador* de la fórmula del IRI se halla siempre el número de las *co-identidades* y en el *denominador* la suma de las *co-identidades* y *co-diferencias*. Los resultados de todos los cálculos IRI se recogen en la matriz de similitud. Ésta es siempre cuadrada, tiene a lo largo de la diagonal siempre el valor uno (o cien) y está compuesta de dos mitades totalmente idénticas. Los valores de un vector de la matriz de similitud son la base de una mapa de similitud.

Véanse los mapas 3 y 4.

El mapa de similitud es el instrumento heurístico más importante, y a la vez el más simple, de la dialectometría de Salzburgo. Al mismo tiempo es la base de todos los complejos instrumentos heurísticos de los que disponemos. Cada uno de los mapas de similitud tiene un punto de referencia que siempre aparece como un polígono blanco: aquí son la ciudad piamontesa de Turín (mapa 3) y la pequeña localidad grisona de Domat/Ems (mapa 4). El resto del mapa muestra la muy variable estratificación del espacio en cuanto a las similitudes cuantitativas de los locolectos de los restantes 381 puntos de encuesta con respecto al locolecto de la localidad de referencia. De este modo el mensaje principal de cada uno de los mapas de similitud es cuantitativo, y ya no cualitativo, como es el caso de los mapas de trabajo (véanse los mapas 1 y 2). Los colores ordenados según el arco iris sirven para representar la variación numérica de los valores de medición de la distribución de similitud.

En el mapa 3 los valores de la distribución de la similitud varían entre 45,88% y 83,10%; en el mapa 4 la misma variación se sitúa entre 38,34% y 84,11%. La formación de las seis clases (o intervalos) se lleva a cabo utilizando un algoritmo especial que aquí se llama MINMWMAX. Para la confección de los mapas 17-20 y 27-32 utilizamos el algoritmo MEDMW que se diferencia ligeramente de MINMWMAX creando perfiles coropléticos más accidentados.[18]

Las técnicas de visualización que se verán a continuación no son, en principio, nada especial: se corresponden con los estándares internacionales de la cartografía cuantitativa. A pesar de ello he realizado aquí una selección especial, que se corresponde exactamente con las necesidades de nuestra dialectometría y que en su totalidad fue adoptada por el programa dialectométrico VDM.

El valor heurístico de los mapas de similitud reside en la generación de patrones espaciales cuantitativos, que son típicos para el correspondiente punto de referencia. En el mapa 3 se advierte que las mayores similitudes con el locolecto de Turín aparecen en el Piamonte y las mayores diferencias con este dialecto se encuentran en Cerdeña, en amplias zonas del sur de Italia, en el Valle de Aosta y en los Grisones. Son significativos en este sentido los polígonos de color azul claro y azul oscuro. Además todos los polígonos con colores calientes se encuentran por encima de la media arit-

[18] Para una presentación detallada de estos algoritmos cfr. Goebl (1981: 361-368), (1983: 370-376), (1984: I, 93-98) y (1987: 79-83).

métrica de la correspondiente distribución de similitud;[19] los polígonos con colores fríos se encuentran por debajo.

En los mapas 3-14 se constata, además, que la disminución de la similitud lingüística en el espacio se corresponde claramente de alguna manera con el aumento de la distancia del punto de referencia. El patrón de disminución de un mapa de similitud puede calificarse como típico del punto de referencia seleccionado y de sus alrededores. Perfiles coropléticos similares al del mapa 3 se pueden obtener también de los otros puntos de encuesta del Piamonte. Lo mismo vale por lo que respecta al resto de los mapas de similitud.

La cantidad de datos en la que se basan los mapas de este artículo es enorme. Se trata de 3.911 mapas de trabajo y de 43.564 áreas taxatorias pertenecientes a todas las categorías lingüísticas. Además, es importante constatar que los mapas de similitud también se pueden interpretar de forma extralingüística. De esta manera es posible hacer comparaciones analógicas con el comportamiento de abonados de una red telefónica o de misioneros que intentan difundir sus ideas en el interior de una red comunicativa espacialmente definida. Es importante tener en cuenta esto para comprender la relevancia interdisciplinar de la dialectometría.

Consideremos ahora los mapas 4 y 6: cada uno de estos dos mapas de similitud tiene muy diferentes puntos de referencia. El mapa 4 tiene un punto de referencia en la parte central de los Grisones (P. 5: Domat/Ems); el mapa 6, por el contrario, en la ciudad de Venecia (P. 376). Las diferencias entre ambos perfiles coropléticos se corresponden con la totalidad de sus seis intervalos. Mientras en el mapa 4 las similitudes superiores a la media aritmética se extienden desde el locolecto de Domat/Ems hasta aproximadamente los Apeninos septentrionales (es decir la conocida "línea La Spezia-Rimini"), en el mapa 6 las similitudes por encima de la media aritmética abarcan casi la totalidad del norte de Italia e, incluso, la totalidad de la Italia central hasta aproximadamente la "línea Roma-Ancona". Obsérvese, además, con atención en ambos casos la extensión en el espacio de los polígonos de color rojo y naranja: por ellos se reconoce el núcleo del correspondiente tipo dialectal.

Los mapas 7 y 8 muestran de una parte la posición relacional de un dialecto toscano septentrional y de la otra la del italiano estándar, el cual ha sido agregado a nuestra red de investigación como punto artificial. ¡Nótese la práctica identidad de los dos perfiles coropléticos!

Se puede variar la forma de los perfiles coropléticos de los mapas de similitud mediante la selección de otros algoritmos de intervalación u otras medidas de similitud: véanse los mapas 9 y 10.

A la izquierda (mapa 9) se ve un perfil de similitud típico de la Campania, que ha sido realizado mediante el algoritmo de intervalación estándar MINMWMAX. A la derecha (mapa 10) se visualizaron los mismos valores de medición mediante otro algoritmo de intervalación (MEDMW), que hace destacar de forma más clara la división interna del perfil coroplético. El programa VDM nos permite realizar este tipo de variaciones en la visualización de manera muy rápida.

---

[19] En el mapa 3 la media aritmética es de 61,88%. Este valor constituye sempre el umbral superior de la clase 3. En el mapa 4 el mismo parámetro es de 50,13%.

4.1. La posición dialectométrica de una isla lingüística occitana situada en el sur de Italia: Guardia Piemontese (P. 760)

Préstese atención, en los perfiles coropléticos de los mapas 9 y 10, a los polígonos diseminados de color verde y amarillo en medio de la zona roja del sur de Italia y de Sicilia. Se trata en su mayoría de polígonos que pertenecen a las islas lingüísticas de tipo galorrománico (en Sicilia: PP. 865, 836, 817), occitano (en Calabria: P. 760) y francoprovenzal (en Apulia: P. 715).

Ahora pasemos a considerar los mapas 11 y 12, los cuales constituyen un caso muy especial de utilización de la medida de similitud. Se trata de mostrar el grado de integración lingüística de la isla lingüística occitana de Guardia Piemontese en la totalidad de la red del AIS. La localidad de Guardia Piemontese fue fundada en el siglo XIII por colonos del sudoeste del Piamonte, quienes llevaron su dialecto occitano-alpino hacia Calabria. Este dialecto está, pues, desde hace ya más de siete siglos en contacto con su entorno calabrés.

En el mapa 11 se ve un típico perfil de similitud calabrés: las mayores similitudes abarcan hasta la "línea La Spezia-Rimini", y el respectivo punto de referencia (P. 761-Mangone) se encuentra en los alrededores inmediatos de Guardia Piemontese (P. 760).

A la derecha (mapa 12) se halla el perfil de similitud de Guardia Piemontese, el cual es muy diferente del de Mangone. Las mayores similitudes —representadas por medio de los polígonos de color rojo y naranja— se encuentran en la parte occidental del norte de Italia. El valor de similitud más grande (65,28%) se halla en una localidad (P. 181)[20] que se encuentra apenas a unos 60 kilómetros, desde la que hace más de 700 años fue colonizada Guardia Piemontese.[21] Este ejemplo nos muestra que, en ciertas condiciones, es posible determinar la patria lingüística de una isla lingüística mediante la medición de la similitud. Pero esto presupone que esta patria ha de encontrarse en el interior de la red analizada.

Véanse los mapas 13 y 14.

Sabiendo que la intensidad de un contacto lingüístico es normalmente mayor en el léxico que en la fonética, puede plantearse también en este caso la misma pregunta. Y de hecho es así también aquí: el perfil de similitud de la izquierda (mapa 13) muestra que la vinculación fonética del dialecto de Guardia Piemontese con su antigua patria está todavía totalmente intacta, mientras que el mapa coroplético de la derecha (mapa 14) muestra de manera totalmente clara que —por lo que se refiere al léxico— el contacto lingüístico con el ambiente calabrés ha llevado a una mayor asimilación.

---

[20] En el mapa 12 el polígono del punto 181 está marcado por un rayado blanco cruzado y una flecha.

[21] Se trata de la parte superior de la Val Pellice (en el Piamonte occidental).

5. Al otro lado de la matriz de similitud

Otro capítulo muy interesante de la dialectometría de la Escuela de Salzburgo es el de los "mapas de parámetros", de los cuales aquí sólo voy a presentar por falta de espacio el mapa de sinopsis de los "coeficientes de asimetría de Fisher".

En los mapas de parámetros se trata de realizar una sinopsis de determinados parámetros de la distribución de similitud para luego visualizarlos. Así, aparecen perfiles coropléticos muy útiles para aclarar determinados fenómenos de la geografía lingüística.

Véanse otra vez los mapas 3 y 4.

Consideremos de nuevo los mapas 3 y 4 que constituyen visualizaciones de dos distribuciones de similitud diferentes. Lo que aquí nos interesa de manera especial son las diferencias estadísticas que subyacen entre las dos distribuciones de similitud: préstese atención a las diferentes formas de ambos histogramas en la mitad inferior derecha de los dos mapas.

El histograma del mapa 3 muestra una distribución de similitud muy simétrica, mientras que el histograma del mapa 4 muestra una distribución de similitud asimétrica y desplazada hacia la izquierda. Las diferencias en la simetría de una distribución de similitud tienen un significado geolingüístico especial. Nos muestran hasta qué punto un dialecto está integrado en la totalidad de la red, y también si toma parte en muchos fenómenos de contacto lingüístico o si, por el contrario, está aislado. Con la ayuda del "coeficiente de asimetría" propuesto por el estadístico inglés Ronald A. Fisher (1890-1962) [CAF] se puede medir muy bien la variación de la simetría en una distribución de similitud.[22]

Véanse los mapas 15 y 16.

El mapa 15 muestra la visualización del resultado de la sinopsis de los 382 valores del CAF. Para poder comprender claramente su perfil coroplético tan bien estructurado, es necesario poder interpretar correctamente desde el punto de vista lingüístico los valores de medición así como los colores correspondientes: los polígonos azules —los cuales visualizan siempre los valores del CAF más bajos— muestran una gran interacción con el resto de la red del AIS, mientras los polígonos de color rojo —correspondientes a los valores del CAF más altos— muestran por su parte una muy escasa interacción. Denomino esta interacción "compromiso lingüístico", siguiendo el ejemplo de la germanística, en la que este fenómeno está muy bien descrito bajo el nombre alemán de "Sprachausgleich".[23]

La ubicación geográfica de los polígonos de color azul oscuro es típica: los encontramos a lo largo de los Apeninos, en Liguria, en los Abruzos y en el Lacio, así como en el interior del norte de Italia a lo largo del río Adigio y justo al sur de los dialectos retorrománicos de los Grisones y Ladinia. Todos estas zonas siempre han llamado la atención de los dialectólogos por su intenso contacto e intercambio lingüístico.

En el norte de Cerdeña y en el sur de Italia hay también polígonos de color azul oscuro. Los de norte de Cerdeña indican una intensiva toscanización de la zona desde hace alrededor de 300 años, y los de sur de Italia así como los de Sicilia apuntan a las islas lingüísticas galorrománicas (de tipo galoitálico: PP: 865, 836, 817, de tipo occi-

[22] Para una presentación pormenorizada del CAF cfr. Goebl (1981: 394-401) y (1984: I, 150-153).

[23] Véase el libro inspirador de Besch (1967).

tano: P. 760 y de tipo francoprovenzal: P. 715) que allí se encuentran. Puesto que cada isla lingüística tiene que ver *per definitionem* con contacto e intercambio lingüístico, es evidente que allí tienen que aparecer valores de CAF muy bajos: el valor mínimo del CAF se encuentra de hecho en la isla lingüística de Guardia Piemontese (P. 760).

Los valores de CAF muy elevados por encima de cero se corresponden con los polígonos rojos; estos nos muestran las zonas que tienen poco que ver con contacto lingüístico y que, por lo tanto, disponen de una homogeneidad lingüística interior relativamente alta. Añado también que en el mapa 15 la visualización no se refiere a seis, sino a ocho clases de valores: esto permite hacer más evidentes las múltiples gradaciones del efecto del "compromiso lingüístico".

Debido a la fina división del perfil del mapa puede ser preferible realizar una visualización no en la forma acostumbrada de un mapa coroplético sino en forma de un estereograma: véase el mapa 16 de la misma página.

Los mapas 17 y 18 muestran que los perfiles coropléticos logrados mediante la sinopsis de los coeficientes de asimetría de Fisher pueden variar según las diferentes categorías lingüísticas a las que pertenecen los mapas de trabajo analizados. Desde el punto de vista lingüístico las diferencias entre los mapas 17 y 18 son muy interesantes: mientras en el mapa 17 (relativo a datos fonéticos) la ubicación de los polígonos de color azul oscuro crea algunas agrupaciones lineares que fraccionan ante todo el espacio central de Italia de manera muy clara, en el mapa 18 (relativo a datos léxicos) los polígonos de color azul oscuro —típicos de un alto nivel de interacción lingüística— se concentran sobre todo en Emilia, Romaña y Liguria. Surge así la certeza de que los intercambios léxicos (mapa 18) tienen una implantación espacial diferente de sus equivalentes fonéticos (mapa 17).

6. La dialectometría interpuntual: mapa isoglótico y mapa de rayos

Véanse los mapas 19 y 20.

El mapa 19 constituye una síntesis dialectométrica de isoglosas, cuyo mensaje se corresponde de manera amplia con los obtenidos por la síntesis de isoglosas tradicionales como las que, por ejemplo, fueron realizadas para la red del AIS por los lingüistas Gerhard Rohlfs (1947) o Robert A. Hall Jr. (1943). El mapa isoglótico de tipo dialectométrico es naturalmente mucho más exacto y rico en detalles. Además se basa en muchos más datos de los que se pueden abarcar por medio de un análisis hecho a mano.

La totalidad de la sintaxis de la imagen se basa en 970 lados de polígonos, en los que se visualiza un número igual de medidas no de similitud sino de distancia. Estas medidas de distancia se refieren siempre a las diferencias lingüísticas entre dos puntos de encuesta contiguos: de ahí el nombre de *análisis inter-puntual*.[24]

El principio cartográfico es sencillo: cuanto más grandes sean las medidas de distancia, más gruesas y azules se visualizan. Y a la inversa: las medidas de distancia más pequeñas se visualizan más finas y más rojas. Se reconocen efectos muy claros de división del espacio sobre todo por medio del grosor de los lados de los polígonos de color azul oscuro y su distribución espacial; préstese atención a la múltiple división interna del

[24] El término "interpoint" fue acuñado por el lingüista francés Théodore Lalanne (1953: 266).

norte de Italia, la clara separación de las variantes galorrománicas del Valle de Aosta y del Piamonte Occidental y el aislamiento de las variedades retorrománicas en el norte.

Señalo además que la "línea La Spezia-Rimini" está mucho mejor marcada que la no menos citada "línea Roma-Ancona". Finalmente se puede ver claramente, como era de esperar, que las islas lingüísticas del sur de Italia y del norte de Cerdeña están rodeadas por gruesos haces de isoglosas.

A la derecha (mapa 20) se ve un mapa de rayos, que es la contrapartida cartográfica del mapa isoglótico. La sintaxis de la imagen se basa en 970 lados de triángulo, a lo largo de los cuales se visualiza el mismo número de valores de similitud. La concentración de gruesos lados de triángulo de color rojo nos señala paisajes lingüísticos interpuntuales especialmente bien enlazados. En este sentido, sobresalen claramente la Toscana, parte de Lombardía y el Véneto, así como Sicilia y el sur de Cerdeña. Se pueden apreciar, sin embargo, zonas que están marcadas por un mínimo contacto entre las localidades vecinas: estas son las zonas donde aparecen muchos triángulos finos de color azul. De este modo, allí donde en el mapa de la izquierda (mapa 19) se ven lados de polígono gruesos de color azul, en el mapa de la derecha (mapa 20) se ven lados de triángulo finos de color azul.[25]

Hemos utilizado para la visualización de los cálculos interpuntuales el algoritmo MEDMW, con ocho clases para conferir una estructura más accidentada a los perfiles interpuntuales de los dos mapas. El algoritmo MEDMW utiliza, como el algoritmo MINMWMAX, la media aritmética para la separación de los valores "altos" y "bajos". Pero en el interior de los recorridos estadísticos situados entre la media aritmética y el máximo de una parte y la media aritmética y el mínimo de otra, el algoritmo MEDMW[26] crea clases con un número de polígonos tan igual como sea posible.

7. La dialectometría dendrográfica

Paso ahora a la presentación de la dialectometría dendrográfica, cuya práctica es muy sencilla con el VDM. Los análisis dendrográficos son habituales en muchas ciencias humanas y naturales, pero para poder emplearlos de manera conveniente, hay que conocer bien el funcionamiento de los algoritmos en ellos empleados. Esto es imposible, desde luego, si no se tiene un mínimo de conocimientos matemáticos.

El programa VDM puede, basándose en seis algoritmos dendrográficos diferentes,[27] no sólo calcular y visualizar los respectivos árboles muy rápidamente, sino que, además, puede mostrar el resultado de la clasificación de estos árboles en el espacio ("espacialización"), es decir, proyectarlo sobre el mapa correspondiente. Esto sucede de maneras muy variadas y gracias a la utilización de colores.

---

[25] Préstese atención al hecho de que las siluetas de los histogramas de los mapas 19 y 20 son enteramente simétricas.

[26] Para el algoritmo MEDMW cfr. Goebl (1984: I, 95) y (1987: 81-82).

[27] Se trata de los algoritmos jerárquico-aglomerativos siguientes: *Single Linkage, Complete Linkage, Ward Method, Simple Average Linkage, Average Linkage (UPGMA)* y *Centroid Method.* En la mayoría de los casos los algoritmos *Centroid Method* y *Single Linkage* producen resultados que no son útiles para la dialectometría.

A la izquierda (mapa 21) figura un árbol genealógico calculado con el método de "Complete Linkage".[28] El cálculo ha sido realizado a partir de los datos de las matrices de similitud que fueron utilizadas anteriormente. En este caso se trata de una matriz de similitud calculada con el índice de similitud IRI y mediante una matriz de datos que comprende nuestro corpus integral de 3.911 mapas de trabajo (los cuales abarcan todas las categorías lingüísticas).

Todos los árboles calculados con algoritmos jerárquico-aglomerativos tienen una estructura binaria y se construyen de izquierda a derecha, es decir, de las hojas a la raíz. Esto se lleva a cabo mediante parejas de fusiones de los elementos más similares desde el punto de vista cuantitativo, siendo el algoritmo utilizado el que define la correspondiente similitud cuantitativa.

El árbol obtenido puede ser interpretado por los lingüistas en dos sentidos: primero de la raíz a las hojas; una interpretación de este tipo es relevante desde el punto de vista diacrónico y se utiliza a menudo en la estadística léxica y en la glotocronología. El otro sentido interpretativo va de las hojas a la raíz. Éste se revela muy importante, pues sirve para mostrar las correspondientes dependencias o similitudes.

En el presente caso he definido en el árbol de clasificación calculado 16 unidades importantes desde el punto de vista geolingüístico, a las que denomino "dendremas": véase el mapa 21. En la espacialización correspondiente (véase el mapa 22) a cada uno de los dendremas le corresponde un sector espacial muy coherente, al que llamo "corema". Los dendremas y los coremas de los mapas 21 y 22 tienen siempre el mismo número y el mismo color.

Para interpretar el resultado de un análisis dendrográfico de este tipo es necesario utilizar siempre el árbol y el mapa de forma paralela. Al observar el árbol (mapa 21) se aprecia que en la primera bifurcación después de la raíz (A), de las dos ramas superiores (C y D) cuelgan los dendremas 1-8 y de las dos ramas inferiores (E y F) los dendremas 9-16. Sólo al observar la espacialización (mapa 22) se puede apreciar que en el mapa la línea separadora de estos dos grupos (A versus B) *grosso modo* discurre paralela a los Apeninos, es decir, a la conocida "línea La Spezia-Rimini". Es, por tanto, el fraccionamiento más importante de nuestra red que se puede encontrar mediante la utilización de este método.

Cuando pasamos en la parte superior del árbol a la bifurcación C, observamos que la totalidad de este maxi-dendrema relativo a la parte superior de Italia se divide en pasos sucesivos en el romanche de los Grisones y el ladino dolomítico (dendrema 1) y en un gran sub-dendrema, el cual incluye los dendremas (y los respectivos coremas) 2-7 que abarcan el resto del norte de Italia a excepción de las zonas de hablas galorrománicas a lo largo de la cadena alpina occidental (dendrema/corema 8 = rama D).

Interesantes son también las conexiones de zonas complejas desde el punto de vista clasificatorio: así los Grisones romanches y la Ladinia dolomítica (dendrema/corema 1) están ligados a un bloque nordoccidental de la parte superior de Italia, el de los dendremas 2 a 7 y que contiene Lombardía, el Véneto, Emilia-Romaña y el Friul.

---

[28] Para una presentación pormenorizada de los algoritmos jerárquico-aglomerativos en general y del algoritmo "Complete Linkage" en particular, véanse Sneath-Sokal (1973: 356 s.), Bock (1974: 356 s.) y Chandon-Pinson (1981: 94 s.).

Las zonas galorrománicas del Piamonte constituyen un propio dendrema-corema, el número 8, que, a su vez, está segregado del resto del Piamonte, el número 3.

No menos interesantes son las conexiones de las islas lingüísticas del sur de Italia: véanse a este respecto los dendremas/coremas 13 (que abarcan los dialectos galoitálicos de los PP. 817, 836, 865) y 14 (que comprende el occitano de Guardia Piemontese [p. 760] y el francoprovenzal de Faeto [P. 715]).

Pero el gran valor heurístico del análisis dendrográfico sólo se puede apreciar cuando se utiliza de forma comparativa y se comparan árboles que, basados en los mismos datos, han sido calculados utilizando diferentes algoritmos.

8. La dialectometría correlativa

Llegamos así a un capítulo especialmente sugestivo de la dialectometría de Salzburgo, que existe desde 2004, cuando el creador de VDM, Edgar Haimerl, lo incluyó en su programa: se trata de la "dialectometría correlativa".[29] Esta permite comparar respectivamente los N vectores de dos matrices de similitud (o de proximidad) establecidas, por parejas, mediante un cálculo de las correlaciones y visualiza de forma inmediata los N valores de correlación obtenidos. Los mapas de correlaciones producidos de esta manera nos informan de si las dos dimensiones correlacionadas gestionan el espacio analizado de manera convergente o de manera divergente. Recuerdo aquí a este respecto el concepto mencionado al principio de "la gestión dialectal del espacio por el hablante", que ha sido desarrollado en Salzburgo.

Además, la experiencia adquirida hasta el momento en el análisis dialectométrico de diferentes córpora lingüísticos corrobora la hipótesis de que la gestión dialectal del espacio por el hablante —que en todas las ocasiones nos ha dado claros y muy ordenados perfiles espaciales— viene regida por "leyes espaciales específicas", semejantes a las regularidades o "leyes" válidas para el cambio fonético o la variación lingüística a lo largo del eje diacrónico.

Como es sabido, las leyes fonéticas fueron descubiertas en el último cuarto del siglo XIX por los Neogramáticos de Leipzig, siendo más tarde corroboradas en todas las ocasiones. Las leyes espaciales mencionadas serían la correspondencia diatópica de las leyes descubiertas por los Neogramáticos de Leipzig, las cuales funcionan en la dimensión diacrónica.

Véanse los mapas 23-26.

Demuestro primero las posibilidades de correlación entre el lenguaje y el espacio. En el mapa 23 se observa un nuevo perfil de similitud: se basa en la medición de similitud correspondiente al punto de referencia de Nápoles (en Campania). En el gráfico opuesto (mapa 24) se halla un tipo de mapa que aún no hemos presentado: se trata de un mapa de proximidad. Visualiza la estratificación de las proximidades geográficas a partir de un punto de referencia. En él se calculan las proximidades espaciales —no las distancias— entre los 382 puntos de encuesta del AIS mediante la utilización del teorema de Pitágoras. Para ello se utilizan las coordena-

[29] Véanse a este propósito nuestra contribución introductora de 2005 y los respectivos esquemas de cálculo en Goebl (2005: 328) y (2008: 53).

das *x* e *y* de todos los puntos de encuesta, las cuales están archivadas desde el principio en la base de datos del VDM. Desde los puntos de vista taxométrico y cartográfico las matrices calculadas de proximidad euclídea se pueden tratar como cualquier otra matriz de similitud lingüística.

¿Qué significa, pues, la diferencia entre ambas visualizaciones? A la izquierda (mapa 23) se encuentra un mapa de similitud basado en los principios de la gestión basilectal de Italia a través de sus habitantes; a la derecha (mapa 24) se halla un mapa que se refiere a la gestión euclídea del mismo espacio. En ambos casos se puede apreciar una disminución de los valores de medición en tanto que aumenta la distancia con respecto a Nápoles (P. 721) como punto de referencia. Cada uno de estos mapas se basa en 381 valores de medición. Cuando correlacionamos las dos series de valores de medición mediante la utilización del coeficiente de correlación r(BP) de (Auguste) Bravais (1811-1863) y (Karl) Pearson (1857-1936), obtenemos un valor de +0,891, que se puede observar en la parte central de la respectiva página (véanse los mapas 23 y 24). Este valor, en principio, como lingüistas, no nos dice nada.

En los mapas 25 y 26 se puede observar el mismo proceder a partir de un punto de referencia en Friul (P. 318). Las tendencias que aparecen en ambos perfiles coroplético s son las mismas que antes: disminución de los valores de medición con el aumento de la distancia con respecto al punto de referencia: a la izquierda (mapa 25) podríamos decir de forma "humana" y a la derecha (mapa 26), si se quiere, de forma "natural". El valor de correlación r(BP) correspondiente es +0,803. Viendo esto nos puede venir la idea de realizar el mismo cálculo de correlación para la totalidad de los 382 puntos de encuesta y visualizar sobre el mapa los 382 valores obtenidos de forma sinóptica.

¿Qué aspecto tiene el perfil coroplético que se obtiene de la correlación de los N vectores acoplados de una matriz lingüística de similitud y una matriz euclídea de proximidad?

Los mapas 27 y 28 proporcionan la respuesta: constituyen dos visualizaciones que, aun basadas en los mismos datos estadísticos, son diferentes en la fineza de su presentación. La visualización de la derecha (mapa 28) —hecha con ocho clases icónicas— muestra más detalles y una más fina escala de niveles que la de la izquierda (mapa 27), hecha con sólo seis clases icónicas. De nuevo se necesita, para poder interpretar estos dos mapas, una exacta comprensión del significado lingüístico de los valores de medición y de los colores.

En general, el color azul se refiere a valores pequeños —algunas veces incluso negativos— de la medición de r(BP). El color rojo se refiere a valores grandes de la medición r(BP).

Valores de medición pequeños, es decir, polígonos azules, significan que la gestión lingüística del espacio y la euclídea suceden según principios diferentes y por lo tanto no van a la "misma cadencia" ni en armonía.

Elevados valores de medición, es decir, polígonos de color rojo, significan lo contrario: el hecho de que la difusión de las similitudes lingüísticas en el espacio se sucede de una forma amplia según principios euclidianos o "naturales".

Allí donde encontramos polígonos azules y, por tanto, tiene lugar una fisura entre las gestiones del espacio y de la lengua, tienen que entrar en juego fuerzas que no son de origen natural, sino que tienen un origen antrópico. Realmente la distribución de los colores rojo y azul es, así como la de todos los semitonos, muy indicativa.

Zonas con elevado disturbio antrópico —aquí de color azul oscuro (correspondiente a las clases icónicas 1 y 2)— se encuentran en el Véneto, el norte de la Toscana y en Liguria, cuando se dejan de lado las cinco islas lingüísticas galorrománicas del sur de Italia y Sicilia, así como la parte norte de Cerdeña. Además, aparece el valor de medición más pequeño de todo el mapa en la isla lingüística de Guardia Piemontese (P. 760), ya conocida. En realidad las islas lingüísticas son productos antrópicos, que *per definitionem* se basan en una desvinculación artificial de las relaciones surgidas de forma natural entre la lengua y el espacio.

El Véneto y muy especialmente el norte de la Toscana son, según nuestro diagnóstico, zonas que en la historia de la lengua italiana tienen que haber jugado el papel de mediadores lingüísticos debido a un fenómeno de importación y exportación lingüística masivo, el cual ha alterado de forma radical su incorporación inicial al espacio de Italia.

De hecho existen puntos de referencia históricos que nos muestran que, especialmente la ciudad de Venecia y sus alrededores, han desarrollado una relación muy estrecha con la Toscana y la Italia media en general, que es totalmente atípica para el resto del norte de Italia. Hay que precisar, sin embargo, que todavía no se sabe con exactitud cuándo sucedió esto. Se puede tratar de un fenómeno acaecido en tiempos de la romanización del Véneto en el siglo II antes de nuestra era. Se puede tratar también, por otra parte, de un fenómeno que se remonta al siglo XIV, cuando la ciudad de Venecia empezó a conquistar su *terraferma*, es decir, el Véneto, y asimilarla lingüísticamente.[30]

Se podría decir que el sur de Italia, con Sicilia y la mayor parte de Cerdeña, se encuentra en una situación de armonía "natural" entre lengua y espacio, mientras que la dinámica evolutiva del desarrollo lingüístico diacrónico se concentra en ambas laderas de los Apeninos del Norte.

Véanse los mapas 29 y 30.

Se puede diferenciar este análisis de nuevo según la fonética y el léxico. En este caso se dan resultados parcialmente diferentes. Mientras que el perfil de la izquierda (mapa 29) se corresponde ampliamente con aquel que ya conocemos, el perfil de la derecha (mapa 30), sobre todo en la zona noroccidental de Italia, tiene claramente otro trazado. Pero el mensaje global permanece invariable: a un sur armónico y quieto desde el punto de vista evolutivo, se contrapone un norte agitado y turbulento desde el mismo punto de vista.

La dialectometría correlativa puede ser empleada también para comparar entre sí categorías intra-lingüísticas como el léxico, la fonética o el vocalismo y el consonantismo. Se puede comprobar así de qué manera estos componentes lingüísticos convergen o divergen en la "gestión dialectal del espacio por parte de los hablantes".

Véanse los mapas 31 y 32.

Nuevamente se dan aquí perfiles claramente estructurados, que deben de ser interpretados de forma conveniente y comparados con la información ya disponible sobre la historia de la lengua italiana y de sus dialectos. Mientras que en el mapa 31 (a la izquierda) aparece un gran dinamismo en el norte y el este de la Toscana, éste aparece en el mapa 32 (a la derecha) de nuevo en el Véneto, Liguria, la franja de transición entre Roma y Ancona, y en algunas islas lingüísticas del sur de Italia.

---

[30] Véanse, en este sentido, las respectivas discusiones en Goebl (2008: 58-61).

De hecho la ya mencionada "centralización" del Véneto parece haber sucedido sobre todo en el ámbito del vocalismo, que así se ha alejado del consonantismo por lo que se refiere a su típica gestión del espacio. Estas conclusiones se deducen de la ubicación de las zonas de color azul oscuro en el mapa 32.

9. Observaciones finales

Llegado al final de mi exposición termino con la presentación de las siete tesis siguientes.

1. El uso de la "faculté langagière" del *Homo loquens* está estrechamente vinculado a las dimensiones naturales del espacio geográfico.
2. El *Homo loquens* se encuentra, pues, en la necesidad de apropiarse —es decir: de gestionar— lingüísticamente el espacio geográfico.
3. La gestión lingüística del espacio lingüístico por parte del *Homo loquens* se efectúa gracias a todos los aspectos de su "faculté langagière".
4. Los atlas lingüísticos representan instrumentos muy útiles para el estudio empírico de la gestión basilectal del espacio.
5. Como los datos de los atlas lingüísticos son (por definición) datos masivos, es preciso utilizar, para su análisis global, métodos cuantitativos.
6. La dialectometría representa la aplicación, a los datos de cualquier atlas lingüístico, de métodos cuantitativos (estadísticos) y cartográficos (visualizadores) debidamente escogidos. Permite el descubrimiento de estructuras geolingüísticas «escondidas» (o «profundas») e incluso de «leyes diatópicas». Estas leyes representan el equivalente epistemológico de las bien conocidas «leyes fonéticas» descubiertas por los Neogramáticos de Leipzig a finales del siglo XIX.
7. La orientación teórica, empírica y metódica de la dialectometría es claramente interdisciplinar. Disciplinas afines son: la clasificación numérica, la estadística, la cartografía y la geografía cuantitativas, la genética de poblaciones y todas la ciencias humanas ocupadas del estudio del espacio como, por ejemplo, la antropología, la etnografía, la sociología, la ecología, etc.

10. Agradecimientos

Quiero manifestar mi agradecimiento sincero a cuantas personas me han ayudado en la tarea de elaboración tanto de la ponencia oral de Vitoria-Gasteiz como de este artículo:

— Traducción castellana del original alemán: Xavier Casassas (Salzburgo).
— Supervisión de la corrección estilística de la versión final: Ramón de Andrés Díaz (Oviedo/Uviéu).
— Producción de los gráficos: Werner Goebl (Viena) y Slawomir Sobota (Salzburgo).
— Mantenimiento y perfeccionamiento continuo del programa VDM: Edgar Haimerl (Seattle) y Slawomir Sobota (Salzburgo).

11. Referencias bibliográficas

AIS: Jaberg, K. & J. Jakob (eds.), 1928-1940, *Sprach- und Sachatlas Italiens und der Südschweiz*, Ringier, Zofingen, 8 vols. (reimpresión: Kraus, Nendeln, 1971).

ALG: Séguy, J., 1954-1974, *Atlas linguistique et ethnographique de la Gascogne*, CNRS, París, 6 vols.

Aurrekoetxea, G., 1992, «Nafarroako euskara: azterketa dialektometrikoa», *Uztaro* 5, 59-109.

Bauer, R., 2009, *Dialektometrische Einsichten. Sprachklassifikatorische Oberflächenmuster und Tiefenstrukturen im lombardo-venedischen Dialektraum und in der Rätoromania*, Istitut ladin «Micurà de Rü», S. Martin de Tor (Ladinia monografica 01).

Besch, W., 1967, *Sprachlandschaften und Sprachausgleich im 15. Jahrhundert. Studien zur Erforschung der spätmittelalterlichen Schreibdialekte und zur Entstehung der neuhochdeutschen Schriftsprache*, Francke, Munich.

Bock, H. H., 1974, *Automatische Klassifikation. Theoretische und praktische Methoden zur Gruppierung und Strukturierung von Daten (Cluster-Analyse)*, Vandenhoeck & Ruprecht, Gotinga.

Brun-Trigaud, G., Le Berre, Y. & J. Le Dû, 2005, *Lectures de l'Atlas linguistique de la France de Gilliéron et Edmont. Du temps dans l'espace. Essai d'interprétation des cartes de l'Atlas linguistique de la France de Jules Gilliéron et Edmond Edmont augmenté de quelques cartes de l'Atlas linguistique de la Basse-Bretagne de Pierre Le Roux*, Éditions du CTHS, París.

Chandon, J.-L. & S. Pinson, 1981, *Analyse typologique. Théories et applications*, Masson: París, Nueva York, Barcelona, Milán.

Goebl, H., 1971, «Projekt einer sprachstatistischen Auswertung von in Sprachatlanten gespeicherter linguistischer Information mit Hilfe elektronischer Rechenanlagen, *Linguistische Berichte* 14, 60-61.

—, 1975, «Dialektometrie», *Grazer linguistische Studien* 1, 32-38.

—, 1976, «La dialectométrie appliquée à l'ALF (Normandie)», in A. Várvaro (ed.), *Atti del XIV Congresso Internazionale di Linguistica e Filologia Romanza*, Macchiaroli, Nápoles / Benjamins, Ámsterdam, vol. 2, 165-195.

—, 1981, «Éléments d'analyse dialectométrique (avec application à l'AIS)», *Revue de linguistique romane* 45, 349-420.

—, 1983, «Parquet polygonal et treillis triangulaire: les deux versants de la dialectométrie interponctuelle», *Revue de linguistique romane* 47, 353-412.

—, 1984, *Dialektometrische Studien. Anhand italoromanischer, rätoromanischer und galloromanischer Sprachmaterialien aus AIS und ALF*, Niemeyer, Tubinga, 3 vols.

—, 1987, «Points chauds de l'analyse dialectométrique: pondération et visualisation», *Revue de linguistique romane* 51, 63-118.

—, 1992, «Problèmes et méthodes de la dialectométrie actuelle (avec application à l'AIS)», in Euskaltzaindia/Académie de la Langue Basque (ed.), *Nazioarteko Dialektologia Biltzarra. Agiriak/Actes du Congrès International de Dialectologie*, Bilbo/Bilbao, 429-475.

—, 2003, «Regards dialectométriques sur les données de l'Atlas linguistique de la France (ALF): relations quantitatives et structures de profondeur», *Estudis Romànics* 25, 59-120.

—, 2005, «La dialectométrie corrélative: un nouvel outil pour l'étude de l'aménagement dialectal de l'espace par l'homme», *Revue de linguistique romane* 69, 321-367.

—, 2008, «La dialettometrizzazione integrale dell'AIS. Presentazione dei primi risultati», *Revue de linguistique romane* 72, 25-113.

Haag, C., 1898, *Die Mundarten des oberen Neckar- und Donautales (Schwäbisch-alemannisches Grenzgebiet: Baarmundarten)*, Hutzler, Reutlingen.

Hall, R. A. Jr., 1943, «The Papal States in Italian Linguistic History», *Language* 19, 125-140.

Jaberg, K., 1906, «Zum Atlas linguistique de la France», *Zeitschrift für romanische Philologie* 30, 512.

—, 1908, *Sprachgeographie. Beitrag zum Verständnis des Atlas linguistique de la France*, Aarau: Sauerländer (versión española: *Geografía lingüística. Ensayo de interpretación del «Atlas lingüístico de Francia»*, traducción de A. Llorente y M. Alvar, Universidad de Granada; Secretariado de Publicaciones, Granada, 1959).

— & J. Jud, 1928, *Der Sprachatlas als Forschungsinstrument. Kritische Grundlegung und Einführung in den Sprach- und Sachatlas Italiens und der Südschweiz*, Niemeyer, Halle (reimpresión: Kraus, Nendeln, 1973).

— & —, 1987, *Atlante linguistico ed etnografico dell'Italia e della Svizzera meridionale, vol. I: L'atlante linguistico come strumento di ricerca. Fondamenti critici e introduzione*, edición italiana por G. Sanga y S. Baggio, Unicopli, Milán.

Lalanne, T., 1953, «Indice de polyonymie. Indice de polyphonie», *Français moderne* 21, 263-274.

Okabe, A., Boots, B. & K. Sugihara, 1992, *Spatial Tesselations. Concepts and Applications of Voronoi Diagrams*, Wiley, Chichester/Nueva York/Brisbane/Toronto/Singapore.

Polanco Roig, L. B., 1984, «Llengua o dialecte: solucions teòriques i aplicació al cas català», in *Actes du XVII[ème] Congrès International de Linguistique et Philologie Romanes (Aix-en-Provence, 29 août-3 septembre 1983)*, Université de Provence, Aix-en-Provence / Jeanne Laffitte, Marsella, vol. 5: 13-30.

Pop, S., 1950, *La dialectologie. Aperçu historique et méthodes d'enquêtes*, Duculot, Gembloux / chez l'auteur, Lovaina, 2 vols.

Ravier, X., 1976, «Jean Séguy et la traversée du langage gascon. Réflexions sur une topogenèse géolinguistique», *Revue de linguistique romane* 40, 389-402.

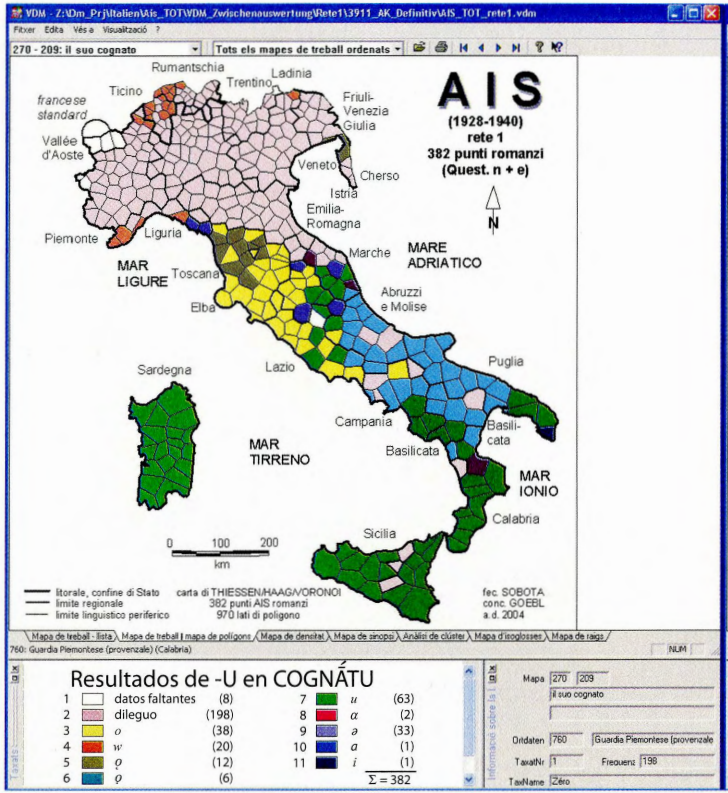
Rohlfs, G., 1947, «Sprachgeographische Streifzüge durch Italien». [con 4 figuras y 29 mapas lingüísticos], Munich, in *Sitzungsberichte* [Memorias] der *Bayerischen Akademie der Wissenschaften, philosophisch-historische Klasse, Jahrgang 1944/46, Heft 3*, 1-67.

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

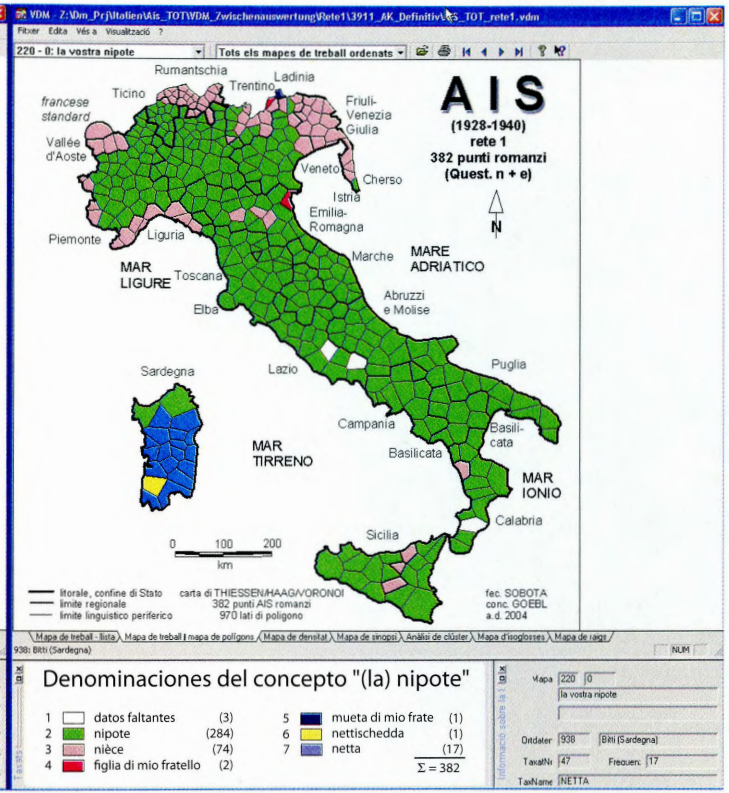
Sousa Fernández, X., 2006, «Aproximación á análise dialectométrica da variedades xeolingüísticas galegas: un estudo comparativo», in M. C. Rolão Bernardo & H. Mateus Montenegro (ed.), *Actas do I Encontro de Estudos Dialectológicos*, Instituto Cultural de Ponta Delgada, Ponta Delgada, 345-362.

Veny, J., 2007-2009, *Petit atles lingüístic del domini català*, Institut d'Estudis Catalans, Barcelona.

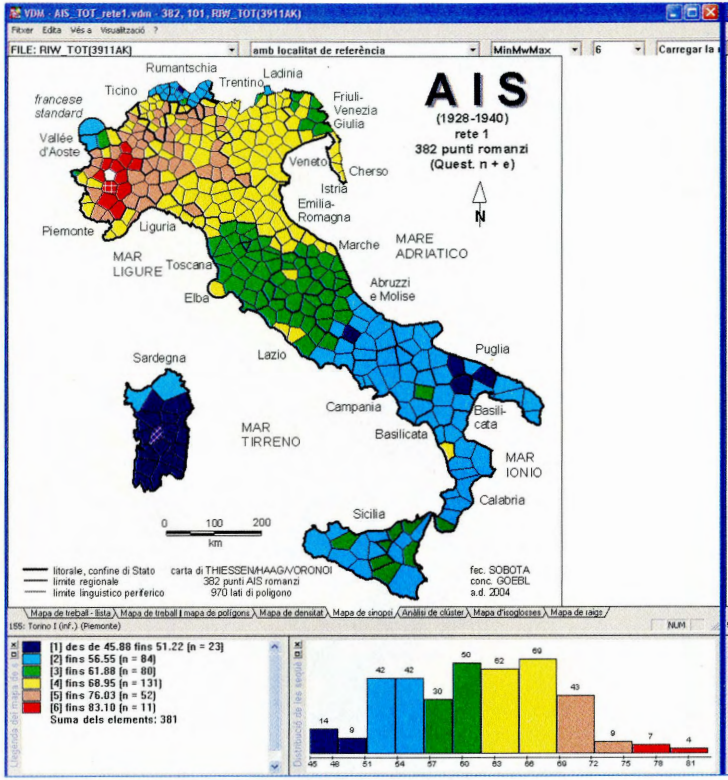
Videsott, P., 2009, *Padania scrittologica: analisi scrittologiche e scrittometriche di testi in italïano settentrionale antico dalle origini al 1525*, Niemeyer, Tubinga.



Mapa 1. Mapa de trabajo fonético: distribución geográfica de los resultados italo-, sardo- y retorrománicos del nexo -U en COGNÁTU latino («cuñado») (según AIS 27 il suo cognato)

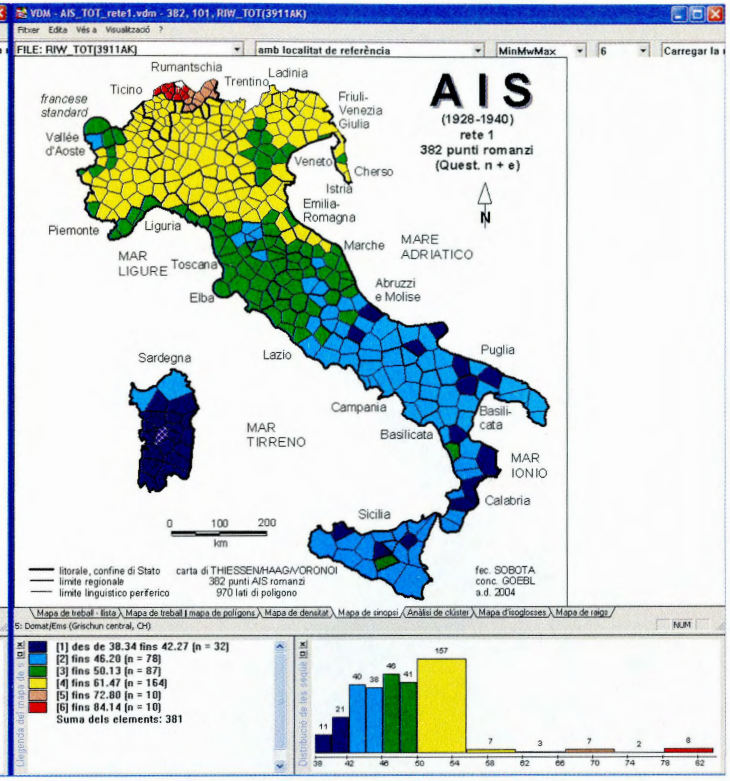


Mapa 2. Mapa de trabajo léxico: distribución geográfica de las denominaciones italo-, sardo- y retorrománicas del concepto «sobrina» (según AIS 22 la vostra nipote)



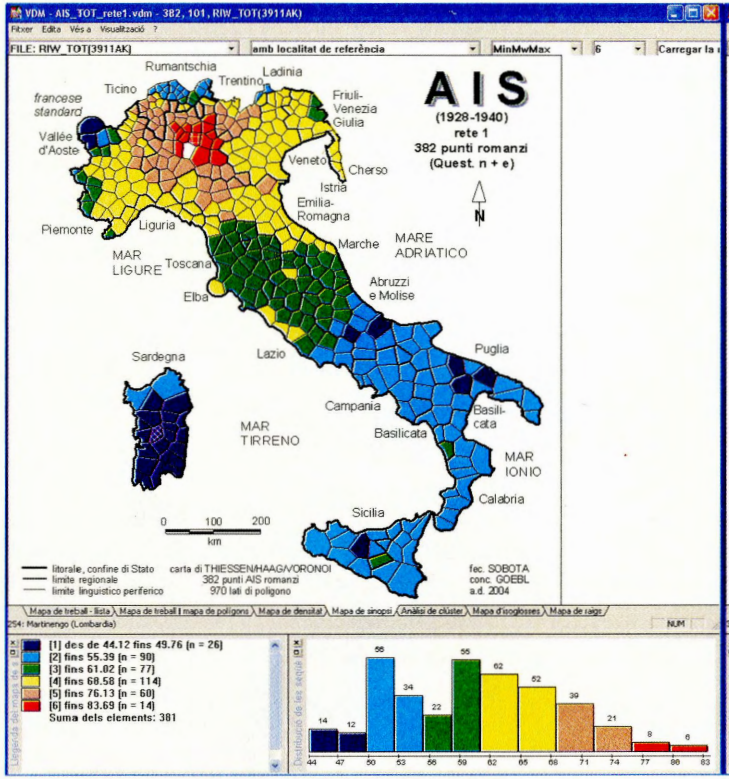
Mapa 3. Mapa de similitud relativo al P.-AIS 155 Turín (Piamonte)

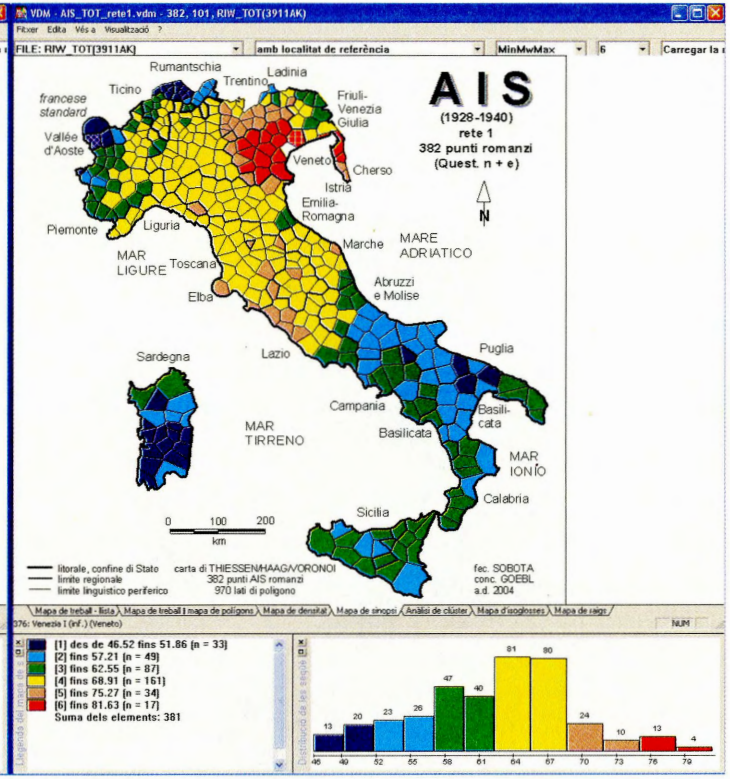
Algoritmo de visualización: MINMWMAX 6-tuplo
Medición de similitud: IRI_{jk}
Corpus: corpus integral (3.911 mapas de trabajo)



Mapa 4. Mapa de similitud relativo al P.-AIS 5 Domat/Ems (Grisones, CH)

Algoritmo de visualización: MINMWMAX 6-tuplo
Medición de similitud: IRI_{jk}
Corpus: corpus integral (3.911 mapas de trabajo)



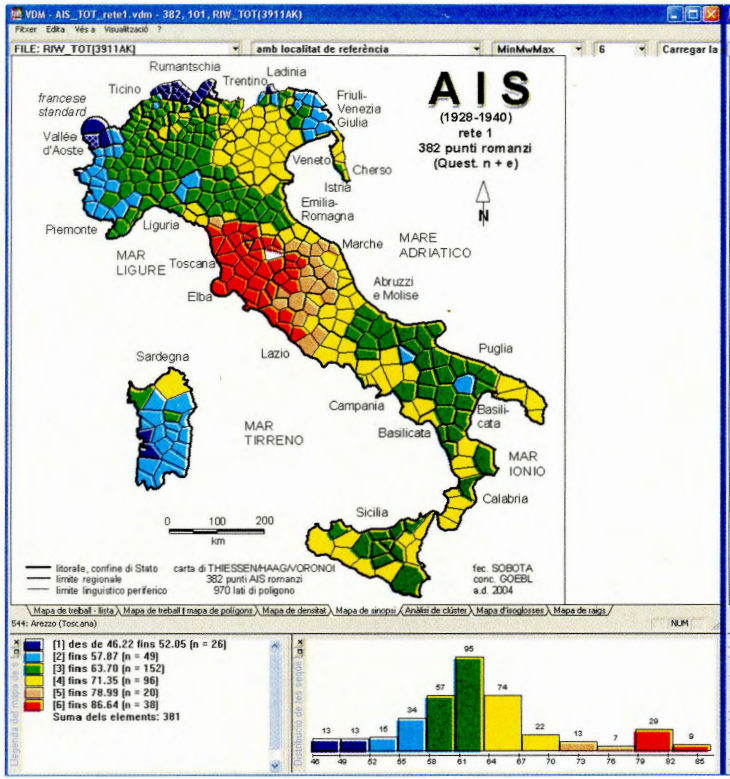


Mapa 5 Mapa de similitud relativo al P.-AIS 254 Martinengo (Lombardía)

Algoritmo de visualización: MINMWMAX 6-tuplo
Medición de similitud: IRI_{jk}
Corpus: corpus integral (3.911 mapas de trabajo)

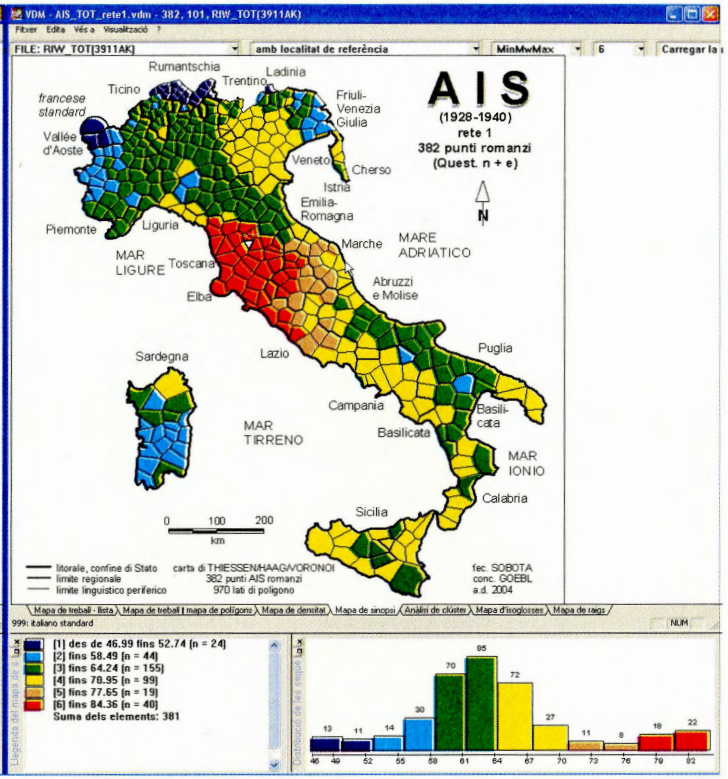
Mapa 6. Mapa de similitud relativo al P.-AIS 376 Venecia

Algoritmo de visualización: MINMWMAX 6-tuplo
Medición de similitud: IRI_{jk}
Corpus: corpus integral (3.911 mapas de trabajo)



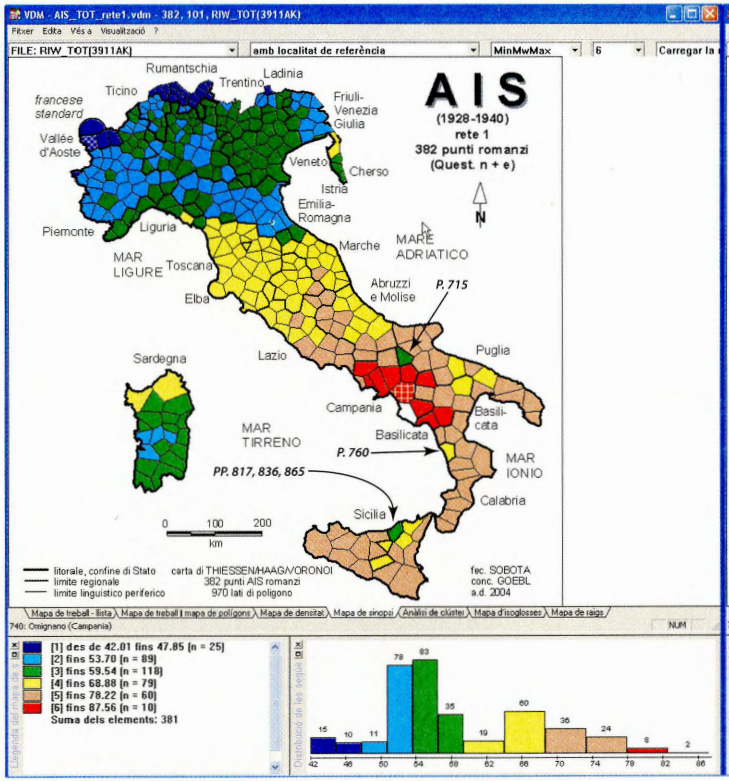
Mapa 7. Mapa de similitud relativo al P.-AIS 544 Arezzo

Algoritmo de visualización: MINMWMAX 6-tuplo
Medición de similitud: IRI_{jk}
Corpus: corpus integral (3.911 mapas de trabajo)



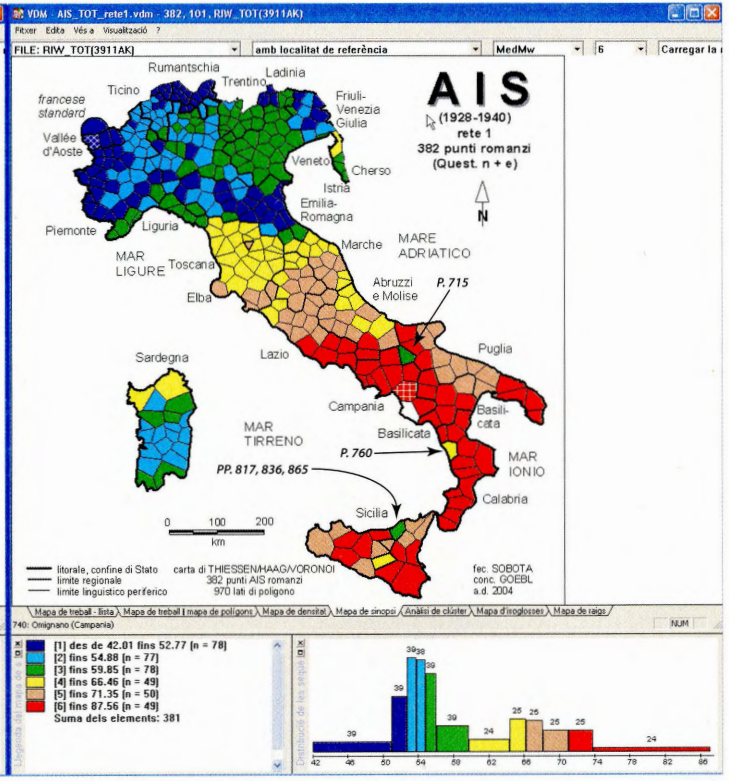
Mapa 8. Mapa de similitud relativo al P.-AIS 999 (italiano estándar)

Algoritmo de visualización: MINMWMAX 6-tuplo
Medición de similitud: IRI_{jk}
Corpus: corpus integral (3.911 mapas de trabajo)



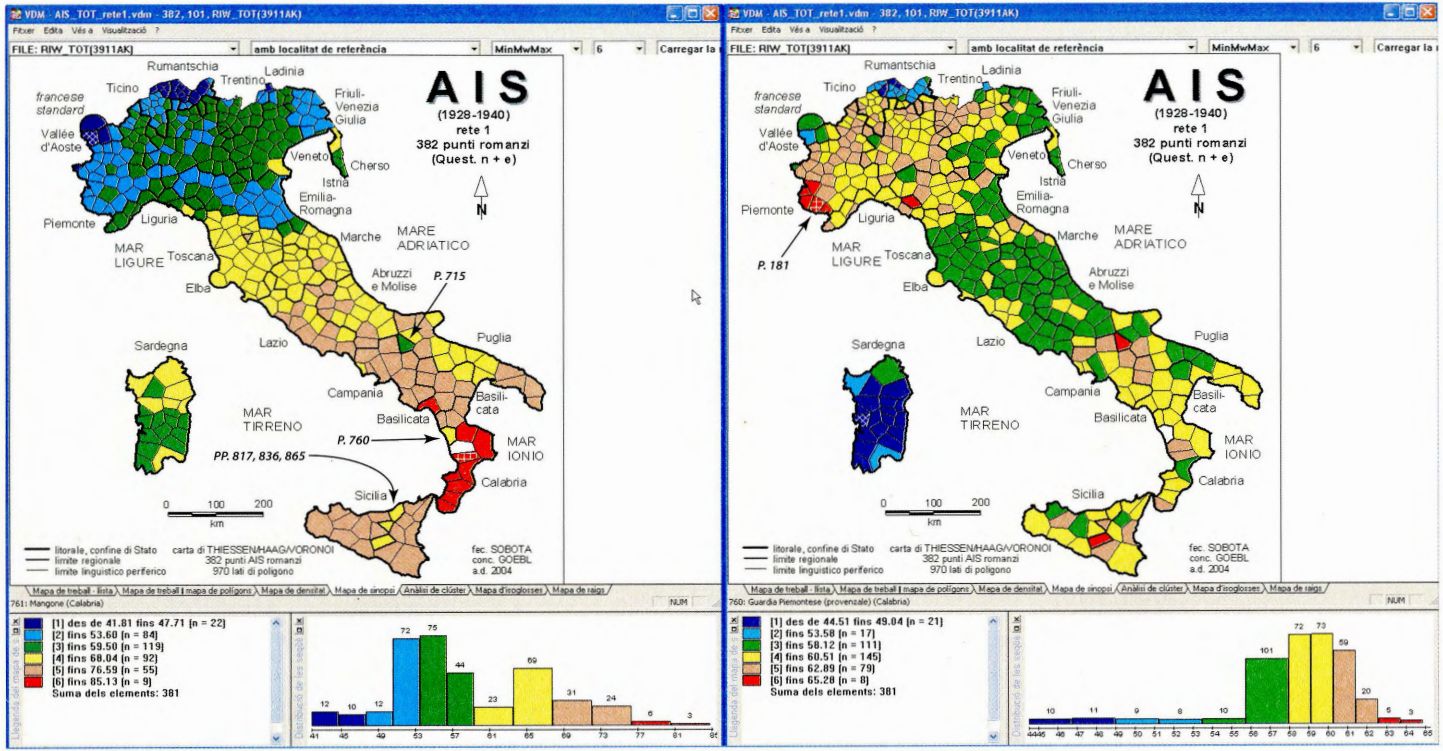
Mapa 9. Mapa de similitud relativo al P.-AIS 740 Omignano (Campania)

Algoritmo de visualización: MINMWMAX 6-tuplo
Medición de similitud: IRI_{jk}
Corpus: corpus integral (3.911 mapas de trabajo)



Mapa 10. Mapa de similitud relativo al P.-AIS 740 Omignano (Campania)

Algoritmo de visualización: MEDMW 6-tuplo
Medición de similitud: IRI_{jk}
Corpus: corpus integral (3.911 mapas de trabajo)

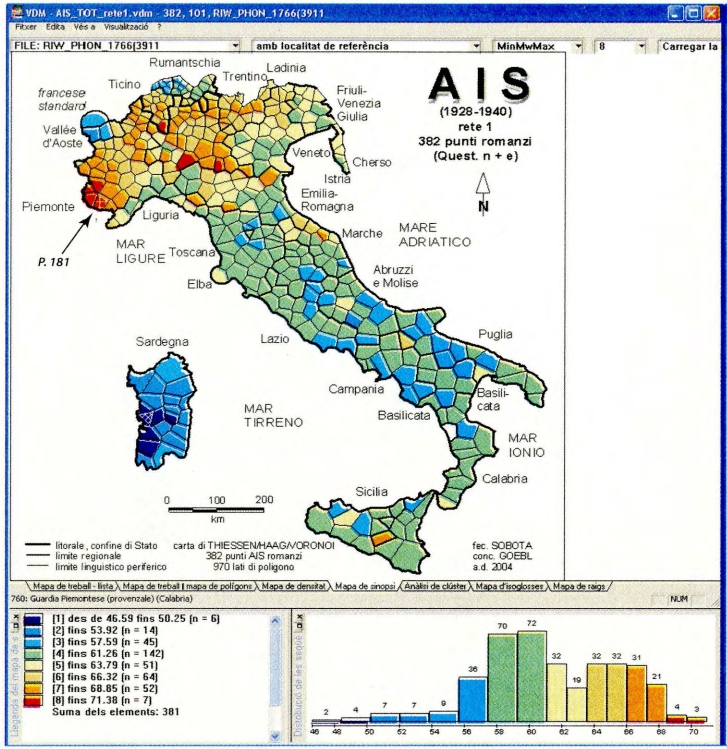


Mapa 11. Mapa de similitud relativo al P.-AIS 761 Mangone (Calabria)

Algoritmo de visualización: MINMWMAX 6-tuplo
Medición de similitud: IRI_{jk}
Corpus: corpus integral (3.911 mapas de trabajo)

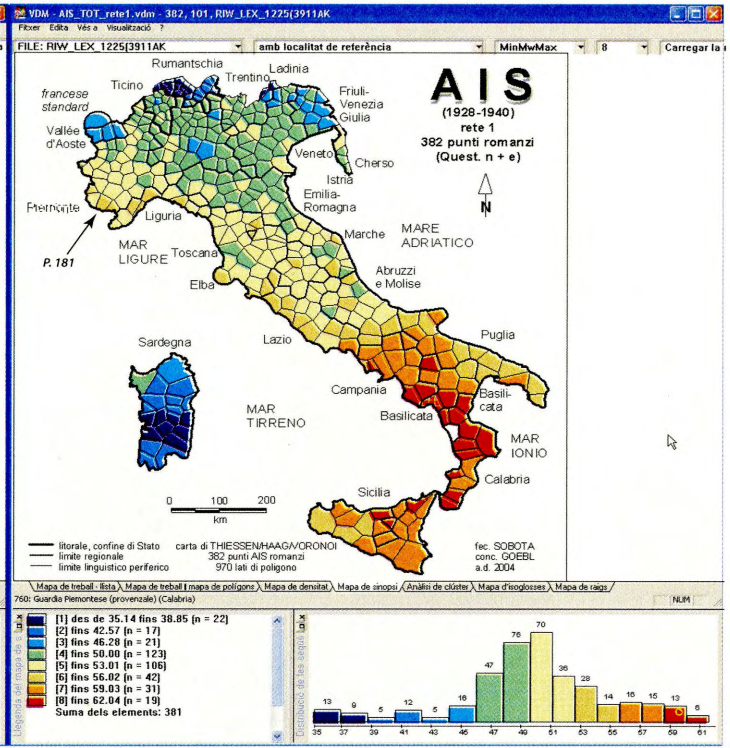
Mapa 12. Mapa de similitud relativo al P.-AIS 760 Guardia Piemontese (Calabria)

Algoritmo de visualización: MINMWMAX 6-tuplo
Medición de similitud: IRI_{jk}
Corpus: corpus integral (3.911 mapas de trabajo)



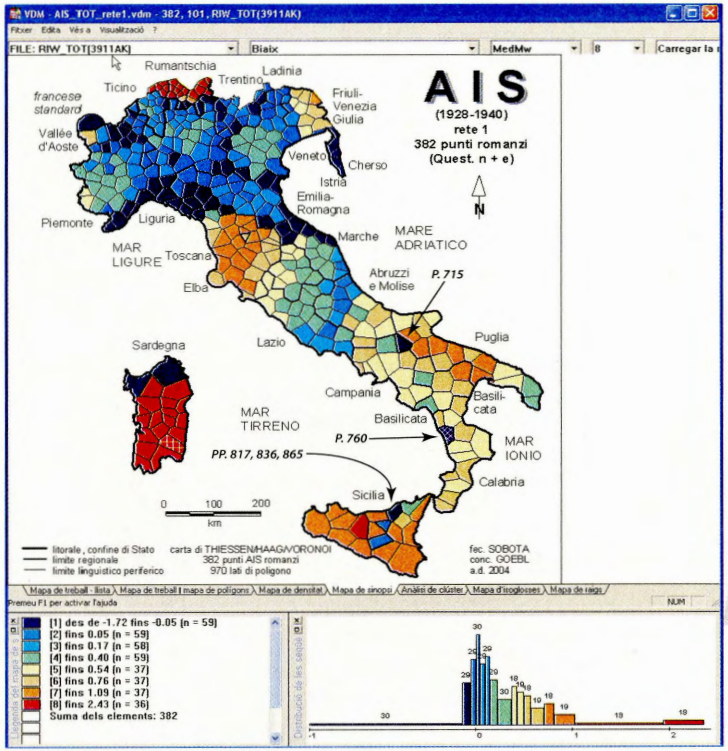
Mapa 13. Mapa de similitud relativo al P.-AIS 760 Guardia Piemontese (Calabria)

Algoritmo de visualización: MINMWMAX 6-tuplo
Medición de similitud: IRI_{jk}
Corpus: subcorpus fonético (1.766 mapas de trabajo)



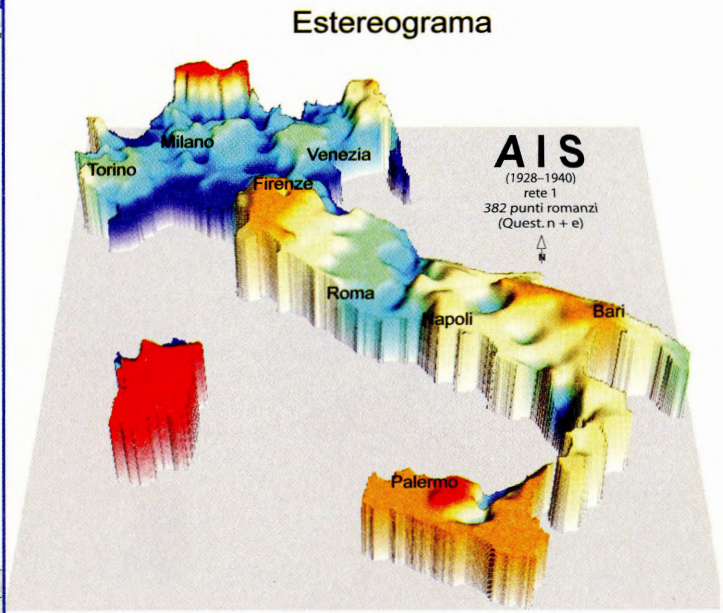
Mapa 14. Mapa de similitud relativo al P.-AIS 760 Guardia Piemontese (Calabria)

Algoritmo de visualización: MINMWMAX 6-tuplo
Medición de similitud: IRI_{jk}
Corpus: subcorpus léxico (1.225 mapas de trabajo)



Mapa 15. Mapa de parámetros: sinopsis de 382 coeficientes de asimetría de Fisher (CAF)

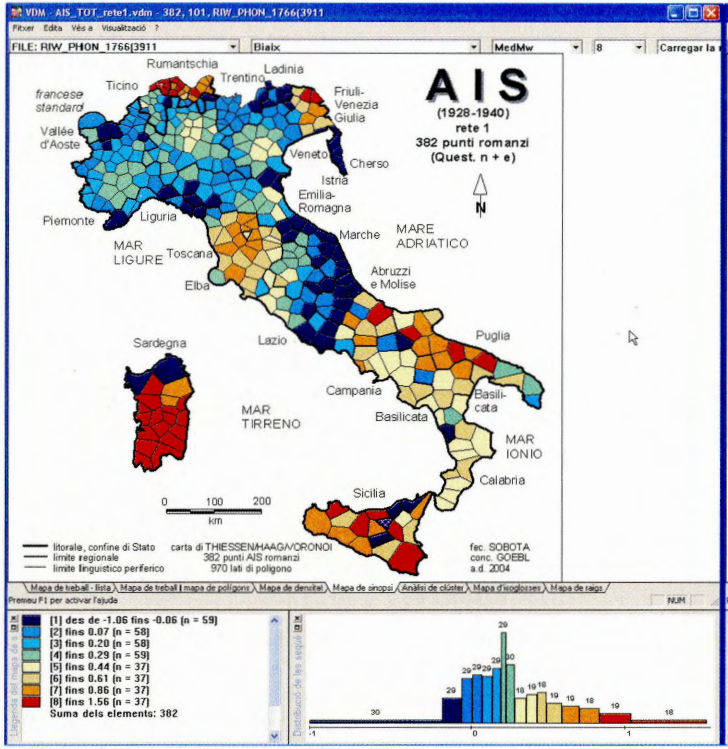
Algoritmo de visualización: MEDMW 6-tuplo
Medición de similitud: IRI_{jk}
Corpus: corpus integral (3.911 mapas de trabajo)



Punto de observación: ***desde sur***
Inclinación: ***45°***

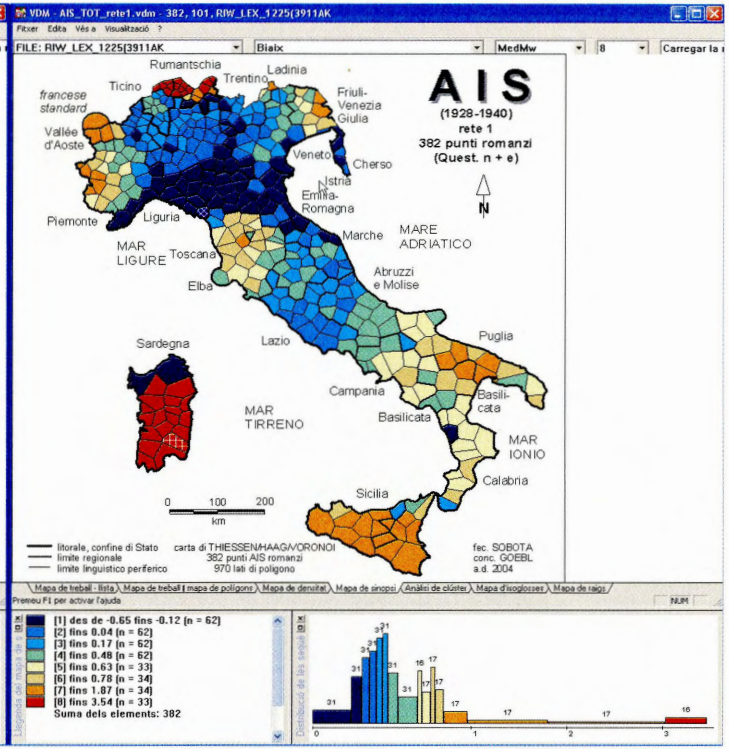
Mapa 16. Estereograma de la sinopsis de 382 coeficientes de asimetría de Fisher (CAF)

Punto de observación: desde el sur
Inclinación: 45°



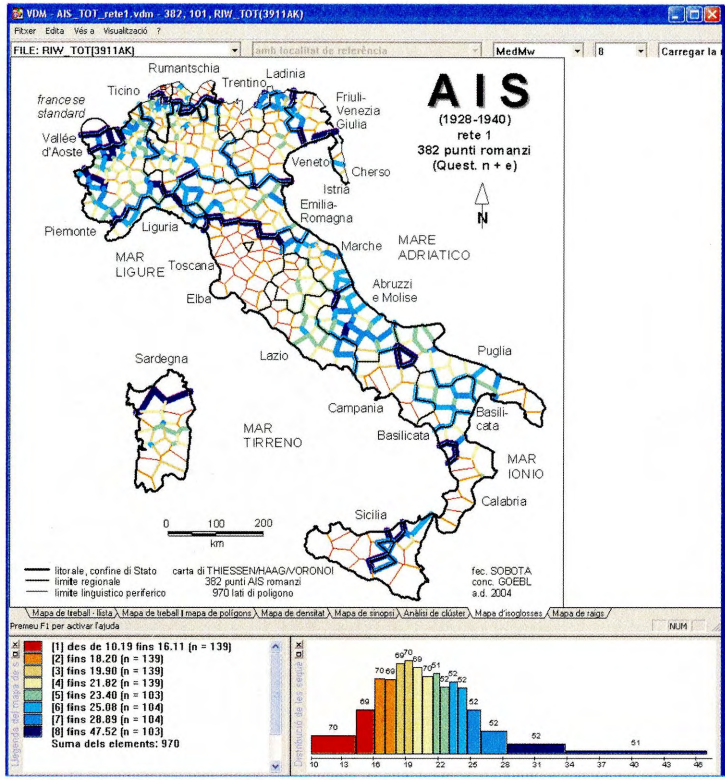
Mapa 17. Mapa de parámetros: sinopsis de 382 coeficientes de asimetría de Fisher (CAF)

Algoritmo de visualización: MEDMW 6-tuplo
Medición de similitud: IRI_{jk}
Corpus: subcorpus fonético (1.766 mapas de trabajo)



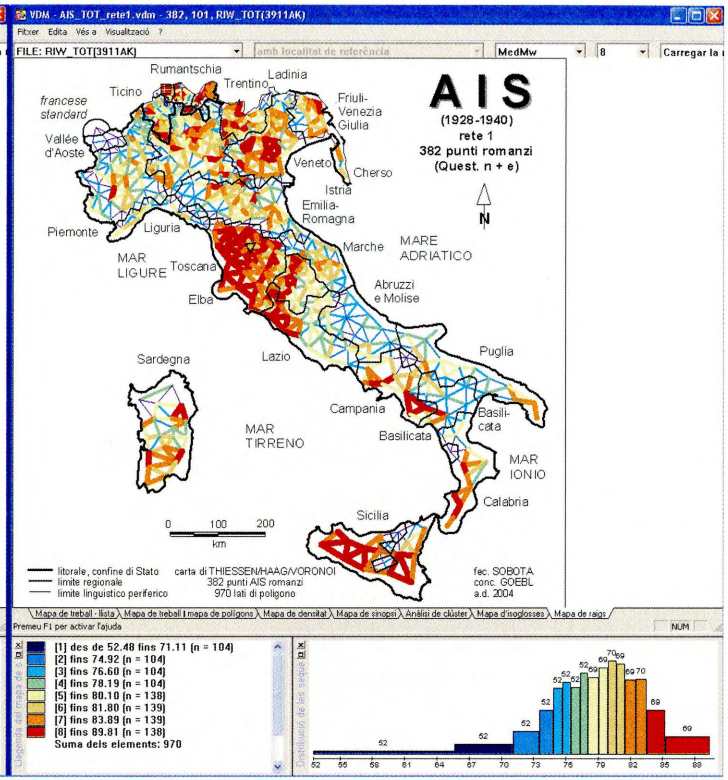
Mapa 18. Mapa de parámetros: sinopsis de 382 coeficientes de asimetría de Fisher (CAF)

Algoritmo de visualización: MEDMW 6-tuplo
Medición de similitud: IRI_{jk}
Corpus: subcorpus léxico (1.225 mapas de trabajo)



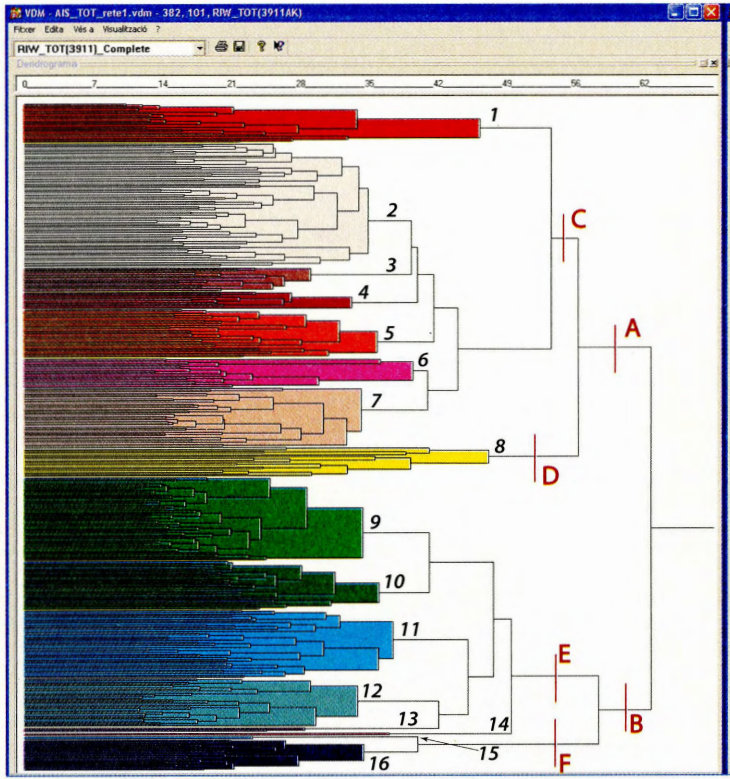
Mapa 19. Mapa interpuntual A: mapa isoglótico

Algoritmo de visualización: MEDMW 6-tuplo
Medición de similitud: IRD_{jk}
Corpus: corpus integral (3.911 mapas de trabajo)



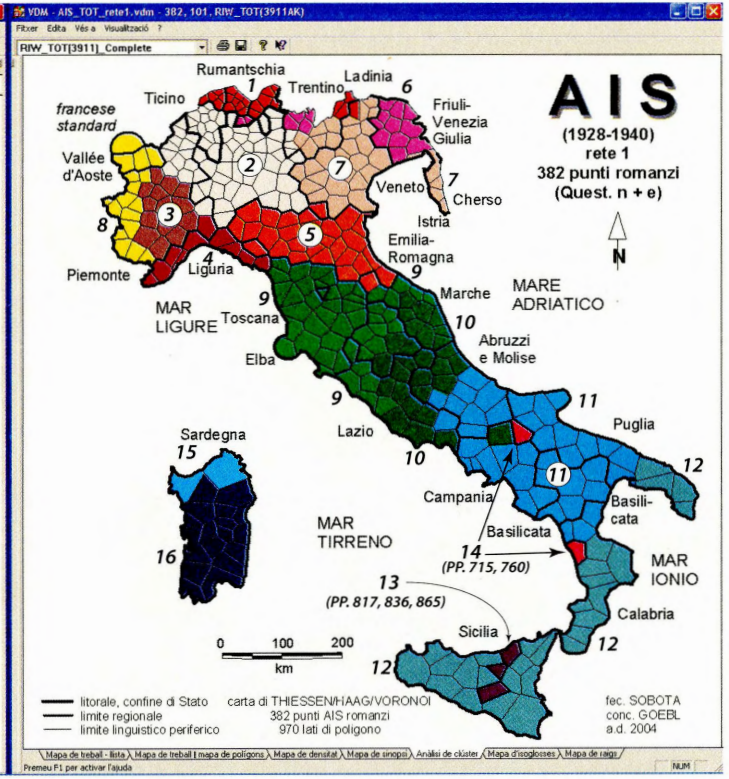
Mapa 20. Mapa interpuntual B: mapa de rayos

Algoritmo de visualización: MEDMW 6-tuplo
Medición de similitud: IRI_{jk}
Corpus: corpus integral (3.911 mapas de trabajo)

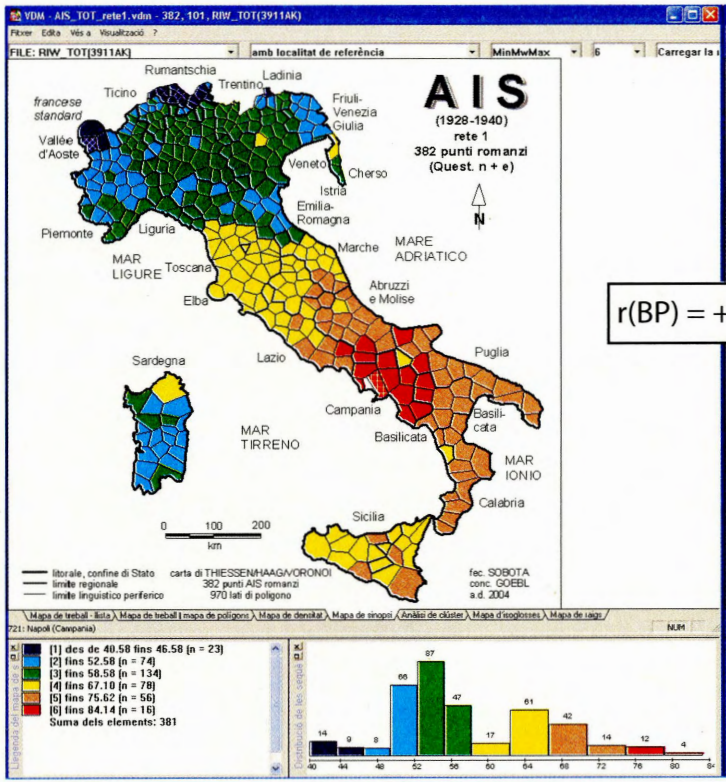


Mapa 21. Clasificación jerárquica aglomerativa: árbol genealógico

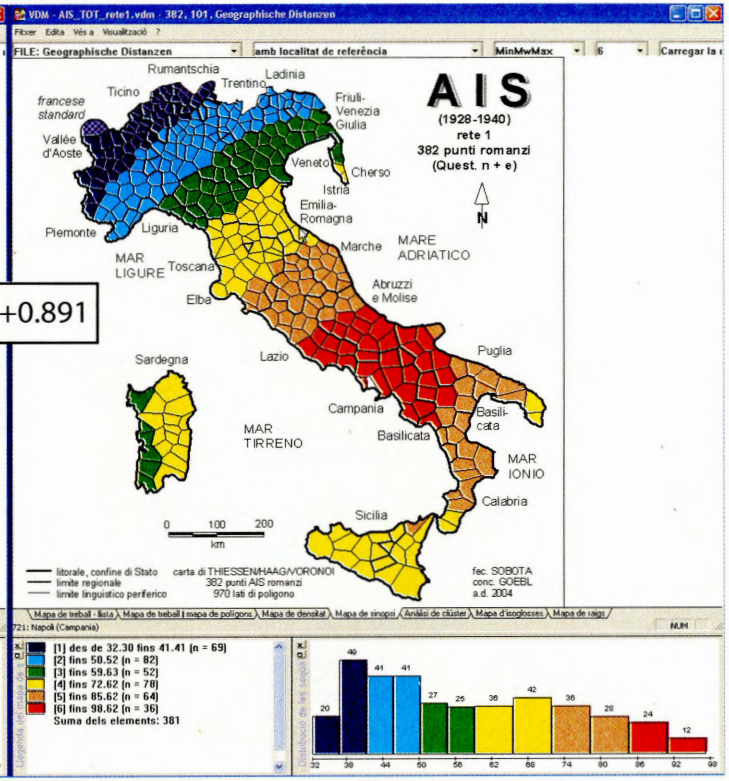
Algoritmo de clasificación: Complete Linkage
Medición de similitud: IRI_{jk}
Corpus: corpus integral (3.911 mapas de trabajo)



Mapa 22. Clasificación jerárquica aglomerativa: espacialización del mapa 21



r(BP) = +0.891

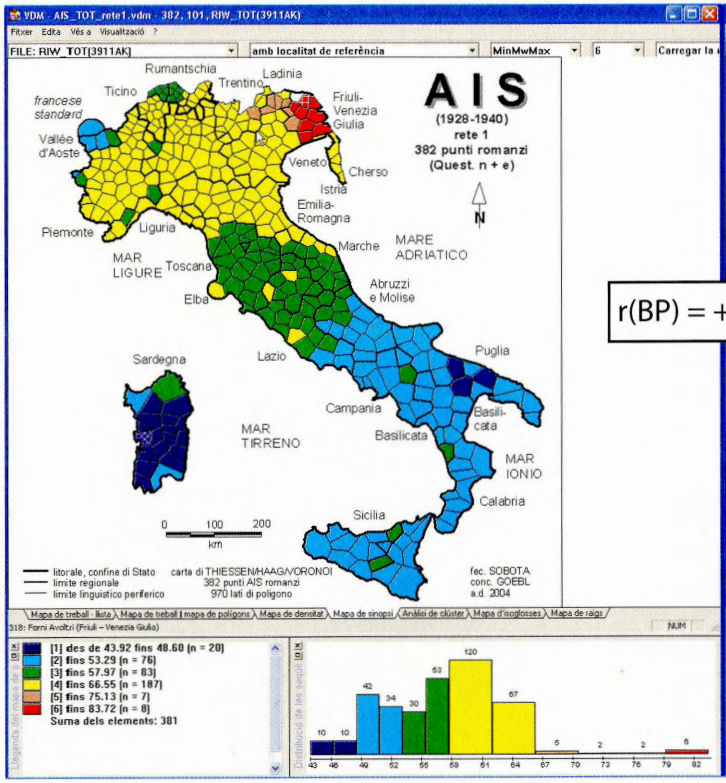


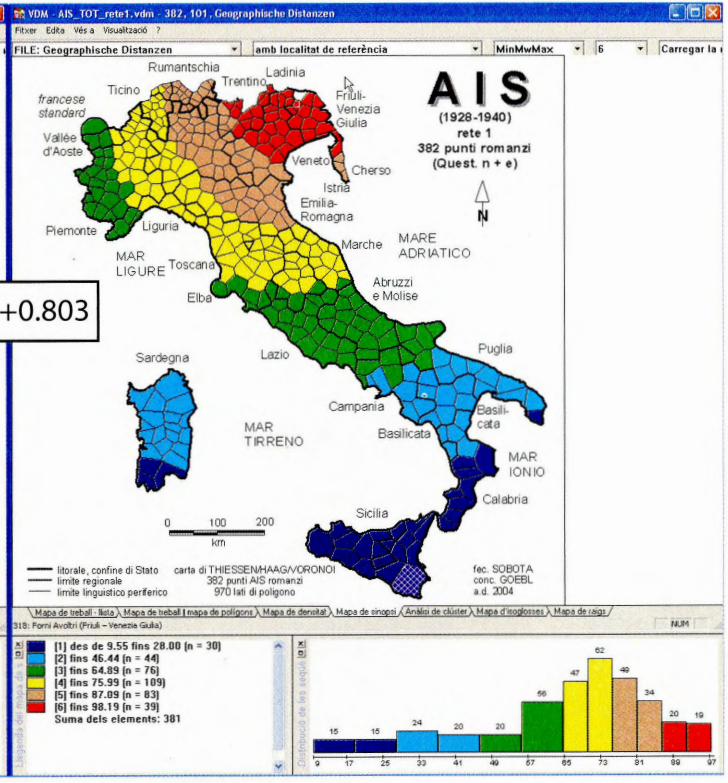
Mapa 23. Mapa de similitud relativo al P.-AIS 721 Nápoles (Campania)

Algoritmo de visualización: MINMWMAX 6-tuplo
Medición de similitud: IRI_{jk}
Corpus: corpus integral (3.911 mapas de trabajo)

Mapa 24. Mapa de proximidad relativo al P.-AIS 721 Nápoles (Campania)

Algoritmo de visualización: MINMWMAX 6-tuplo
Medición de proximidad: teorema de Pitágoras





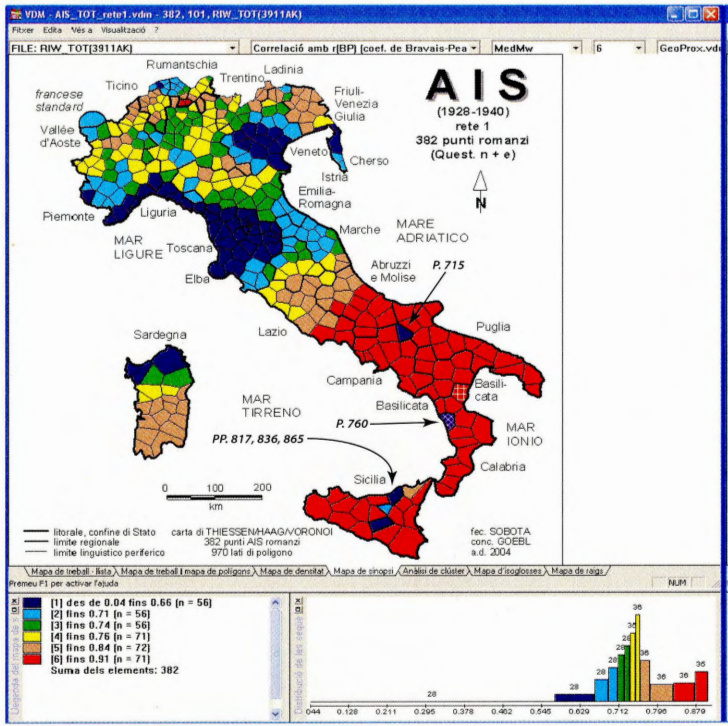
r(BP) = +0.803

Mapa 25. Mapa de similitud relativo al P.-AIS 318 Forni Avoltri (Friul)

Algoritmo de visualización: MINMWMAX 6-tuplo
Medición de similitud: IRI_{jk}
Corpus: corpus integral (3.911 mapas de trabajo)

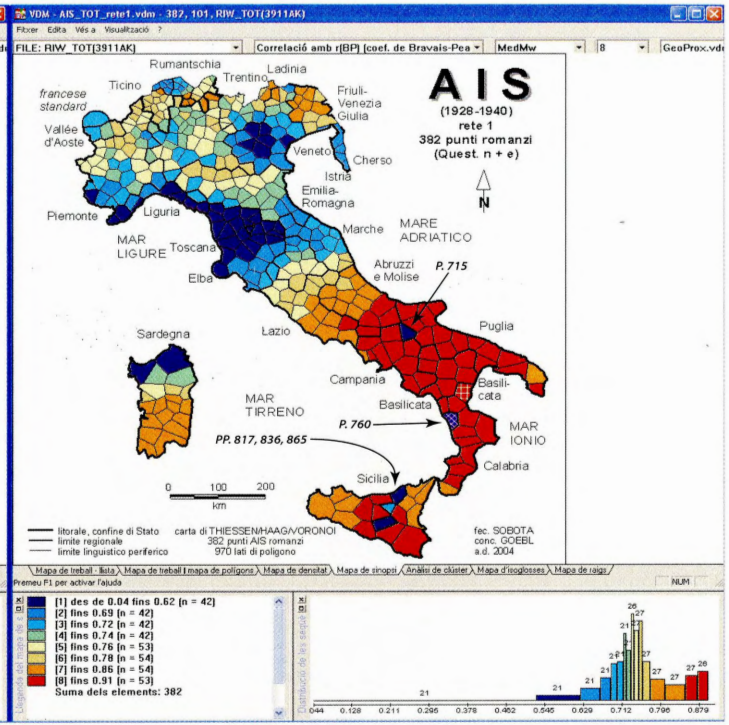
Mapa 26. Mapa de proximidad relativo al P.-AIS 318 Forni Avoltri (Friul)

Algoritmo de visualización: MINMWMAX 6-tuplo
Medición de proximidad: teorema de Pitágoras



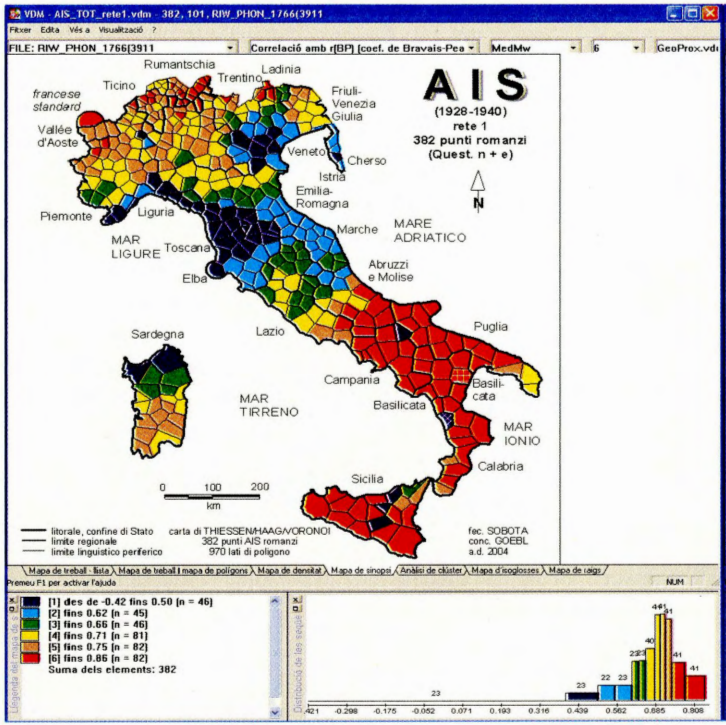
Mapa 27. Mapa de correlaciones: similitudes lingüísticas y proximidades geográficas

Algoritmo de visualización: MEDMW 6-tuplo
Medición de similitud: IRI_{jk}
Corpus: corpus integral (3.911 mapas de trabajo)
Medición de proximidad: teorema de Pitágoras



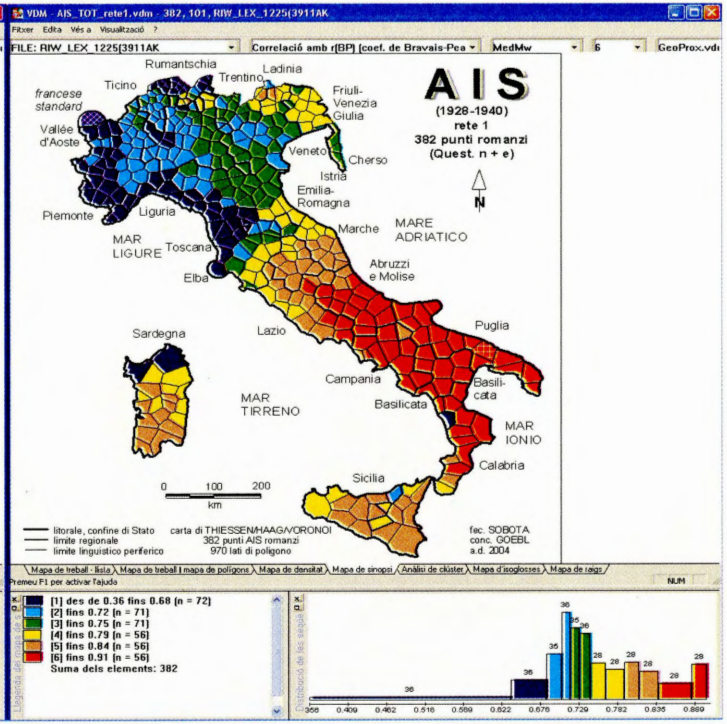
Mapa 28. Mapa de correlaciones: similitudes lingüísticas y proximidades geográficas

Algoritmo de visualización: MEDMW 8-tuplo
Medición de similitud: IRI_{jk}
Corpus: corpus integral (3.911 mapas de trabajo)
Medición de proximidad: teorema de Pitágoras



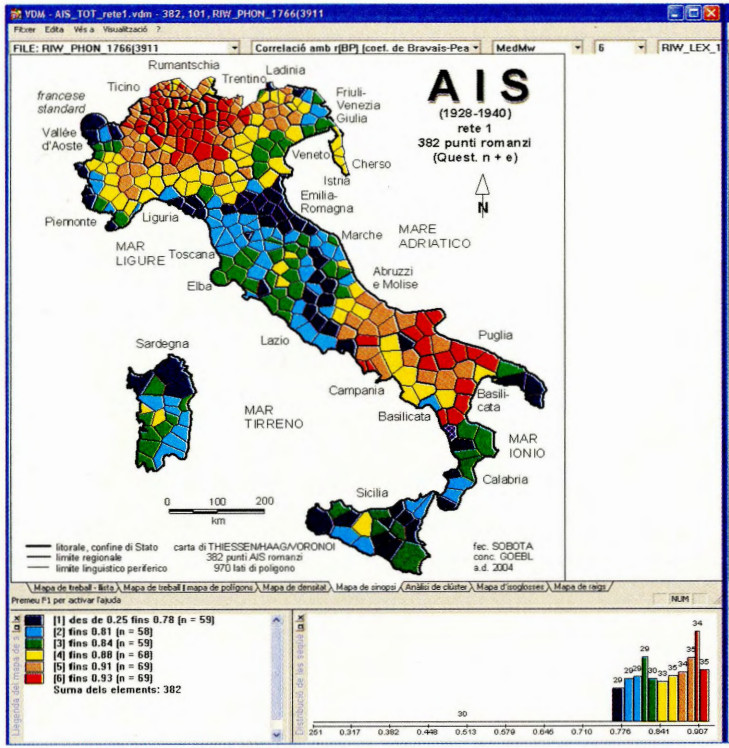
Mapa 29. Mapa de correlaciones: similitudes fonéticas y proximidades geográficas

Algoritmo de visualización: MEDMW 6-tuplo
Medición de similitud: IRI_{jk}
Corpus: subcorpus fonético (1.766 mapas de trabajo)
Medición de proximidad: teorema de Pitágoras



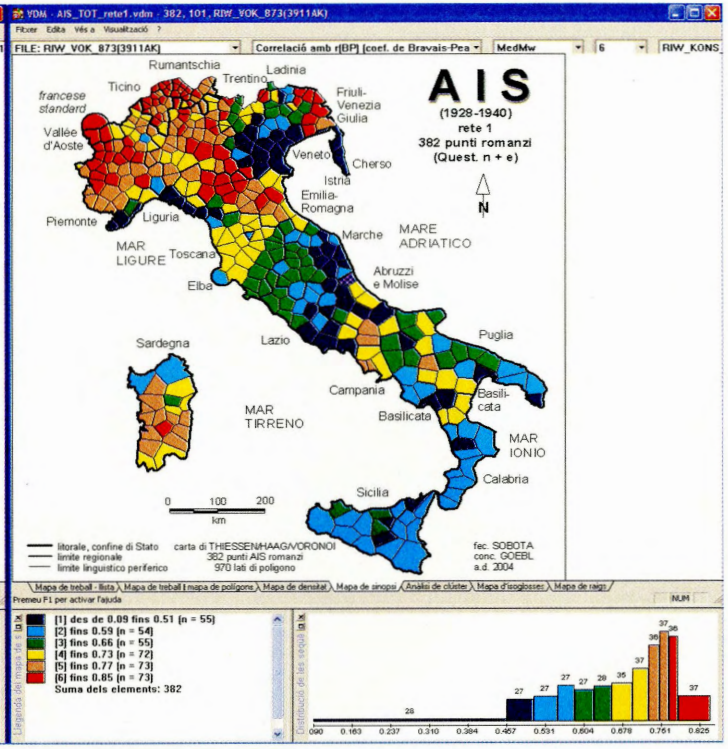
Mapa 30. Mapa de correlaciones: similitudes léxicas y proximidades geográficas

Algoritmo de visualización: MEDMW 6-tuplo
Medición de similitud: IRI_{jk}
Corpus: subcorpus léxico (1.225 mapas de trabajo)
Medición de proximidad: teorema de Pitágoras



Mapa 31. Mapa de correlaciones: similitudes fonéticas y léxicas

Algoritmo de visualización: MEDMW 6-tuplo
Medición de similitud: IRI_{jk}
Subcorpora lingüísticos:
fonética: 1.766 mapas de trabajo
léxico: 1.225 mapas de trabajo



Mapa 32. Mapa de correlaciones: similitudes vocálicas y consonánticas

Algoritmo de visualización: MEDMW 6-tuplo
Medición de similitud: IRI_{jk}
Subcorpora lingüísticos:
vocalismo: 873 mapas de trabajo
consonantismo: 805 mapas de trabajo

SOME FURTHER DIALECTOMETRICAL STEPS

John Nerbonne, Jelena Prokić, Martijn Wieling & Charlotte Gooskens

Center for Language and Cognition
University of Groningen

Abstract

This article surveys recent developments furthering dialectometric research which the authors have been involved in, in particular techniques for measuring large numbers of pronunciations (in phonetic transcription) of comparable words at various sites. Edit distance (also known as Levenshtein distance) has been deployed for this purpose, for which refinements and analytic techniques continue to be developed. The focus here is on (i) an empirical approach, using an information-theoretical measure of mutual information, for deriving the appropriate segment distances to serve within measures of sequence distance; (ii) a heuristic technique for simultaneously aligning large sets of comparable pronunciations, a necessary step in applying phylogenetic analysis to sound segment data; (iii) spectral clustering, a technique borrowed from bio-informatics, for identifying the (linguistic) features responsible for (dialect) divisions among sites; (iv) techniques for studying the (mutual) comprehensibility of closely related varieties; and (v) Séguy's law, or the generality of sub-linear diffusion of aggregate linguistic variation.

Keywords: *Phonetic alignment, multi-alignment, spectral clustering, mutual comprehensibility, linguistic diffusion*

1. Introduction

The dialectometric enterprise (Goebl 1982) need not be introduced in a paper in this volume, which is largely dedicated to presenting its contemporary form. We shall only take care to note points at which there may not yet be consensus. The authors of the present contribution view the introduction of exact techniques into dialectology as an extension of the *methods* available within the discipline but do not suggest that dialectology change in its central questions, in particular, questions on the nature of the influence that geography has on language variation. More exact techniques, and especially computational techniques, serve to broaden the empirical base that dialectology can effectively build on, improve the replicability of data analysis techniques, and enable more abstract questions to be addressed with empirical rigor. Nerbonne (2009) elaborates on these opportunities for dialectometry.

We have collaborated especially on developing and applying measures of pronunciation distance derived from edit distance or Levenshtein distance. The distance be-

tween two transcriptions t_1 and t_2 is defined as the sum of costs associated with the least costly set of operations needed to transform t_1 into t_2, and typically one makes use of only three operations, substitution, insertion and deletion. A by-product of the calculation is an alignment of the two transcriptions, where the segments which have been involved in the operations are written one above the other:

[æ ə f t ə n ʉ n] 'afternoon', Georgia (LAMSAS)
[æ f t ə r n u n] 'afternoon', Pennsylvania

In the example here, we see that schwa [ə] corresponds with ø (the null segment), [ø] with [r], and [ʉ] with [u]. The correspondences are extracted automatically from the digitized transcriptions. See Nerbonne & Heeringa (2009) and references there for more extensive explanation and illustration. The technique was developed with a view to analyzing the sort of data typically found in dialect atlases, in particular where the pronunciation of a single word or phrase is elicited, recorded and ultimately transcribed for later analysis. The "Salzburg school" of dialectometry typically devotes a good deal of time to manually extracting correspondences from dialect atlases (a phase Goebl refers to as *Taxierung*, roughly 'appraisal'), a step in methodology which the application of edit-distance largely obviates. So this is a point at which we feel there has been a contribution to dialectometric technique. Nerbonne and Heeringa (2009) reviews work devoted to pronunciation difference measurement.

We see an additional point in favor of the deployment of edit distance, namely that it provides a broader view of the variation in typical atlas data because it incorporates *entire* pronunciations in its measurements instead of relying on the analyst's choice of variables to extract. This means that dialectometrists using edit-distance measures of pronunciation are less likely to fall prey to choosing their variables in a way that biases results. Since we, too, normally deal with dialect atlas data, we are not shielded from the danger of biased data collection completely, but whereas other approaches typically focus on the set of variables the atlas was designed to assay, the edit-distance measurements incorporate all the pronunciations, not merely a one or two segments per word.

In addition we note that pronunciation distance is a true metric, obeying the relevant mathematical axioms. This means that the distances assayed are all non-negative, zero only in case of identity, and that they satisfy the so-called triangle inequality: for all t_1, t_2, there is no t' such that:

$$d(t_1,t') + d(t',t_2) < d(t_1,t_2)$$

The fact that genuine measurements are made instead of predications of identity vs. non-identity has a consequence that less data is required for reliable assessment of the relations among sites in a dialect landscape. Heeringa (2004) shows that about 30 transcriptions per site yields consistent measurements for Dutch (Cronbach's $\alpha \geq 0.8$). One typically needs 100 or more items to attain comparable levels of consistency when comparing at a categorical level. The consistency measure is empirical, and therefore must be recalculated on each new data set, but Heeringa's result has been confirmed in several other data collections. We typically work with sets of 100 words per variety, because we are not merely interested in the distance relations

at an aggregate level, but also in the concrete segmental correspondences which the analysis also yields. These may not turn up in small data sets.

Although we share the general interest in finding groups of most similar sites in language areas, we are also sensitive to the "French" worry that dialect areas may lead ephemeral existences. Following Goebl and others, we have used clustering regularly as a means of seeking groups in our data. We have also been sensitive to the charges of the statistical community that no clustering technique works perfectly (Kleinberg 2003), and have therefore explored versions of clustering that remove the instability inherent in the technique (i.e. the problem that very small differences in input data may lead to major differences in output clusterings), using both bootstrap clustering and "noisy" clustering, a technique we developed independently (Nerbonne et al. 2008, Prokić & Nerbonne 2008). As the last reference demonstrates, there is good reason to be wary of cluster analyses even when they are accompanied by stability-enhancing augmentations.

Stimulated by the difficulties of finding groups reliably using clustering, we have emphasized the use of multi-dimensional scaling (MDS) in analyzing the results of dialectometric measurements (Nerbonne, Heeringa and Kleiweg 1999). Black (1973) introduced MDS to linguistics, where it has been applied sporadically since. It is a remarkable fact that the very complicated data of linguist variation, potentially varying in hundreds of dimensions (one for each point of comparison) may normally be faithfully rendered in just two or three dimensions. But this remarkable fact means that it is possible to visualize language variation effectively in scatterplots of two dimensions that make use of a single color dimension, e.g., grey tones to plot the third. Nerbonne (to appear, b) reviews approaches to mapping aggregate linguistic variation, focusing especially on techniques involving MDS.

The remainder of this paper sketches very recent work building on the lines set out above. Section 2 aims to answer a question that we have been wrestling with since the very earliest work applying edit distance to variation data (Nerbonne & Heeringa 1997), namely how to weight operations in such a way that phonetically natural substitutions (and also insertions and deletions) cost less than unnatural ones. Focusing on substitutions, we should prefer to see that the substitution of [i] for [e] should cost less than the substitution of [i] for [a]. Section 3 reports on multi-alignment, the attempt to extend the process of aligning two strings to that of aligning many, potential hundreds. This is potentially very interesting in historical linguistics, where regular sound correspondences play an important role. Section 4 summarizes work on a new effort to identify not only the important groups of varieties, but also —and simultaneously— the linguistic basis of the group. Section 5 reports on efforts to link the work we have done on pronunciation distance to the important issue of the comprehensibility of varieties, effectively asking whether phonetically distant varieties are also less comprehensible. Section 6 attempts to use the dialectometric perspective to view language variation from a more abstract perspective, and asks whether this might allow the formulation of more general laws of linguistic variation. Finally, the concluding section attempts to identify areas which are promising for dialectometric inquiry in the near future.

2. Inducing Segment Distances Empirically

Many studies in dialectometry based on the Levenshtein distance use very simple segment distances, only distinguishing vowels and consonants; then substitution costs for each pair of vowels (or pair of consonants) are the same (e.g. see Wieling et al. 2007). In such studies, the substitution of [e] for [i] has the same weight as a substitution of [e] for any other vowel. While it would clearly be rewarding to use linguistically sensible segment distances, obtaining these is difficult and seems to require some arbitrary decisions. If this seems surprising given the relatively simple charts of phonetic symbols found e.g. in the IPA Handbook, one might consider that dialect atlases employ hundreds of symbols (LAMSAS distinguishes over 1,100 different vowels). A complete segment distance table thus requires tens of thousands of specifications, minimally (and in the case of LAMSAS over 500,000).

In his thesis Heeringa (2004) calculated segment distances using two different approaches. In Chapter 3 he represented every phone as a bundle of features where every feature is a certain phonetic property. In Chapter 4 Heeringa measured the transcription distances using acoustic segment distances calculated from the recordings in Wells and House (1995). This is less arbitrary than the feature representation since it is based on physical measures, but both of these approaches have their disadvantages. The former relies on the selection of a feature system, while the latter requires acoustic data to be available.

In order to avoid these problems, Wieling et al. (2009) proposed using pointwise mutual information (PMI) to automatically induce segment distances from phonetic transcriptions. PMI is a measure of association between two events x and y:

$$PMI(x,y) = log_2(P(x,y) / P(x)P(y))$$

where the numerator $P(x,y)$ tells us how often we have observed the two events together, while the denominator $P(x)P(y)$ tells us how often we would expect these two events to occur together if we assumed that their occurrence were statistically independent. The ratio between these two values shows us if two events co-occur together more often than just by chance. Wieling et al. (2009) use it to automatically learn the distances between the phones in aligned word transcriptions and also to improve the automatically generated alignments. Applied to aligned transcriptions, $P(x,y)$ represents the relative frequency of two segments being aligned together, while $P(x)$ and $P(y)$ are relative frequencies of segments x and y.

The procedure of calculating the segment distances and improving the alignments is iterative and consists of the following steps: a) align all word transcription using the Levenshtein algorithm b) from the obtained alignments calculate the PMI distances between the phones c) align all word transcriptions once more using the Levenshtein algorithm, but based on the generated segment distances d) repeat the previous two steps until there are no changes in segment distances and alignments.

Wieling et al. (2009) evaluated the alignments based on the PMI procedure on a manually corrected gold standard set of alignments of Bulgarian dialect data. The results indicated that both at the segment level and at the word level the novel algorithm was a clear improvement over the Levenshtein algorithm with hand-coded segment distances.

Qualitative error analysis has shown that both versions of the Levenshtein algorithm make most errors due to the restriction that vowels and consonants cannot be aligned. Apart from this error, the simple Levenshtein algorithm is not able to distinguish between aligning a vowel with one of the two neighboring vowels, since in the simple version of the algorithm all vowels are equally distant from each other. This also holds for the consonants. Using PMI-induced segment distances solves this problem since the algorithm learns that the distance between [n] and [ŋ] is smaller than the distance between [n] and [k] (see Figure 1). Correction of these types of errors is where the PMI procedure improves the performance of the simple Levenshtein algorithm and generates more correct alignments.

v	ˈɤ	-	n	-		v	ˈɤ	n	-	-
v	ˈɤ	ŋ	k	ə		v	ˈɤ	ŋ	k	ə

Figure 1

Erroneous alignment produced by the simple Levenshtein algorithm (left) and the correct alignment produced by Levenshtein PMI algorithm (right)

The alignments produced using the segment distances arrived at via the pointwise mutual information procedure improves the alignment accuracy of the Levenshtein algorithm, and consequently enables us to obtain better distances between each pair of sites calculated from the transcriptions. The next step would be to improve this procedure and enable the algorithm to calculate the distances between vowels and consonants. In that way, the quality of the alignments could be further improved by minimizing the number of errors caused by the restriction on the alignments between vowels and consonants.

3. Multi-Alignment

While the technique described in the previous section aims at improving pairwise alignments, Prokić et al. (2009) introduced an algorithm that is used to produce multiple sequence alignments. It is an adapted version of the ALPHAMALIG algorithm (Alonso et al. 2004) modified to work with phonetic transcriptions.

Pairwise string alignment methods compare two strings at a time, while in multiple string alignment (MSA) all strings are aligned and compared at the same time. MSA is an especially effective technique for discovering patterns that can be hard to detect when comparing only two strings at once. Both techniques are widely used in bioinformatics for aligning DNA, RNA or protein sequences in order to determine similarities between the sequences. However, as noted in Gusfield (1997), the multiple string alignment method is more powerful. Gusfield calls it "the holy grail" of sequence algorithms in molecular biology.

In recent years there has been increasing interest in using phylogenetic methods to analyze linguistic data, especially language change and variation (Gray and Atkinson 2003; Atkinson et al. 2005; Warnow et al. 2006). This is possible because of the similarities between the evolution of languages and the evolution of species —they

are both passed on from generation to generation accompanied by changes during the process. As they change, both languages and biological populations can split into new subgroups, becoming more and more distant from each other and from common ancestor(s). In order to apply methods from biology directly to pronunciation data, which we are particularly interested in, it is essential to preprocess the data in order to identify all the correspondences. This means that we need to be able to derive multiply aligned strings of phonetic transcriptions. An example of multiply aligned transcriptions can be seen in Figure 2:

village 1:	j	ˈɑ	-	-	-	-
village 2:	j	ˈɑ	z	e	-	-
village 3:	-	ˈɑ	s	-	-	-
village 4:	j	ˈɑ	s	-	-	-
village 5:	j	ˈɑ	z	e	k	a
village 6:	j	ˈɛ	-	-	-	-
village 7:	-	ˈɒ	s	-	-	-

Figure 2

Multiply aligned phonetic transcriptions
of the Bulgarian word *az* 'I' collected at 7 villages

The advantage of multiply aligned strings over pairwise alignments is two-fold: a) it is easier to detect and process sound correspondences (e.g. [ɑ], [ɛ] and [ɒ] are very easy to detect and extract from the second column in Figure 2); b) the distances between strings are more precise if calculated from multiple aligned strings since they preserve information on the sounds that were lost (the last two columns in all transcriptions —except for village 5's transcription— preserve the information that the villages commonly lack the last two sounds, which would be lost in the pairwise alignments).

The automatic alignment of strings was carried out using the ALPHAMALIG algorithm, originally designed for bilingual text alignment. It is an iterative pairwise alignment program that merges multiple alignments of subsets of strings. Although originally developed to align words in texts, it can work with any data that can be represented as a sequence of symbols of a finite alphabet. In Prokić et al. (2009) the algorithm was adapted in order to work with phonetic transcriptions. The distances between the phones were set in such a way that vowels can be aligned only with vowels and consonants only with consonants.

Since there is no widely accepted way to evaluate the quality of multiple alignments Prokić et al. (2009) suggested two new methods for comparing automatically produced alignments against manually aligned strings, the so called *gold standard.* Both methods compare the content of corresponding columns in two alignments. One method, called the *column dependent method,* takes into account the order of columns in two alignments and the content of the columns as well. In other words, it looks for a perfect match. The other method is not sensitive to the order of col-

umns and takes into account only the content of two corresponding columns. It is based on Modified Rand Index (Hubert and Arabie 1985), one of the most popular methods for comparing two different partitions.

The results for the Bulgarian data set show that automatically generated alignments are of a very high quality, scoring between 92% (first method) and 98% (latter method). Error analysis has revealed that most of the alignment errors are due to the restriction that vowels and consonants cannot be aligned. In order to avoid this problem, the algorithm would need information on the distances between the phones, which is not straightforward to obtain (see Section 2). Although both evaluation methods could be improved further, they both estimate alignment quality well.

Studies in historical linguistics and dialectometry where string comparison is used could benefit from tools for multiple sequence alignment by speeding up the process of string aligning and making it suitable to work with large amounts of data.

4. Spectral Graph Clustering

Until recently, almost all aggregate dialectometric analyses have focused on identifying the most important geographical groups in the data. While it is important to identify the areas which are linguistically similar and those which differ, the aggregate approach does not expose the linguistic basis of the groups.

The aggregate approach averages over the distances between pairs of large numbers of aligned words to obtain pairwise dialect distances. After obtaining an MDS map of Dutch dialects, Wieling et al. (2007) correlated the distances based on each individual word in the dataset with all MDS dimensions to find the most characteristic words for each of the three MDS dimensions. While finding the most characteristic word is certainly informative, linguists are also interested in finding the most important sound correspondences. Nerbonne (to appear, a) used factor analysis to identify the linguistic structure underlying the aggregate analysis of southern American dialects, focusing his analysis on vowels and showing that aggregate distances based only on vowels correlated highly with distances based on all sound segments.

Prokić (2007) went a step further and extracted the ten most frequent non-identical sound correspondences from the aligned transcriptions of pairs of Bulgarian dialects and used the relative frequency of each of these sound correspondences to assign a score to each site (each site had multiple scores; one for each sound correspondence). When the pairwise distances were calculated on the basis of these scores and correlated with the aggregate distances, she was able to identify how characteristic each sound correspondence was for the aggregate view.

All the aforementioned methods have in common that the process of determining the linguistic basis is *post-hoc*; the aggregate distances are calculated first and the linguistic basis is determined later. This is less than optimal as we wish to know which features or sound correspondences really serve as the linguistic basis for the site grouping. Consequently, linguists have also not been convinced by these approaches and have been slow to embrace the aggregate approach.

Another approach was taken by Shackleton (2007), who used principal component analysis to group features (instead of sound correspondences) in the pronunciation of English dialects. For each variety the component scores were calculated and

groups of varieties were distinguished based on the presence of the grouped features. Hence, in this case, first the groups of features are determined, after which the geographical groups are identified. Shackleton did not use sound correspondences, but he used self-selected features of both consonants and vowels and also (in a separate experiment) variants determined by English linguists. While this is certainly insightful, there remains a great deal of subjectivity in categorizing and selecting the features.

To counter these drawbacks, Wieling and Nerbonne (2009, to appear) introduced a new method to simultaneously cluster geographic varieties as well as the concomitant sound correspondences (compared to a reference variety). The BIPARTITE SPECTRAL GRAPH PARTITIONING method they applied was first introduced by Dhillon (2001) to co-cluster words and documents and is based on calculating the singular value decomposition (SVD) of a word-by-document matrix. The left and right singular vectors obtained from this procedure are merged and clustered into the desired number of groups using the *k*-means algorithm. A detailed explanation as well as an example is shown in Wieling and Nerbonne (to appear).

The variety-by-sound correspondence matrix of Wieling and Nerbonne (2009, to appear) was based on alignments for 423 Dutch varieties with respect to a reference pronunciation close to standard Dutch using the PMI algorithm discussed in Section 2. All sound correspondences present in the alignments for a variety were counted and in the matrix the presence (frequency of at least 1) or absence of a sound correspondence in a variety was stored. We did not use the frequencies in the matrix as this seemed to have a negative impact on performance, possibly because of the presence of some very high frequencies. In their first study, Wieling and Nerbonne (2009) reported a fair geographical clustering in addition to sensible sound correspondences, based on "eyeballing" the data. In a subsequent study, Wieling and Nerbonne (to appear) developed a method to rank the sound correspondences to identify the most important ones for each cluster based on representativeness (i.e. the proportion of varieties in a cluster containing the sound correspondences) and distinctiveness in a cluster (i.e. the number of varieties within as opposed to outside the cluster containing the sound correspondence). They concluded that their method to rank the sound correspondences conformed to a great extent with the subjectively selected sound correspondences in the previous study (Wieling and Nerbonne 2009). While this method still has some drawbacks (e.g., incorporating frequency information has a negative effect on the results), it is certainly a step forward in identifying the linguistic basis of aggregate dialectometric analyses.

5. Intelligibility of contrasting varieties

Dialectometrical techniques are useful tools for investigating the role that linguistic distances play in the mutual intelligibility among speakers of closely related language varieties. The techniques allow the researcher to test the relationship between intelligibility on the one hand and objective linguistic similarity on the other. It seems likely that the greater the linguistic resemblance is between two languages or dialects, the greater the degree of mutual intelligibility will be. However, only a moderately strong correlation ($r = -0.65$, $p < 0.01$) was found between intelligibil-

ity scores of 17 Scandinavian dialects by Danish listeners and the perceived distances to these dialects from the listeners' own varieties (Beijering, Gooskens and Heeringa 2008). This suggests that perceived distance and intelligibility scores are two different measurements that cannot be equated with each other. In other words, the (dis)similarity of another language variety to one's own, is only a moderately successful predictor of the how intelligible this variety is.

Methods for testing and measuring the effect of linguistic distance are becoming increasingly sophisticated. Methods include web-based experiments and computational techniques. Intelligibility can be measured by asking listeners to answer open or closed questions about the content of a spoken text or by having subjects translate a spoken text or word lists. By means of open questions about a text, the mutual intelligibility of three Scandinavian standard languages (Danish, Norwegian and Swedish) and three West-Germanic languages (Afrikaans, Dutch and Frisian) were tested in Gooskens (2007). The percentage of correctly answered questions per listener-language combination (e.g. Danes listening to Swedish) was correlated with the corresponding phonetic distances measured with the Levenshtein algorithm. There was a negative correlation of –0.64 ($p < 0.01$) between linguistic distance and intelligibility when all data were included but a stronger correlation ($r = -0.80$, $p < 0.01$) when only the Scandinavian data were included.

In another study (Beijering, Gooskens and Heeringa 2008), the intelligibility of 17 Scandinavian dialects by speakers of Standard Danish was tested using a translation task. The percentage of correctly translated words in a short story (per language variety) was correlated with phonetic distances from Standard Danish to each of the 17 dialects. Previous applications of the Levenshtein algorithm typically employed a word length normalization, which means that the total number of operations (insertions, deletions and substitutions) is divided by the number of alignment slots for a word pair. The effect of normalization is that a pronunciation difference counts more in a short word than in a long word. In our investigation, we correlated the intelligibility scores with normalized as well as non-normalized Levenshtein distances. The results showed higher correlations than in the previous study ($r = -0.86$, $p < 0.01$ for the normalized and $r = -0.79$, $p < 0.01$ for the non-normalized distances). The phonetic distances between each of the 17 dialects and Standard Danish were correlated with perceptual distances as judged by the listeners on a 10-point scale. The normalized Levenshtein distances correlated more strongly with the intelligibility scores than with perceived distances, and the difference was significant ($r = -0.86$ versus $r = 0.52$, $p < 0.05$). The non-normalized Levenshtein distances showed the same tendency, but the difference between normalized and non-normalized distances was not significant ($r = -0.79$, $p < 0.01$ versus $r = -0.62$, $p < 0.01$ respectively). These results suggest that Levenshtein distance is a good predictor of both intelligibility and perceived linguistic distances. However, the algorithm seems to be a better predictor of intelligibility than of perceived linguistic distances. Word length normalization is important when modeling intelligibility since segmental differences in short words presumably have a larger impact on intelligibility than segmental differences in long words. On the other hand, perceived distance is likely to be dependent on the total number of deviant sounds regardless of word length and therefore correlations are higher with the non-normalized distances.

Mutual intelligibility among the Scandinavian languages is fairly high, comparable to the mutual intelligibility in many dialect situations. Several investigations have been carried out in order test how well Scandinavians understand each other (e.g. Maurud 1976; Börestam 1987; Delsing & Lundin Åkesson 2005; Gooskens, Van Heuven, Van Bezooijen and Pacilly accepted). Results repeatedly show asymmetric intelligibility scores between Scandinavian language pairs. Especially Swedes have more difficulties understanding Danes than vice versa. The techniques for distance measurements used for the investigations discussed above cannot capture this asymmetry. A conditional entropy measure has therefore been developed as a measure of remoteness to model asymmetric intelligibility (Moberg, Gooskens, Nerbonne and Vaillette 2007). In the conditional entropy measure semantically corresponding cognate words are taken from frequency lists and aligned. The conditional entropy of the phoneme mapping in aligned word pairs is calculated. This approach aims to measure the difficulty of predicting a phoneme in a native language given a corresponding phoneme in the foreign language. The results show that a difference in entropy can be found between language pairs in the direction that previous intelligibility tests predict. Conditional entropy as a measure of remoteness thus seems a promising candidate for modeling asymmetric intelligibility.

In the investigations mentioned above, intelligibility was measured on the basis of whole texts and aggregate phonetic distance measures were applied. As a consequence, no conclusions could be drawn about the nature of the phonetic differences that contribute most to intelligibility. However, it is desirable to gain more detailed knowledge about which role various linguistic factors play in the intelligibility of closely related languages. Gooskens, Beijering and Heeringa (2008) reanalyzed the data from the intelligibility tests with 17 Scandinavian dialects (see above), now measuring the consonant and the vowel distances separately. A higher correlation was found between intelligibility and consonant distances ($r = -0.74$, $p < 0.01$) than between intelligibility and vowel distances ($r = -0.29$, $p < 0.05$), which confirms the claim that consonants play a relatively important role for intelligibility of a closely related language.

Kürschner, Gooskens and Van Bezooijen (2008) focused on the comprehension of 384 isolated spoken Swedish words among Danes, and examined a wide variety of potential linguistic predictors of intelligibility, including the similarity of the foreign word's pronunciation to one's own variety's pronunciation. The strongest correlation was found between word intelligibility and phonetic distances ($r = -0.27$, $p < 0.01$). This is rather low in comparison with the correlations found for the intelligibility of whole texts with aggregate phonetic distances. Including more linguistic factors like word length, foreign sounds, neighborhood density and word frequency in a logistic regression analysis improved the predictive power of the model. However, a large amount of variation still remains to be explained. We attribute this to the high number of idiosyncrasies of single words compared with the aggregate intelligibility and linguistic distance used in earlier studies. While at the aggregate level we are well able to predict mutual intelligibility between closely related language varieties, it is a challenge for future work to develop phonetic distances that are better able to express the communicative distances between closely related languages at the word level.

6. Séguy's law

In the paper which launched the dialectometric enterprise, Jean Séguy (1971) observed that the measure of aggregate lexical distance which he defined in the same paper increased directly with geographic distance, but in a sub-linear fashion. Not long after Séguy's paper, Trudgill (1974) advocated that "gravity" laws of the sort then popular in the social sciences might underlie the dynamics of linguistic diffusion. He explicitly advocated that dialectology seek more general accounts of linguistic variation. But where Séguy had measured a sublinear effect of geography on linguistic diversity, Trudgill postulated an attractive force which weakened with the square of linguistic distance. Trudgill clearly investigated geography as a means of studying the effect of social contact, a point at which we, and we suspect nearly all dialectologists, heartily agree. This can be seen in the fact that Trudgill also predicted that population size would play a role in strengthening the tendency of two sites to adopt each other's speech habits. Nerbonne & Heeringa (2007) explicitly contrasted the two views of geography, arguing their incompatibility based on a study of 52 locations in the northeast Netherlands, and showing that aggregate Dutch dialect distance followed the a logarithmic curve not unlike the sublinear curve that Séguy had used to model his Gascony data.

Nerbonne (to appear, c) examines five more dialect areas, namely Forest Bantu (Gabon), Bulgaria, Germany, Norway, and the U.S. Eastern seaboard, and shows that the distribution of aggregate linguistic distance in each of these is a logarithmic function of geography, just as the Netherlands and Gascony. Figure 3 shows the six curves.

Six areas of linguistic variation that display the same sub-linear relation between geographic distance and aggregate linguistic distance, first noted by Séguy (1971). Logarithmic curves are drawn. Because different measuring schemes were applied, the *y*-axes are not comparable.

The paper goes on to examine the relation between the dynamic influencing individual linguistic variables and the curve representing aggregate variation, showing that the sub-linear aggregate curve is indeed incompatible with a dynamic obeying an inverse square law, but that it is also compatible with a dynamic in which the force to differentiate decreases linearly with distance.

The general, cross-linguistic relation between geography and linguistic variation is clearly of great potential interest to dialectology and the study of linguistic diffusion and deserves further attention. Progress can be made by obtaining a broader selection of linguistic case studies on which to base the general views, by the development of a metric that is applicable cross-linguistically without confounding effects, and by the development of hypotheses concerning the more exact form of the curves.

7. The proximate horizon

Dialectology has always based its appeal on its attention to the incredible detail and multi-faceted nature of linguistic variation. Dialectometry serves dialectology by improving the efficiency of its data archiving techniques, and by developing efficient and replicable data analysis techniques, which in turn broaden the empirical base on which theoretical dialectology can build. Finally, the opportunity to measure linguis-

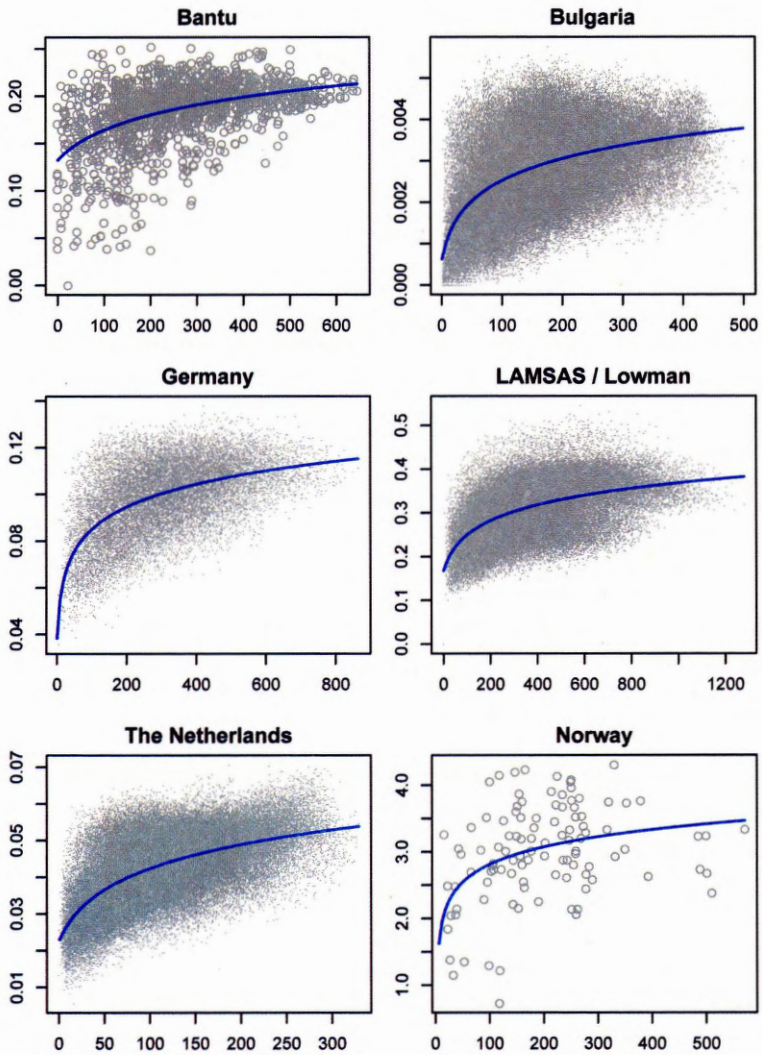


Figure 3

Séguy's law

tic variation validly and consistently may open the door to more abstract and general characterizations of dialectology's central ideas.

Several topics deserve spots high on the dialectometrical agenda. We mention first two practical matters. First, although the Salzburg VDM package and the Groningen L04 package are being used profitably at other centers, it would be worthwhile to develop new packages or new versions of these packages which are easier for linguists with standard dialectological training to use. Second, it would be worthwhile devel-

oping techniques to extract variation from the increasingly available corpora which are collected for the purposed of pure and applied linguistic research (Szmrecsanyi 2008).

Turning to the more abstract scholarly topics, we should mention third, the relation between synchronic variation and historical linguistics. There is widespread consensus that language variation and language change belong together as topics, and there is increasing interest in the use of exact techniques in historical linguistics (Nakleh, Ringe & Warnow 2005). It would be only natural to see some cross fertilization in these two areas, for example in the deployment of the alignment techniques discussed here and the application of phylogenetic inference, which to-date has mostly used lexical data and/or manually prepared phonetic or phonological data.

A fourth opportunity for progress in dialectometry lies in the further validation of techniques. By validation, we mean the proof that the techniques are indeed measuring what they purport to measure, an undertaking which presupposes that one has been explicit about what this is. A good number of studies are undertaken without being explicit on this point: the researchers seem content to show that 81% of the vocabulary is shared between sites s_1 and s_2, etc. But the ease with which alternatives are developed and implemented makes further reflection on this point absolutely imperative. Even in the case of share vocabulary, we ask whether it is enough that some sort of cognate is found, whether the order of preferences for certain lexicalizations above others has been taken into account, whether the significance of shared vocabulary might not need to be corrected for frequency effects, etc. These considerations led Gooskens and Heeringa (2004) to suggest that dialectometric measurements be understood as measuring the signals of provenance which dialect speakers provide, which in turn led them to view to validate the measurements on the basis of dialect speakers' abilities to identify a dialect as like or unlike their own. Further studies along these lines would be most welcome, if for no other reason, than to guard against depending too much on a single behavioral study done on speakers of a single, Scandinavian language.

A fifth, but related opportunity certainly lies in the further investigation of the relation between comprehensibility (as discussed above in Section 5) and the signal of provenance that was central in Gooskens and Heeringa (2004). It is clear that comprehensibility and the signal of "likeness" correlate to a high degree, giving rise to the question of whether the two are ultimately the same, or, alternatively where they part their ways and with respect to which (linguistic) phenomena. More empirical studies investigating the overlap would certainly advance the field.

Sixth, we do not mean to suggest that the spectral clustering techniques presented in Section 4 above should be regarded as closing the book on the issue of how to identify the linguistic basis of dialect distributions. We are optimistic that these techniques are promising, but they should be compared *inter alia* to techniques such as Shackleton's (2007) use of principal component analysis. The quality of the groups detected needs further investigation, and the impact of factors such as frequency would be worthwhile examining closely.

Seventh, and finally, we hope that further research into the general relation between geographical distance and dialectal difference is a promising avenue for further work, as we hope to have suggested in Section 6 above.

References

Alonso L., Castellon, I., Escribano, J., Messegeur, X. & L. Padro, 2004, «Multiple sequence alignement for characterizing the linear structure of revision», in *Proceedings of the 4th International Conference on Language Resources and Evaluation.*

Atkinson, Q., Nicholls, G., Welch, D. & R. Gray, 2005, «From words to dates: water into wine, mathemagic or phylogenetic», *TPS* 103, 193-219.

Beijering, K., Gooskens, C. & W. Heeringa, 2008, «Predicting intelligibility and perceived linguistic distances by means of the Levenshtein algorithm», in M. van Koppen & B. Botma (eds.), *Linguistics in the Netherlands*, John Benjamins, Amsterdam, 13-24.

Black, P., 1976, «Multidimensional Scaling applied to Linguistic Relationships», *Cahiers de l'Institut de linguistique de Louvain* 3, 43-92.

Börestam Uhlmann, U., 1991, *Språkmöten och mötesspråk i Norden* [Language meetings and the language of meetings], Nordisk språksekretariat, Oslo.

Delsing, L-O. & K. Lundin Åkesson, 2005, *Håller språket ihop Norden? En forskningsrapport om ungdomars förståelse av danska, svenska och norska* [Does the Language Keep the Nordic Countries Together? A Research Report on How Well Young People Understand Danish, Swedish and Norwegian], Nordiska ministerrådet, Copenhagen.

Dhillon, I., 2001, «Co-clustering documents and words using bipartite spectral graph portioning», in *Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining*, ACM, New York, 269-274.

Goebl, H., 1982, *Dialektometrie: Prinzipien und Methoden des Einsatzes der Numerischen Taxonomie im Bereich der Dialektgeographie.* Osterreichische Akademie der Wissenschaften, Wien.

Gooskens, C., 2007, «The contribution of linguistic factors to the intelligibility of closely related languages», *Journal of Multilingual and Multicultural Development* 28 (6), 445-467.

—, Beijering, K. & W. Heeringa, 2008, «Phonetic and lexical predictors of intelligibility», *International Journal of Humanities and Arts Computing* 2 (1-2), 63-81.

— & W. Heeringa, 2004, «Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data», *Language Variation and Change* 16(3), 189-207.

—, Heuven, V. Van, Bezooijen, R. van & J, Pacilly, accepted, «Is spoken Danish less intelligible than Swedish?», *Speech Communication.*

Gray, R. & Q. Atkinson, 2003, «Language-tree divergence times support Anatolian theory of Indo-European origin», *Nature* 405, 1052-1055.

Gusfield, D., 1997, *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology.* CUP.

Heeringa, W., 2004, *Measuring Dialect Pronunciation Differences using Levenshtein Distance.* PhD thesis, Rijksuniversiteit Groningen.

Hubert, A. & P. Arabie, 1985, «Comparing partitions», *Journal of Classification* 2, 193-218.

Kleinberg, J., 2003, «An impossibility theorem for clustering», in S. Becker, S. Thrun & K. Obermaier (eds.), *Advances in Neural Information Processing Systems* 15. Avail. at http://books.nips.cc/papers/files/nips15/LT17.pdf.

Kürschner, S., Gooskens, C. & R. van Bezooijen, 2008, «Linguistic determinants of the intelligibility of Swedish words among Danes», *International Journal of Humanities and Arts Computing* 2 (1-2), 83-100.

Maurud, Ø., 1976, *Nabospråksforståelse i Skandinavia. En undersøkelse om gjensidig forståelse av tale- og skriftspråk i Danmark, Norge og Sverige* [Mutual intelligibility of languages in Scandinavia. A study of the mutual understanding of written and spoken language in Denmark, Norway and Sweden], Nordiska Rådet, Stockholm.

Moberg, J., Gooskens, C., Nerbonne, J. & N. Vaillette, 2007, «Conditional Entropy Measures Intelligibility among Related Languages» in P. Dirix, I. Schuurman, V. Vandeghin-

ste & F. van Eynde (eds.), *Computational Linguistics in the Netherlands 2006: Selected papers from the 17th CLIN Meeting*, LOT, Utrecht, 51-66.

Nakleh, L., Ringe, D. & T. Warnow, 2005, «Perfect phylogenetic networks: A new methodology for reconstructing the evolutionary history of natural languages», *Language* 81(2), 382-420.

Nerbonne, J., 2009, «Data-Driven Dialectology», *Language and Linguistic Compass*. 3(1), 2009, 175-198. DOI: 10.1111/j.1749-818x2008.00114.x

— (to appear, a), «Various Variation Aggregates in the LAMSAS South», in C. Davis & M. Picone (eds.) *Language Variety in the South III,* University of Alabama Press, Tuscaloosa.

— (to appear, b), «Mapping Aggregate Variation», in S. Rabanus, R. Kehrein & A. Lameli (eds.), *Mapping Language,* De Gruyter, Berlin.

— (to appear, c), «Measuring the Diffusion of Linguistic Change», *Philosophical Transactions of the Royal Society B: Biological Sciences.*

— & W. Heeringa, 1997, «Measuring Dialect Distance Phonetically», in J. Coleman (ed.), *Workshop on Computational Phonology,* ACL, Madrid, 11-18.

— & —, 2007, «Geographic Distributions of Linguistic Variation Reflect Dynamics of Differentiation», in S. Featherston & W. Sternefeld (eds.), *Roots: Linguistics in Search of its Evidential Base*, De Gruyter, Berlin, 267-297

— & —, 2009, «Measuring Dialect Differences», in J.-E. Schmidt & P. Auer (eds.), *Language and Space: Theories and Methods* in series *Handbooks of Linguistics and Communication Science*, Chap. 31, De Gruyter, Berlin, 550-567.

—, — & P. Kleiweg, 1999, «Edit Distance and Dialect Proximity», in D. Sankoff & J. Kruskal (eds.), *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison,* 2nd ed., CSLI Press, Stanford, v-xv.

—, Kleiweg, P., Heeringa, W. & F. Manni, 2008, «Projecting Dialect Differences to Geography: Bootstrap Clustering vs. Noisy Clustering», in Ch. Preisach, L. Schmidt-Thieme, H. Burkhardt & R. Decker (eds.), *Data Analysis, Machine Learning, and Applications. Proc. of the 31st Annual Meeting of the German Classification Society,* Springer, Berlin, 647-654.

Prokić, J., 2007, «Identifying linguistic structure in a quantitative analysis of dialect pronunciation», *Proceedings of the ACL 2007 Student Research Workshop*, ACL, Prague, 61-66.

— & J. Nerbonne, 2008, «Recognizing Groups among Dialects», *International Journal of Humanities and Arts Computing*, 153-172. DOI: 10.13366/E1753854809000366.

—, Wieling, M. & J. Nerbonne, 2009, «Multiple string alignments in linguistics», in L. Borin & P. Landvai (chairs), *Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education (LaTeCH-SHELT&R 2009),* EACL Workshop.

Séguy, J., 1971, «La relation entre la distance spatiale et la distance lexicale», *Revue de Linguistique Romane* 35, 335-357.

Shackleton, R. G., 2007, «Phonetic Variation in the Traditional English Dialects», *Journal of English Linguistics* 35(1), 30-102. DOI: 10.1177/0075424205297857.

Szmrecsanyi, B., 2008, «Corpus-Based Dialectometry: Aggregate Morphosyntactic Variability in British English Dialects», *International Journal of Humanities and Arts Computing* 2, special issue on *Computing and Language Variation*, ed. by J. Nerbonne, C. Gooskens, S. Kürschner & R. van Bezooijen, 261-278.

Trudgill, P., 1974, «Linguistic change and diffusion: Description and explanation in sociolinguistic dialect geography», *Language in Society* 2, 215-246.

Warnow, T., Evans, S., Ringe, D. & L. Nakhleh, 2006, «A stochastic model of language evolution that incorporates homoplasy and borrowing», in P. Foster and C. Renfrew (eds.), *Phylogenetic Methods and the Prehistory of Languages*, MacDonald Institute for Archeological Research, Cambridge.

Wells, J. & J. House, 1995, *The Sounds of the International Phonetic Alphabet.* Dept. Phonetics & Linguistics, University College London. Booklet with tape.

Wieling, M., Heeringa, W. J. and J. Nerbonne, 2007, «An aggregate analysis of pronunciation in the Goeman-Taeldeman-van Reenen-Project data», *Taal en Tongval* 59, 84-116.

— & J. Nerbonne, 2009, «Bipartite spectral graph partitioning to co-cluster varieties and sound correspondences in dialectology», in M. Choudhury et al. (eds.) *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing (TextGraphs-4)*, ACL-IJCNLP, Singapore, 7 August 2009, 14-22.

— & — (to appear), «Bipartite spectral graph partitioning for clustering dialect varieties and detecting their linguistic features», *Computer Speech and Language.*

—, Prokić, J. and J. Nerbonne, 2009, «Evaluating the pairwise string alignment of pronunciations», in L. Borin & P. Landvai (chairs), *Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education (LaTeCH-SHELT&R 2009)*, EACL Workshop.

TOOLS FOR DIALECT SYNTAX: THE CASE OF CORDIAL-SIN (AN ANNOTATED CORPUS OF PORTUGUESE DIALECTS)

Ernestina Carrilho
Universidade de Lisboa

Abstract

This paper addresses methodological issues of concern to the study of morphosyntactic variation. While the empirical basis of dialect syntax is still a matter of elaboration, the focus will be here on the role of dialect corpora as tools for the study of linguistic variation in this particular domain. The case of CORDIAL-SIN, an annotated corpus of Portuguese dialects, will be presented along with some initial advances in Portuguese dialect syntax. Two levels of tools for the study of linguistic variation will thus be addressed here: (i) corpora as general tools for dialect syntax; and (ii) tagging and syntactic annotation within a dialect corpus as tools that ease the way how variation in morphosyntax can be studied.

Section 1 introduces methodological remarks concerning the empirical ground for dialect syntax; the CORDIAL-SIN is presented in section 2; section 3 briefly illustrates how this tool has enhanced the development of Portuguese dialect syntax.

Key words: *Dialect syntax; morphosyntactic variation; dialect corpus; annotated corpus; European Portuguese dialects.*

1. On the empirical basis of dialect syntax

It is well known and often mentioned that the study of syntax has only played a very marginal role in traditional dialectology. Dialectologists have mainly been concerned with the study of phonological and lexical variation, and it was for this purpose that data were systematically collected in dialect surveys and linguistic atlases, which represent the main data source for traditional dialectology.[1] In fact, in the atlas projects all over the world only a scarce part of the published dialect maps involve syntactic data (Cornips and Jongenburger 2001: 1).

This neglect of syntax in dialect studies certainly owes much to the methodological difficulties that syntactically oriented fieldwork raises. The classical method of dialectological interviews, conducted with the help of a questionnaire, hardly combines with the gathering of specific syntactic constructions, for which naming ques-

[1] To this respect, it is worth remembering some notable exceptions such as Remacle's (1952-60) work on the syntax of the Walloon dialect of *la Gleize*.

tions or even completing questions happen to be fairly unhelpful. Also, oral translation from the standard language, a method used for eliciting syntactic properties in most European linguistic atlases, is far from unproblematic. It has been acknowledged that this kind of elicitation raises several problems, among which a high risk of getting an answer influenced by the standard construction (a.o. Bucheli and Glaser 2002: 3). Besides, such a method is only conceivable (despite its imperfections) for those areas where variation meets different linguistic systems (which may be the case of Italy, France or Switzerland, but is not the case for the most part of the Portuguese territory, for instance).

The last decade of the 20th century witnessed however a general and renewed interest in syntactic variation[2] and, concomitantly, new methodological concerns were brought about the empirical ground for this domain of linguistic inquiry. In fact, over the last two decades, several projects dealing with the syntax of dialects have been independently established in different European countries, some of which are still under development: among others, the *Syntactic Atlas of the Dutch Dialects* (SAND); the *Syntactic Atlas of Northern Italy* (ASIS); the *Audible Corpus of Spoken Rural Spanish* (COSER); the project *English Dialect Syntax from a Typological Perspective*; and, more recently, the supranational projects *ScanDiaSyn (Scandinavian Dialect Syntax)*, across the Scandinavian dialect continuum, and *Edisyn* (*European Dialect Syntax*), an European project specifically aimed at developing cooperation among dialect syntax projects in Europe through similar or common methodologies (regarding data collection, data storage and annotation, data retrieval, cartography). Among these projects, the data that feed the empirical demands of dialect syntax range from a corpus of independently collected speech (as in the project *English Dialect Syntax from a Typological Perspective*) to written questionnaires requesting translations into the informants' dialect (as in the first phase of ASIS). The advantages and disadvantages of both naturalistic methods behind corpus-data and elicitation techniques have been recurrently discussed. It may easily be acknowledged that none of them is exempt of problems.

Although naturalistic corpus-based data can hardly circumvent problems such as the lack of negative evidence and the weak representation (if any) of sentence types that are rare in spontaneous speech, the experience of fieldwork data collection through elicitation has however proven that this method is not free of difficulties either (see also Labov 1996).

> Every elicitation situation is artificial, because the subject is being asked for a sort of behavior that is entirely different from everyday conversation (cf. Schütze 1996: 3). Sociolinguistic research has clearly shown that the response of subjects on direct judgement tasks ('Is this a good sentence in your dialect?') often tends to reflect the form which they believe to have prestige or obeys the learned norm, rather than the form they actually use (Labov 1972: 213). A reasonable alternative is to use more indirect elicitation tasks (e.g. 'Do you encounter this sentence in your dialect?') Different levels of speech style (informal and formal) yield another complicating factor for syntactic data elicitation. (Barbiers and Cornips 2002: 8-9)

[2] Actually, such an interest was already sketched in the late seventies within the international geolinguistic project ALE, *Atlas Linguarum Europae*, which already stated the convenience for syntactic theory to count on a comparative inquiry into dialect syntax (see Lehman 1980, Kruijsen 1983).

In fact, past experiences have often shown that the results obtained through elicitation data may differ from those appearing in spontaneous speech (Cornips 2003); also, different elicitation methodologies may often lead to different results (Auckle, Buchstaller, Corrigan and Holmberg 2007). At least, such results appear to suggest that more research is still needed in order to decide on the reliability of the different elicitation techniques.

The pratice developed within the SAND project (Cornips and Poletto 2005, Barbiers et al. 2007), known as a "layered methodology", may be taken as exemplar. The phases of planning the SAND data collection involved, as a first step, a comprehensive literature study. As a second step, a written questionnaire has been prepared on the basis of the syntactic phenomena described in the literature. As Cornips and Jongenburger (2001) report, this questionnaire was carefully prepared to provide insight into (i) the geographic distribution of syntactic variation; (ii) the validity of each type of (written) elicitation; (iii) areas particularly interesting with respect to syntactic variation. As such, the questionnaire served as the input for the next phase, which consisted of oral fieldwork. Preparing the oral fieldwork for the SAND project involved the consideration of an appropriate elicitation task for each syntactic variable to be investigated. The results with respect to the usability of both written and oral elicitation techniques show that not every task is easy to perform and that not every task is adequate to every type of syntactic variable. Fundamental aspects concerning the reliability and the workability of elicitation in dialect syntax may thus be evoked: (i) though useful, elicitation tasks are not without problems; (ii) the negative effects associated to each elicitation task must always be carefully evaluated; (iii) combining different types of data collection methods may help obviating some problems; (iv) dialect syntax analysis requires careful consideration of the relation between the collected data and the effects induced by the method by means of which the data are obtained.

Dialect syntax projects may thus take advantage from a layered methodology that can combine different sources of data collection tasks: besides the appropriate elicitation tests, also interview techniques that can generate more naturalistic speech. Such a practice has recently been adopted in large-scale dialect syntax projects, such as the *ScanDiaSyn*.

As a matter of fact, naturalistic data have also played a non-negligible role in setting up recent advances in dialect syntax. Within different linguistic domains, dialect corpora became also important heuristic tools for the study of non-standard syntax. For instance, we may refer to the *Freiburg English Dialect Corpus* (FRED), a computerized corpus of English dialects, within the project *English Dialect Syntax from a Typological Perspective* (Kortmann 2002), or to the above mentioned COSER, seminal to recent research on several aspects of Spanish dialect syntax (see Fernández-Ordóñez 2009).

Thus, even if we acknowledge the importance of introspective syntactic data that only elicitation tasks may provide, the limits of such a data collecting methodology are also to be remembered when it comes to the study of non-standard syntax. The linguist preparing such data collecting tasks can hardly be familiar with the different varieties of his native language.

> One point which might be made is that this method [the introspective method, EC] cannot be used to study any language or language variety not known to the in-

> vestigator, and since academic linguists are seldom competent speakers of non-standard dialects or uncodified languages, can in practice be used for describing only fully codified languages. This is not of course to deny that those who have grown up as native speakers of a dialect (for example, Peter Trudgill in Norwich [...]) may have intuitions about its structure; so also might non-native speakers who have developed an intimate knowledge of the structure of a dialect (see J. Milroy 1981 for an example). But descriptions of non-standard dialects generally use intuition as an aid to focusing the investigation, rather than a basic method; [...]. (Milroy 1987: 76)

Above all, introspection alone could hardly be invoked as a source of hypotheses-motivating data central to elicitation tests' design. In this context, thus, dialect corpora can play a different role. Observations sometimes formulated with respect to the empirical basis of linguistic research in general appear especially significant when referring to empirical methodologies applied to the study of dialect syntax:

> The advantage of working with a corpus is, of course, the enhanced objectivity of the data and of all the research that is based on it. In comparison with the other approaches, the possibilities for the researcher to manipulate the data are minimized. Another great advantage is that a corpus the researcher has not produced himself may be varied, heterogeneous, full of surprises and a constant source of inspiration. Exposing oneself to spontaneous data is, in fact, the safest way of discovering those categories of a language [EC: or of a dialect] that are peculiar to it and that the researcher did not expect. (Lehmann 2004: 201)

This is so much so to the extent that comprehensive or specific descriptive dialect syntactic studies are often unavailable for some languages. A dialect corpus, if available, may then be *full of surprises.*

2. CORDIAL-SIN: the syntax-oriented corpus of Portuguese dialects

2.1. Background

By the end of the 20th century, the condition of dialect syntax in Portuguese studies was not significantly different from what happened in other linguistic domains. The major Portuguese dialect surveys had not generally contemplated any kind of syntactic variation, and the questionnaire of the linguistic atlas ALEPG (*Atlas Linguístico-Etnográfico de Portugal e da Galiza*) explicitly stated that "for practical reasons" it did not include syntactic questions (Gottschalk, Barata and Adragão 1974).

Nevertheless, even if no comprehensive description of syntactic variation phenomena was available, there existed sparse allusions to syntactically relevant variation in European Portuguese. These could in fact be found from the pioneering work in Portuguese dialectology by Leite de Vasconcellos (1901) to many different dialect monographs written near the mid-20th century. However, the place for syntax was usually very marginal when compared to that of lexicon, phonology or even morphology.

It was against this background that CORDIAL-SIN began to be compiled, in 1999. The acronym stands for the Portuguese name "*Cor*pus *Dial*ectal para o Estudo da *Sin*taxe" ("Syntax-oriented Corpus of Portuguese Dialects") and it has mainly been conceived as a major empirical resource for the study of dialect syntax.

As a very important condition for the CORDIAL-SIN genesis, I shall mention the existence of a rich collection of tape-recorded dialect speech, gathered by the Centro de Linguística da Universidade de Lisboa. At this Center, the research group working on Linguistic Variation has been committed to several projects of dialect geography, for which fieldwork interviews have been conducted from the mid-seventies till the beginning of 2000. In the course of such interviews, informants often speak about their story of life, make observations on aspects inquired by the questionnaire, comment on etnographic issues, which amounts to an important extent of spontaneous speech, collected in fairly controlled and homogeneous conditions.

The CORDIAL-SIN project aimed precisely at making available for researchers in general (and especially for those interested in dialect syntax) a significant amount of spontaneous and semi-directed speech drawn from these data. More specifically, this project also aimed at providing fast and systematic access to precise morphological and syntactic information —which motivated the building up of an annotated corpus, marked up with morphological and syntactic information—.

By compiling and making available such an empirical resource for dialect syntax, the CORDIAL-SIN team has also been engaged in the enhancement of research activity on syntactic dialect variation in European Portuguese.

2.2. A dialect corpus

A team coordinated by Ana Maria Martins has been committed to the selection, transcription, annotation and publication of this corpus, compiled from sources such as the ALEPG, the ALLP, the ALEAç and Segura (1987). The corpus amounts to 600,000 words, collected from 42 locations within continental Portugal and the archipels of Madeira and Azores.[3]

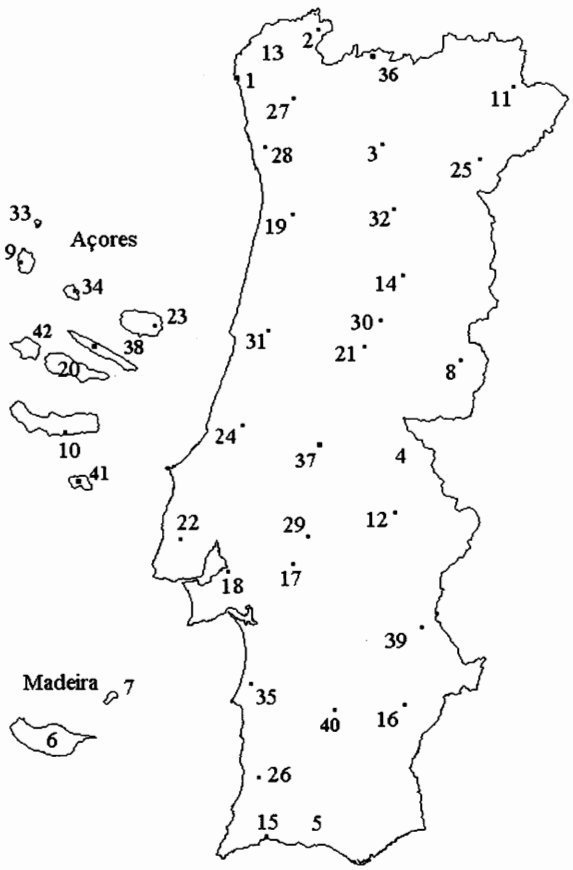
The speakers' sociological profile is fairly constant across locations. Given the sources for this corpus, informants correspond to the traditional type of informant in dialect geography: old, non-educated, rural and born and raised in place of interview.

The corpus is freely available through the internet (http://www.clul.ul.pt/english/sectores/variacao/cordialsin/projecto_cordialsin_corpus.php), under three different formats, for the moment: (i) *verbatim* orthographic transcripts; (ii) 'normalized' orthographic transcripts; and (iii) morphologically tagged texts. In the future, CORDIAL-SIN will also be available as a syntactically annotated corpus.

Verbatim orthographic transcripts include the marking up of some syntactically relevant phonetic and morphological variants, and of generalized spoken language phenomena, such as hesitations, filled and empty pauses, repetitions, rephrased segments, false starts, truncated words, speech overlappings, unclear productions (see (3) below). 'Normalized' orthographic transcripts correspond to a simplified version of *verbatim* transcripts, automatically obtained through elimination of the marked up features of spoken language and of phonetic transcriptions. The normal-

---

[3] CORDIAL-SIN has been funded by FCT (*Fundação para a Ciência e a Tecnologia*), through the following projects: PRAXIS XXI/P/PLP/13046/1998; POSI/1999/PLP/33275; POCTI/LIN/46980/2002; PTDC/LIN/71559/2006.



1. CORDIAL-SIN locations (see Appendix for identification)

ized transcripts are the input for the tagging and the syntactic annotation. An example of this two-layered transcription is given below:[4]

(1) *verbatim* ortographic transcript
Eu sei que aquilo que{fp} {PH|nũ=não} é por mal, sabe? Mas quem ouve... Vem cá uma pessoa estranha, {PH|nũ=não} é, {PH|nũ=não} conhece e diz: "Ah, [AB|são] são *malcriados, os pescadores" (...).* [Vila Praia de Âncora, VPA15]

(2) 'normalized' ortographic transcript (ASCII version)
Eu sei que aquilo que não é por mal, sabe? Mas quem ouve... Vem cá uma pessoa estranha, não é, não conhece e diz: "Ah, <break> (...) </break> são malcriados, os pescadores" <break> (...) </break>.

[4] On the conventions used in *verbatim* transcripts and their relation to normalized transcripts see *Normas de Transcrição*, http://www.clul.ul.pt/english/sectores/variacao/cordialsin/manual_normas.pdf.

(3) Examples of marked-up spoken language phenomena:
{PH|nũ=não} — phonetic variant for the negation word
{CT|pa=para a} — contraction 'to+the.FEM'
{AB|xxx} — false starts, abandoned sequences
{pp} — empty pauses
{fp} — filled pauses
[underlining] — overlapping
(...) — unclear sequences (also: omitted sequence, e.g. [AB|...], in 'normalized' transcripts)

2.3. Tagging and syntactic annotation in a dialect corpus

Further development of CORDIAL-SIN endowed this dialect corpus with tagging for each word. This has been conceived as a first step towards fast and systematic access to precise morphosyntactic information, which will ultimately be achieved with syntactic annotation.

CORDIAL-SIN tagging and syntactic annotation were both made easier by automatic tools already developed and in use within other related projects. More concretely, tagging and syntactic annotation have been largely inspired by the processes and tools used by the *Penn Parsed Corpora of Historical English* —a set of corpora developed at the University of Pennsylvania by Anthony Kroch and his associates—, and also by the *Tycho Brahe Parsed Corpus of Historical Portuguese*, a corpus of Portuguese authors born from the 16th to the 19th centuries, coordinated by Charlotte Galves at the University of Campinas (Brazil).

Collaborative work with the teams developing these corpora has permitted the tuning of already available tagging and annotation tools in such a way that these could satisfactorily apply to dialectal European Portuguese and serve our purpose. Besides accelerating the tagging and annotation phases, this cooperation also ensures the ease of linguistic information retrieval, since a query tool operating on the annotation system in use is already available.

2.3.1. Tagging

The morphological tagging operation has been to a great extent facilitated by the use of an automated morphological tagger, designed for the *Tycho Brahe Corpus* of Portuguese texts (Finger 1998). After training over a sample of 30,000 hand tagged CORDIAL-SIN words, the rate of accuracy of the tagger proved to be satisfactory enough to encourage the use of its output as the basis for a hand refined and corrected tagged version of the corpus. To ensure the precise format of the tags, an additional automatic tool has been used after manual tag correction and refinement.

Thus, CORDIAL-SIN's morphologically tagged transcripts result from a three steps process involving: (i) automatic tagging by the Tycho Brahe tagger; (ii) manual tag correction and refinement using the CORDIAL-SIN's morphological annotation system; (iii) automatic verification of the corrected tags.

The format of the morphological tags and the basics of the tagset of the CORDIAL-SIN essentially stem from the system designed for the Tycho Brahe automatic tagger (Galves and Britto 1999). Tags have an internal structure consisting of an ever-present main tag, which includes part-of-speech tags (such as D, for determiner), and, in certain cases, sub-tags (for instance, F for feminine, P for plural), diacritics attaching different main tags ("+", "!") or main tags to sub-tags ("–"), and figures indicating clusters, as in the following examples:

Tag	Application	Ex.
/D	singular masculine determiner	*o/D*
/D-P	plural masculine determiner	*os/D-P*
/D-F-P	plural feminine determiner	*as/D-F-P*
/P+D-F	preposition plus singular feminine determiner contraction	*da/P+D-F*
/VB+CL	verb (infinitive) plus enclitic pronoun	*dar-lhe/VB+CL*
/VB-R-1S!CL	verb (future) plus mesoclitic pronoun	*dar-te-ei/VB-R-1S!CL*
/P31	first element of a triple prepositional cluster	*por/P31 mor/P32 de/P33*

2. Examples of CORDIAL-SIN tags

The set of sub-tags codifies inflectional information — tense/mood and person/number for verbs or gender and number for nominal categories. It also specifies in more detail some morpho-syntactic information (such as a -NEG sub-tag that identifies different negative categories, like adverbs, quantifiers, prepositions). The system also allows: (i) main tags attachment for contractions or cliticizations; and (ii) tags and figures combination for multiple words behaving as clusters.

The tags thus obtained have a structured format that straightforwardly allows for very detailed morphological information, a very appealing solution when tagging a morphologically rich language, such as European Portuguese. As a welcome result a number of possible structured tags higher than 1.000 can be obtained from the CORDIAL-SIN tagset, which reduces to c. 40 main tags plus a smaller set of 25 sub-tags. (For a detailed description of the entire tagset and its application, see *Manual of the CORDIAL-SIN Morphosyntactic tagging*, http://www.clul.ul.pt/sectores/variacao/cordialsin/manual_anotacao_morfologica.pdf.)

2.3.2. Syntactic annotation

CORDIAL-SIN syntactic annotation is currently under development (2008-2010, within the project DUPLEX).[5] The annotation processes and tools in use have been developed for the *Penn Parsed Corpora of Historical English* (and the same or very similar annotation system is equally used on the *Tycho Brahe* corpus).

[5] Project PTDC/LIN/71559/2006, funded by FCT (http://www.clul.ul.pt/english/sectores/variacao/projecto_duplex.php).

Adopting such a rich annotation system (developed for English corpora) for a Portuguese corpus required a cautious adaptation of the existing system to data that differ from English in many respects. The initial phase of the CORDIAL-SIN syntactic annotation process has thus been devoted to the tuning of the basic annotation system, a task which was carried out in strict collaboration with the *PennCorpora* and the *Tycho Brahe* teams. Hand annotation of a 10,000 words sample of the corpus has served to define and consolidate the main guidelines of the system so as it could apply to Portuguese spoken texts. These general guidelines resulted in the first version of the annotator's manual.

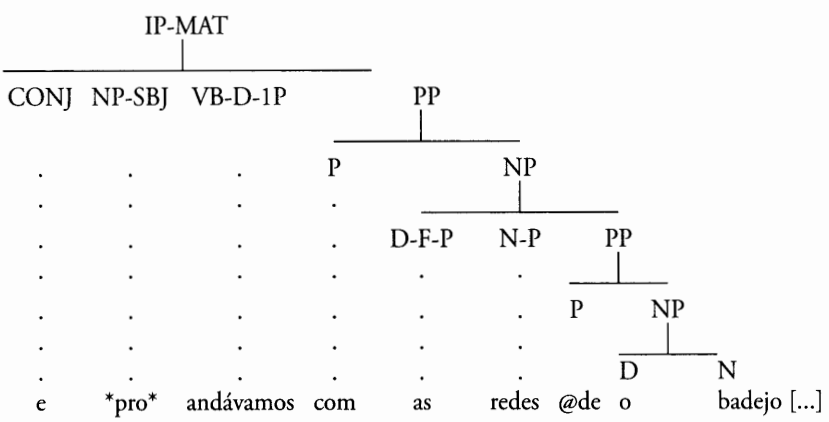
Data annotation itself is usually a very complex task. In the present case, additional complexity was expected, given the spoken and dialectal nature of the corpus. Sentences that call for detailed consideration are frequent, even though the basic lines of the system are already defined. Difficult annotations are decided upon after discussion, and each new difficult example is added to the annotator's manual, in order to assure consistency. Thus, the syntactic annotation guidelines have been progressively enriched during the whole course of the annotation phase, as more data are analysed and as new difficult sentences arise. (The current version of the *Syntactic Annotation Manual* is available at: http://www.clul.ul.pt/sectores/variacao/cordialsin/Syntactic%20annotation%20manual.html.)

The CORDIAL-SIN syntactically annotated transcripts are built on previously tagged texts, such as:

(4) e/CONJ andávamos/VB-D-1P com/P as/D-F-P redes/N-P @de/P @o/D badejo/N ,/, que/WPRO são/SR-P-3P mais/ADV-R baixas/ADJ-F-P .../. [VPA07]

The syntactic annotation produces a tree representation in the form of labeled brackets, where depth of indenting corresponds to depth of structural embedding. As for the *Penn Parsed Corpora*, the annotation represents quite flat trees, allowing for multiple branching nodes and, for some words, projecting only a word-level node (e.g. inflected verbs, negation).

(5) a.



b.

```
(IP-MAT(CONJ e)
        (NP-SBJ *pro*)
        (VB-D-1P andávamos)
        (PP    (P com)
               (NP   (D-F-P as)
                     (N-P redes)
                     (PP     (P @de)
                             (NP    (D @o)
                                    (N badejo))
                     (, ,)
                     (CP-REL        (WNP-1      (WPRO que))
                                    (IP-SUB     (NP-SBJ *T*-1)
                                                (SR-P-3P são)
                                                (ADJP (ADV-R mais)
                                                      (ADJ-F-P baixas))))))
        (. ...)) [VPA07]
```

In addition to constituent boundaries and phrase and clause dependencies, the annotation marks up grammatical relations, clause and sentence type, some empty categories (such as null subject and null object), among others. At the word level, morphological labels are preserved. Phrase and clause labels indicate category (NP, PP...), often specified by an extended label indicating syntactic function (e.g. subject, direct object), clause type (e.g. relative, adverbial, interrogative), or other relevant information (e.g. left dislocation, pragmatic marker).

It is worth noting that such an annotation scheme is to be seen as a fairly theory-neutral representation of constituent structure, to which category and function labels are added. The main goal of the syntactic annotation is thus to facilitate automated searches for various constructions, not to associate every sentence with an adequate structural description. Controversial decisions on annotation are avoided (for instance, by omitting undecidable information —such as the attachment of a PP as complement or as adjunt—; or by using default rules), so that the annotation scheme is completely predictable and so exploitable for automatic searches.

Turning now to the annotation process, three different stages must be mentioned: (i) a stage of automatic parsing of the data, in which the *Penn Corpora* version of a statistical parser (Collins 1999, Bikel 2004) runs over the tagged texts (at Univ. Pennsylvania); (ii) a stage of human editing of the parsed output, a time-consuming task carried out with the help of *CorpusDraw*, an editing annotation tool; (iii) finally, the result is a parsed version of the corpus in such a format that allows data retrieval through syntactic configurations (automated searches become possible with the aid of *CorpusSearch*, a search engine for parsed corpora). Both *Corpus Search* and *Corpus Draw* are components of *CorpusSearch2* —a Java program that supports research in corpus linguistics, developed by Beth Randall at University of Pennsylvania (Randall 2005-2007)—. This program, which is freely available from http://corpussearch.sourceforge.net, is thus useful both for the construction of syntactically parsed corpora and for searching them.

CorpusDraw gives support to the human editing of the parser output, which may involve: changing syntactic tags, adding subcategory information, changing attachment level, adding empty categories, for instance. *CorpusSearch,* in turn, is a dedicated engine for parsed corpora permitting basic search functions that are linguistically intuitive: for instance, *(immediately)precedes, (immediately)dominates, exists, hassister,* among others.

CORDIAL-SIN is now in the intermediate stage of the annotation process: the output of the automated parser is under manual correction with the aid of *CorpusDraw*. As mentioned above, this task is also that of defining the details of CORDIAL-SIN's annotation system within the standards already operative in other corpora and readable by the automatic parser. Adapting the already defined system mainly involves finding solutions required by those grammatical aspects where Portuguese and English differ (a task shared with the Tycho Brahe team) or by other aspects that are characteristic of CORDIAL-SIN's dialectal and spoken data. Also at the level of the label set, a very small number of extended labels have been added, all of them relating to aspects particular to spoken language.

3. CORDIAL-SIN as a tool for Portuguese dialect syntax

The annotated dialect corpus CORDIAL-SIN has been —and still is— the main empirical source for a number of studies on different aspects of Portuguese dialect syntax (see http://www.clul.ul.pt/english/sectores/variacao/cordialsin/projecto_cordialsin_publicacoes.php). Some of these studies have already counted on the available tagging of this corpus. If we refer to a very simple example, investigating inflected gerunds as a dialectal feature in European Portuguese (as it has been achieved in Lobo 2008) could begin by obtaining a list of inflected gerunds in CORDIAL-SIN. Through the tagged corpus, this can easily be achieved just in a couple of seconds with any concordancing program. Concordances can thus easily operate over the relevant tags and sub-tags (here the sub-tags G for 'gerund' and -F for 'inflected'), providing very precise results in a very short span of time.[6] Given the distribution of all CORDIAL-SIN locations, this corpus may also provide insight into the geographic distribution of syntactic variants (see Carrilho and Pereira 2009). Finally, and perhaps more surprisingly, CORDIAL-SIN has revealed the type of till-then-unknown data without which a researcher could hardly prepare relevant and adequate elicitation tests. This was in fact the case of all the wide spectrum of (mostly unknown) constructions featuring expletive *ele* found in CORDIAL-SIN, which motivated a proposal for the re-evaluation of the received view about the grammatical status of this expletive in European Portuguese (Carrilho 2005).

The importance of making accessible to other researchers detailed syntactic information about dialectal data is twofold: firstly, it eases the way to have a closer look at dialectal data relevant for the study of syntax in general and it permits to know their

[6] Within the European project *Edisyn,* a Search Engine is currently under development. An experimental version of this Search Engine, which allows searches for part-of-speech tags across different corpora and databases (among others, the *SAND,* the *ASIS* and a corpus of Estonian Dialects), can already be operated over the CORDIAL-SIN tagged data.

geographical distribution; and secondly, it provides researchers with a wider range of syntactic phenomena, some of which pervasively appear in the dialects, while being almost unknown in the standard variety. In this respect, a general improvement of the empirical foundations for the study of syntax can be achieved. To the extent that the sentence-based syntactic annotation is compatible with already available tools permitting detailed searches, CORDIAL-SIN syntactic annotation will ensure fast and efficient access to massive dialectal data, capable of responding to the different demands of specific research purposes within the domain of Portuguese dialect syntax.

References

ALE: *Atlas Linguarum Europae*, Van Gorcum, Assen Maastricht.
ALEAç: *Atlas Linguístico e Etnográfico dos Açores* (J. Saramago, coord.) (*http://www.clul.ul.pt/english/sectores/variacao/projecto_aleac.php*)
ALLP: *Atlas Linguístico do Litoral Português* (G. Vitorino, coord.) (*http://www.clul.ul.pt/english/sectores/variacao/projecto_allp.php*)
ALEPG: *Atlas Linguístico-Etnográfico de Portugal e da Galiza* (J. Saramago, coord.) (*http://www.clul.ul.pt/english/sectores/variacao/projecto_alepg.php*)
ASIS: *Syntactic Atlas of Northern Italy* (*http://asis-cnr.unipd.it*).
COSER: *Audible Corpus of Spoken Rural Spanish* (*http://www.uam.es/coser*).
EDISYN: *European Dialect Syntax* (*http://www.meertens.knaw.nl/edisyn*).
FRED: *Freiburg English Dialect Corpus* (*http://www2.anglistik.uni-freiburg.de/institut/lskortmann/FRED*).
Penn Parsed Corpora of Historical English (*http://www.ling.upenn.edu/hist-corpora*).
ScanDiaSyn: Scandinavian Dialect Syntax (*http://uit.no/scandiasyn*).
Syntactic Atlas of the Dutch Dialects (SAND, see Barbiers et al. 2006).
Tycho Brahe Parsed Corpus of Historical Portuguese (*http://www.ime.usp.br/~tycho/corpus*).

Auckle, T., Buchstaller, I., Corrigan, K. & A. Holmberg, 2007, «Speakers can "talk the talk", but can they "walk the walk" too?: Measuring syntactic variability using different instruments.» *Sixth meeting of the UK Language Variation and Change Conference* (UKLVC6), Lancaster University, September 2007.

Barbiers, S. et al., 2006, *Dynamic Syntactic Atlas of the Dutch dialects (DynaSAND)*, Meertens Institute, Amsterdam. (*http://www.meertens.knae.nl/sand*).

— & L. Cornips, 2002, «Introduction to Syntactic Microvariation», in S. Barbiers, L. Cornips & S. van der Kleij (eds.), *Syntactic Microvariation*, Meerteens Institute Electronic Publications in Linguistics. 2. (Available at: *http://www.merteens.knaw.nl/book/synmic*).

—, — & J. P. Kunst, 2007, «The Syntactic Atlas of the Dutch Dialects: A corpus of elicited speech and text as an on-line dynamic atlas», in J. C. Beal, K. C. Corrigan & H. Moisl (eds.), *Creating and digitizing language corpora. Volume 1: Synchronic databases*, Palgrave Macmillan, Hampshire, 54-90.

Bikel, D., 2004, *On the Parameter Space of Generative Lexicalized Statistical Parsing Models*, PhD diss. Univ. Pennsylvania, Philadelphia, PA. (Available at: *http://www.cis.upenn.edu/~dbikel/software.html* «Multilingual Statistical Parsing Engine»).

Bucheli, C. & E. Glaser, 2002, «The Syntactic Atlas of Swiss German Dialects: Empirical and Methodological Problems», in S. Barbiers, L. Cornips & S. van der Kleij» (eds.), *Syntactic Microvariation*, Merteens Institute Electronic Publications in Linguistics. (Available at: *http://www.merteens.knaw.nl/books/synmic/*).

Carrilho, E., 2005, *Expletive* ele *in European Portuguese Dialects*, PhD dissertation. University of Lisbon (Available at: *http://www.clul.ul.pt/equipa/ecarrilho/Carrilho2005.pdf*).

— & S. Pereira, 2006, «On the areal distribution of non-standard syntactic constructions in European Portuguese», paper presented at the *VIth Congress of Dialectology and Geolinguistics*, Univ. Maribor, Slovenia, September.

Collins, M., 1999, *Head-Driven Statistical Models for Natural Language Parsing*. Ph.D. dissertation, Univ. Pennsylvania, Philadelphia, PA.

Cornips, L., 2003, «Contact-induced Varieties, Syntactic Variation and Methodology», presented at *European Dialect Syntax ESF/SCH Explanatory Workshop*, Padova, September.

— & W. Jongenburger, 2001, «Elicitation techniques in a Dutch syntactic dialect atlas project», in: H. Broekhuis & T. van der Wouden (eds.), *Linguistics in The Netherlands* 2001, 18. John Benjamins, Amsterdam/Philadelphia.

— & C. Poletto, 2005, «On Standardizing Syntactic Elicitation Techniques (part 1)», *Lingua* 115, 939-957.

— & —, 2007, «Field linguistics meets formal research: How a microparametric view can deepen our theoretical investigation (sentential negation)», unpublished paper presented at *ICLaVE 4,* University of Cyprus, June.

Fernández-Ordoñez, I., 2009, «Dialect grammar of Spanish from the perspective of the *Audible Corpus of Spoken Rural Spanish* (or *Corpus Oral y Sonoro del Español Rural, COSER*)», *Dialectologia* 3, 23-51. (Available at: *http://www.publicacions.ub.es/revistes/dialectologia3).*

Finger, M., 1998, «Tagging a Morphologically Rich Language», in *Proceedings of the First Workshop on Text, Speech and Dialogue (TSD'98)*, Brno, Czech Republic, 39-44.

Galves, C. & H. Britto, 1999, «A construção do *Corpus Anotado do Português Histórico Tycho Brahe*: o sistema de anotação morfológica», in I. Rodrigues & P. Quaresma (eds.), *Proceedings of the IV PROPOR*, Universidade de Évora, Évora, 55-67.

Gottschalk, M. F., M. da G. Themudo Barata & J. V. Adragão, 1974, «Introdução», *Questionário Linguístico,* Instituto de Linguística, Lisboa.

Kortmann, B., 2002, «New Prospects for the Study of English Dialect Syntax: Impetus from Syntactic Theory and Language Typology», in S. Barbiers, L. Cornips & S. van der Kleij (eds.), *Syntactic Microvariation*, Merteens Institute Electronic Publications in Linguistics. (Available at: *http://www.merteens.knaw.nl/books/synmic/.*)

Kruijsen, J., 1983, «La Syntaxe dans l'*Atlas Linguarum Europae*», in C. Angelet, L. Melis, F. J. Mertens & F. Musarra (eds.) *Langue, Dialecte, Littérature. Études Romanes à la Mémoire de Hugo Plompteux,* Leuven U. P., Leuven. 213-223.

Labov, W., 1972, *Sociolinguistic Patterns*, University of Pennsylvania Press, Philadelphia.

—, 1996, «When Intuitions Fail», *Chicago Linguistics Society* 32. Parasession on Theory and Data in Linguistics, 76-106.

Lehmann, C., 2004, «Data in Linguistics», *The Linguistic Review* 21, 175-210.

Lehmann, W. P., 1980, «Dialect Investigations as Basis for Syntactic Study», in J. Kruijsen (ed.), *Liber Amicorum Weijnen*, AFA, Assen. 379-384.

Leite de Vasconcellos, J., 1901, *Esquisse d'une Dialectologie Portugaise*, Centro de Linguística da Universidade de Lisboa/Instituto Nacional de Investigaçã Científica, 3rd edition, 1987.

Lobo, M., 2008, «Variação morfo-sintáctica em dialectos do Português europeu: o gerúndio flexionado», *Diacrítica, Ciências da Linguagem*, Revista da Universidade do Minho, Braga, 22.1, 25-55.

Milroy, J., 1981, *Regional Accents of English: Belfast*, Blackstaff.

Milroy, L., 1987, *Observing and Analysing Natural Language: A Critical Account of Sociolinguistic Method*, Basil Blackwell, Oxford.

Randall, B., 2005-2007, CorpusSearch2 *(http://corpussearch.sourceforge.net).*

Remacle, L., 1952-1960, *Syntaxe du Parler Wallon de la Gleize*, Société d'Édition «Les Belles Lettres», Paris. vol. 1 (1952), vol. 2 (1956), vol. 3 (1960).

Schütze, C. T., 1996, *The Empirical Base of Linguistics: Grammatical Judgments and Linguistic Methodology*, University of Chicago Press, Chicago.
Segura, M. L., 1987, *A Fronteira Dialectal do Barlavento do Algarve*, Doctoral diss. CLUL.

Appendix: List of CORDIAL-SIN locations

01 VPA Vila Praia de Âncora (Viana do Castelo)
02 CTL Castro Laboreiro (Viana do Castelo)
03 PFT Perafita (Vila Real)
04 AAL Castelo de Vide, Porto da Espada, S. Salvador de Aramenha, Sapeira, Alpalhão, Nisa (Portalegre)
05 PAL Porches, Alte (Faro)
06 CLC Câmara de Lobos, Caniçal (Funchal)
07 PST Camacha, Tanque (Funchal)
08 MST Monsanto (Castelo Branco)
09 FLF Fajãzinha (Horta)
10 MIG Ponta Garça (Ponta Delgada)
11 OUT Outeiro (Bragança)
12 CVB Cabeço de Vide (Portalegre)
13 MIN Arcos de Valdevez, Bade, São Lourenço da Montaria (Viana do Castelo)
14 FIG Figueiró da Serra (Guarda)
15 ALV Alvor (Faro)
16 SRP Serpa (Beja)
17 LVR Lavre (Évora)
18 ALC Alcochete (Setúbal)
19 COV Covo (Aveiro)
20 PIC Bandeiras, Cais do Pico (Horta)
21 PVC Porto de Vacas (Coimbra)
22 EXB Enxara do Bispo (Lisboa)
23 TRC Fontinhas (Angra do Heroísmo)
24 MTM Moita do Martinho (Leiria)
25 LAR Larinho (Bragança)
26 LUZ Luzianes (Beja)
27 FIS Fiscal (Braga)
28 GIA Gião (Porto)
29 STJ Santa Justa (Santarém)
30 UNS Unhais da Serra (Castelo Branco)
31 VPC Vila Pouca do Campo (Coimbra)
32 GRJ Granjal (Viseu)
33 CRV Corvo (Horta)
34 GRC Graciosa (Angra do Heroísmo)
35 MLD Melides (Setúbal)
36 STA Santo André (Vila Real)
37 MTV Montalvo (Santarém)
38 CLH Calheta (Angra do Heroísmo)
39 CPT Carrapatelo (Évora)
40 ALJ Aljustrel (Beja)
41 STE Santo Espírito (Ponta Delgada)
42 CDR Cedros (Horta)

LE PROJET VIVALDI: PRÉSENTATION D'UN ATLAS LINGUISTIQUE PARLANT VIRTUEL

Roland Bauer
Université de Salzbourg

Abstract

The article deals with a speaking linguistic atlas named VIVALDI. *The project aims to cover all over Italy, with a choice of measuring-points, representing Romance as well as non-Romance dialects (like Albanian in Sicily or German in Southern Tyrol). The data is published in form of phonetic transcriptions accompanied by the corresponding sound-files collected by previous fieldwork. The questionnaire contains 359 single items covering all intra-linguistic categories and including a part of the parable of the lost son to be translated. Most of the questions are already documented by the Linguistic Atlas of Italy and Southern Switzerland (AIS 1928-1940) which opens the door to systematic diachronic studies. Actually (2009) data of the following nine regions are available: Aosta Valley (10 measuring-points), Friuli (11), Liguria (13), Molise (16), Piedmont (22), Sardinia (16), Sicily (17), Trentino-South Tyrol (33), Umbria (14). The whole* VIVALDI*-dataset (for now about 55.000 responses) is freely accessible.*

Key words: *Linguistic atlas, Italian dialects, Romance Linguistics, Dialectology.*

1. Introduction[1]

Comme le dit le titre de notre contribution, VIVALDI est un atlas linguistique parlant virtuel, donc disponible uniquement sous forme électronique ou sur le web, ou bien sur CD-ROM et sur DVD. Le son ainsi que les transcriptions phonétiques sont librement accessibles sur le site bilingue du projet, disponible en allemand et en italien, réalisé en langage *Java* et hébergé sur un serveur berlinois.[2]

En italien, l'acronyme VIVALDI se lit comme *VIVaio Acustico delle Lingue e dei Dialetti d'Italia*, ce qui correspond au français "VIVier Acoustique des Langues et des Dialectes de l'Italie". L'objectif central du projet consiste à créer une documentation acoustique du paysage dialectal actuel de l'Italie, y compris non seulement les dialectes gallo- et italoromans mais aussi des variétés alloglottes parlées dans les îlots linguistiques.

---

[1] J'adresse un grand merci à Lily Ditz-Fuhrich (Université de Salzbourg), qui a bien voulu se charger du contrôle stylistique de ce texte.

[2] <http://www2.hu-berlin.de/Vivaldi/>.

Fig. 1

Version française de la page d'accueil de l'atlas ladin parlant <ald.sbg.ac.at/ald/ald-i>

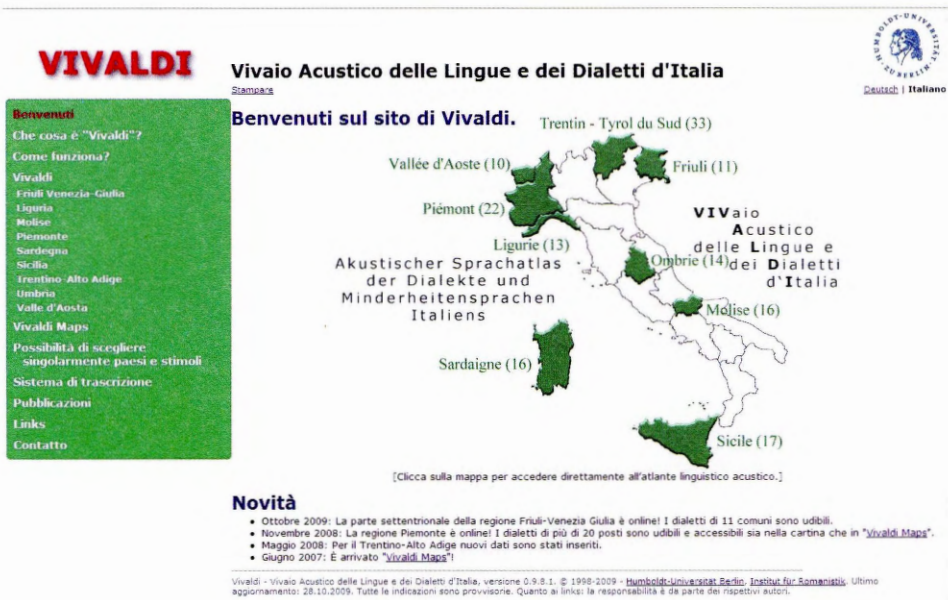


Fig. 2

Version italienne de la page d'accueil du projet VIVALDI <www2.hu-berlin.de/Vivaldi/>; en vert: régions déjà explorées (no. des localités disponibles en ligne)

Les matériaux de VIVALDI ont pour objet soit l'exploitation dialectologique traditionnelle (comme par exemple la comparaison diachronique avec des corpus géolinguistiques antérieurs) soit l'emploi didactique au niveau universitaire. A ce propos on a fait d'excellentes expériences surtout en ce qui concerne les réactions des étudiants. Mis à part le travail fondé exclusivement sur la documentation écrite des transcriptions phonétiques, travail plutôt aride, comme nous le savons tous, le contact complémentaire avec la réalité acoustique (i.e. naturelle) de nos dialectes semble avoir amélioré notamment la motivation des étudiants tant pour la dialectologie que pour la phonétique romanes. En même temps, leur intérêt pour les choses et les gens, c'est-à-dire pour les dialectes et pour leurs parlants, est également augmenté, de façon que plusieurs de nos étudiants ont choisi un paysage linguistique connu pour la première fois à travers VIVALDI soit comme argument pour préparer leur mémoire de maîtrise soit comme destination de stages de formation et/ou de vacances.

Notre projet atlantographique est né d'une coopération austro-allemande, initié au début de l'année 1992 par Dieter Kattenbusch et l'auteur de ces lignes (cf. Bauer 1995 et Kattenbusch 1995), et réalisé, depuis lors, par une équipe de romanistes de l'Université Humboldt de Berlin en collaboration avec le Département d'Etudes Romanes de l'Université de Salzbourg.[3]

VIVALDI prend comme modèle les expériences faites dans les années 80 du siècle passé au sein du projet ALD-I ("Atlas Linguistique du Ladin des Dolomites"), où l'on avait développé le prototype d'un premier atlas linguistique parlant, publié en 1991 sur un disque compact audio.[4] En contrepartie, l'ALD a pu s'appuyer sur la technologie-web développée par un membre du groupe VIVALDI (Marcel L. Müller, Université de Fribourg-en-Brisgau) pour réaliser une version de l'Atlas ladin parlant en ligne (cf. Müller 2008).[5]

2. Réseau d'enquête

Dans notre réseau d'enquête, chacune des 20 régions italiennes est représentée par environ une quinzaine de localités. Actuellement (décembre 2009), il y a neuf régions disponibles en ligne (voir aussi fig. 2). Par ordre chronologique, il s'agit de la Sicile (avec 17 enquêtes menées entre 1992 et 2008, cf. Bauer 1995 et Kattenbusch 2004), de la Sardaigne (avec 16 localités enquêtées entre 1999 et 2003, cf. Kattenbusch & Köhler 2004), de la Ligurie (13 points, enquêtes réalisées entre 2001 et 2002), de la Vallée d'Aoste (10 points, tous explorés en 2003, cf. Kattenbusch 2005), de l'Ombrie (14 points, 2003-2004), du Trentin et du Tyrol du Sud (avec 33 points d'enquête, 2005-2007), du Molise (16 points, 2005),[6] du Piémont (avec 22 enquêtes

---

[3] Cf. Müller & Köhler & Kattenbusch 2001 ainsi que Kattenbusch 2003.

[4] Cf. Bauer (1991) [CD-audio, édition bilingue, en italien et en allemand] et Bauer et al. (1990) [article complémentaire]. Pour la réédition de l'ALD-I parlant sur DVD cf. aussi Bauer & Goebl (2005) et Goebl & Bauer (2005). Pour l'atlas imprimé en sept volumes cf. Goebl, Bauer & Haimerl (1998).

[5] Pour la version française de la page d'accueil du site <http://ald.sbg.ac.at/ald/ald-i/> voir aussi fig. 1.

[6] 14 de ces enquêtes ont été réalisées par F. Tosques (Université Humboldt de Berlin), qui est en train de préparer une thèse de doctorat ("Geolinguistica molisana") sur le paysage linguistique du Molise. Voir aussi les réactions positives dans la presse du chef-lieu Campobasso (N.N. 2008) et sur le web <http://www.toro.molise.it >, lien <dialetto>.

réalisées entre 2007 et 2008) et enfin de la partie septentrionale (surtout carnique) du Frioul (avec 11 points d'enquête visités entre 2007 et 2009). Au total, on dispose actuellement de 152 localités accessibles sur le web, dont chacune est représentée par 359 épreuves dialectales acoustiques. A ce jour, VIVALDI met donc à disposition déjà environ 55.000 réponses dialectales.

Fig. 3

Points d'enquête siciliens dans le réseau VIVALDI (astérisque rouge: îlot linguistique albanais, astérisques jaunes: îlots linguistiques gallo-italiens)

Jetons maintenant un coup d'œil sur le réseau de la première de nos régions, à savoir sur celui de la Sicile (voir à ce propos fig. 3). Quatre des 17 enquêtes y étant réalisées concernent des îlots linguistiques. Au sud de Palerme, et plus précisément à Piana degli Albanesi, c'est l'albanais qu'on parle à la place du dialecte sicilien.[7] La colonisation alloglotte de Piana remonte à la deuxième moitié du XVe siècle, quand, du fait de l'expansion turque, beaucoup d'Albanais, obligés de quitter leur pays, s'installèrent dans le midi de l'Italie et en Sicile. Dans les villages de San Fratello, San Piero Patti et Aidone (cf. Raccuglia 2003 [dictionnaire du dialecte d'Aidone]), par contre, c'est le gallo-italien qui est parlé comme basilecte local.[8] Le gallo-italien fut importé par des colons de l'Italie du Nord (provenant surtout de la Ligurie et du Piémont) appelés par les Normands qui dominaient la Sicile au XIIe et au XIIIe siècle.[9]

---

[7] Voir l'astérisque rouge sur la fig. 3. Cf. aussi Birken-Silverman (1989).

[8] Voir les astérisques bleus sur la fig. 3. N.B.: L'enquête de San Piero Patti (réalisée par nos soins en été 2008) n'est pas encore intégrée dans la version en ligne.

[9] Pour une vue d'ensemble des colonies gallo-italiques de la Sicile cf. Trovato (1998).

La fig. 4 regroupe trois épreuves de transcriptions de trois différents dialectes de la Sicile, à savoir de l'albanais et du gallo-italien, toujours en comparaison avec le sicilien parlé dans le chef-lieu, c'est-à-dire à Palerme.[10] On y trouve les transcriptions phonétiques des réponses pour *l'acqua è calda* ("l'eau est chaude"), pour *il cane è bello* ("le chien est beau") et pour *non dormo mai prima di mezzanotte* ("je ne dors jamais avant minuit").

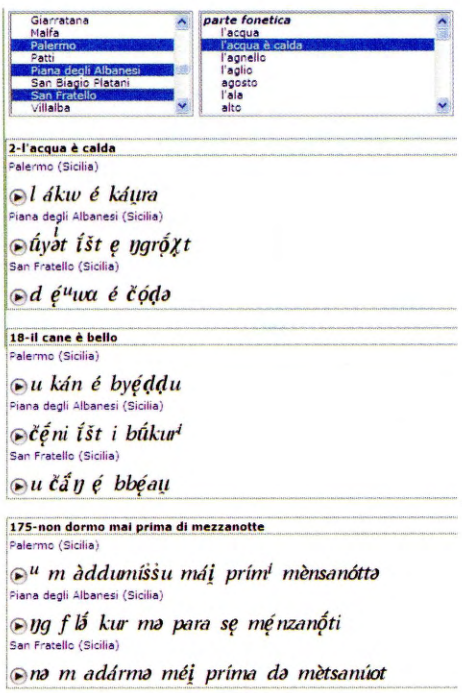


Fig. 4

Epreuves de transcriptions de trois différents dialectes de la Sicile: sicilien (Palerme), siculo-albanais (Piana degli Albanesi) et gallo-italien (San Fratello)

Du point de vue de la phonétique historique, on y note, par exemple dans les réponses de San Fratello, la présence de la palatalisation de C^A en [č] (= affriquée sourde post-palatale), donc d'une évolution retenue typique pour la Gallo-Romania, tandis que les dialectes siciliens maintiennent l'occlusive sourde vélaire [k]:

(1) CÁL(I)DA > gallo-italien [čǫ́ḍə] "chaude" vs. sicilien [káu̯ra]
(2) CÁNE > gallo-italien [čấŋ] "chien" vs. sicilien [kán]

De l'autre coté, on peut observer le phénomène de la rétroflexion du nexus lat. -LL- en [ḍḍ] (= occlusive sonore rétroflexe ou cacuminale), trait retenu typique pour

[10] Piana degli Albanesi: enquête D. Kattenbusch (1992), transcriptions G. Birken-Silverman (Université de Mannheim); San Fratello et Palerme: enquêtes et transcriptions R. Bauer (1994, 1999).

les parlers méridionaux de l'Italie et pour les dialectes septentrionaux sardes, cf. Rohlfs (1966: 328-333):

(3) BÉLLU > sicilien [byę́ḍḍu], sarde [bę́ḍḍu] vs. gallo-italien [bbę́aụ] "beau"

En ce qui concerne l'hétérogénéité lexicale de l'albanais de Piana, caractérisé par des souches autochtones ainsi que par un certain nombre de grécismes, de latinismes et d'italianismes, on remarque, au début de la deuxième phrase l'utilisation du latinisme *qeni* et, à la fin de la troisième phrase, la présence d'un emprunt au sicilien ou bien à l'italien *mezzanotte*:

(4) lat. CÁNE > siculo-albanais [č́ę́ni] "chien"
(5) MEDIANÓCTE > sicilien [mènsanóttə] > siculo-albanais [mènzanǫ́ti] "minuit"

3. Questionnaire

Quant au questionnaire de VIVALDI, on a essayé de le concevoir en majeure partie sur la base de celui de l'*Atlas Italo-Suisse* (AIS), publié par Jaberg et Jud entre 1928 et 1940, pour permettre une confrontation des données de l'AIS, recueillies à partir des années 20 du siècle passé, avec la documentation actuelle de VIVALDI.

Sous forme de mots isolés, d'expressions ou de petits syntagmes, les 359 questions qui figurent dans la version actuelle du questionnaire (2009) concernent toutes les catégories intralinguistiques. On y trouve une partie phonétique contentant 285 questions, 17 concepts sont dédiés au lexique, 18 questions concernent la morphologie et 15 exemples touchent la syntaxe. En plus, il y a une section de traduction qui compte 24 parties de la parabole du Fils Prodigue. Pour un extrait du questionnaire voir fig. 5. Les réponses basilectales qui y sont reproduites remontent à une enquête effectuée en été 2006 à Malfa, petit village situé sur une des îles éoliennes (Salina) juste au nord de la Sicile. Le dialecte éolien est, entre autres, caractérisé par la présence de traits phonétiques inconnus dans les dialectes limitrophes de la Sicile

164	mangiare	manggāri
165	la mano	a mānu
166	le mani	i māni
167	martedì	mārtedè
168	il martello	u martię́ḍḍu
169	marzo	mārtsu
170	**il maschio**	u māskulu

Fig. 5

Questionnaire-VIVALDI de Malfa (Ile de Salina), extrait de la partie phonétique

septentrionale et, parfois, à tort interprétés comme influences de la part des parlers napolitains, comme par exemple la diphtongaison de Ĕ tonique, cf. Fanciullo (1995):

(6) MARTĔLLU > éolien [martɪ̯ɛ̯ḍḍu] vs. messinais [martɛ̯ḍḍu] "marteau"

4. Transcription phonétique

Pour faciliter les analyses diachroniques, le système de transcription de VIVALDI, lui aussi, prend comme modèle celui de l'*Atlas Italo-suisse* (AIS), également utilisé dans les sept volumes de l'*Atlas Ladin* (ALD-I) publié en 1998.[11]

Sur les pages-web de VIVALDI on trouve tous les signes spéciaux de notre transcription phonétique accompagnés d'un bref commentaire. Fig. 6 fait apparaître un choix des signes consonantiques. En bas du tableau on reconnaît la consonne rétroflexe sonore typique de la Sicile [ḍ], rencontrée dans le commentaire à l'exemple (3) et qu'on retrouve également dans l'exemple (6) cité ci-dessus. Fig. 7 reprend les signes diacritiques les plus importants utilisés pour la distinction des voyelles, comme par exemple l'accent aigu (voyelles toniques), le tilde (nasalisation) ou bien les signes qui marquent la quantité vocalique.

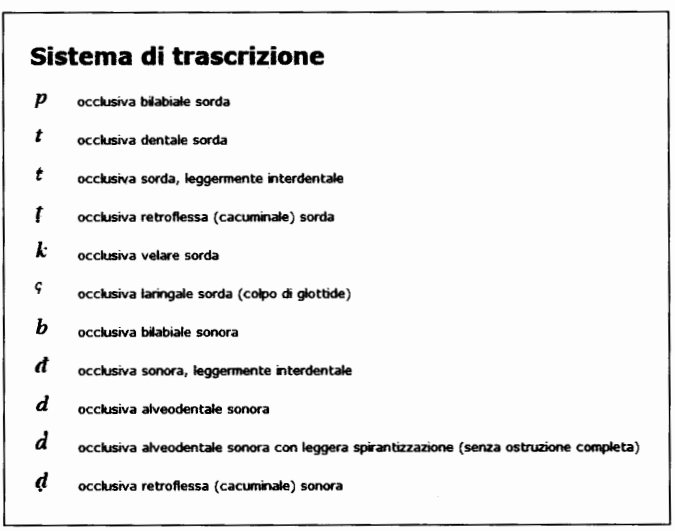


Fig. 6

Système de transcription de VIVALDI, extrait de la partie consonantique

[11] Cf. les explications données par Jaberg et Jud dans l'introduction à l'AIS (1928: 24-36). Cf. aussi ALD-I, vol. I, XVI et XXIV-XXV. Pour les problèmes de transcription appliqués au projet VIVALDI cf. Kattenbusch (2008).

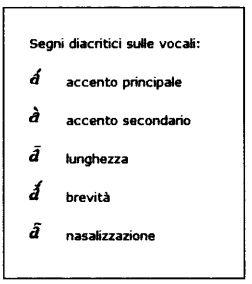


Fig. 7

Système de transcription de VIVALDI, signes diacritiques vocaliques

5. Détails techniques et informatiques[12]

Lors des travaux sur le terrain, les enregistrements acoustiques se font, entretemps, à l'aide d'un lecteur minidisque portable ou bien d'un lecteur enregistreur pour cartes mémoires (du type <cf>, *"compact flash"*). Pendant l'interview, on essaye d'éviter les interruptions d'enregistrement, afin d'obtenir un seul fichier-audio d'une durée de deux à quatre heures.

Pour transmettre les données du minidisque au disc dur de l'ordinateur, on se sert du programme *SonicStage* de Sony,[13] qui permet le choix de plusieurs formats audio. Il convient de choisir le format <pcm-wav> *("Pulse Code Modulation-Waveform")*, étant donné que le format <oma> *("OpenMG Music Format")*, proposé automatiquement par Sony, aboutit à la production de fichiers cryptés qui peuvent être modifiés uniquement à l'aide d'un logiciel Sony.

Les fichiers <wav> de départ (qui correspondent, nous le rappelons, à plusieurs heures de séance d'enregistrement) doivent être élaborés, c'est-à-dire découpés en segments à l'aide d'un programme approprié. Le découpage en segments sert à isoler les réponses basilectales qui correspondent à chacune de nos questions. Nous avons fait de bonnes expériences avec l'éditeur *Goldwave*,[14] à l'aide duquel les fichiers découpés peuvent être stockés ou en format <wav> ou bien directement en format <mp3>.

La représentation des données acoustiques sous forme d'oscillogramme, une fonction de base de *Goldwave*, comporte des avantages considérables pour le travail plutôt épineux de segmentation. D'une part, on y reconnaît assez facilement des détails consonantiques quantitatifs, comme par exemple les doubles par rapport aux consonnes simples. Sur la fig. 8 on voit l'oscillogramme de *cotto* ("cuit") dans un dialecte de la Sardaigne, y compris (en bleu) un [tt] géminé d'une durée de 240 ms environ. Sur la fig. 9 on en voit la variante simple, à savoir un [t] de 130 ms environ contenu dans la variante sarde du numéral *venti* ("vingt") dans la prononciation [bínti].

---

[12] Des remerciements particuliers vont à Fabio Tosques de l'Université Humboldt de Berlin, qui m'a gentiment fourni toutes les données techniques de cette présentation.

[13] <http://support.sony-europe.com/dna/downloads/downloads.aspx?l=de&f=sstage_dl>.

[14] <http://www.goldwave.com/>.

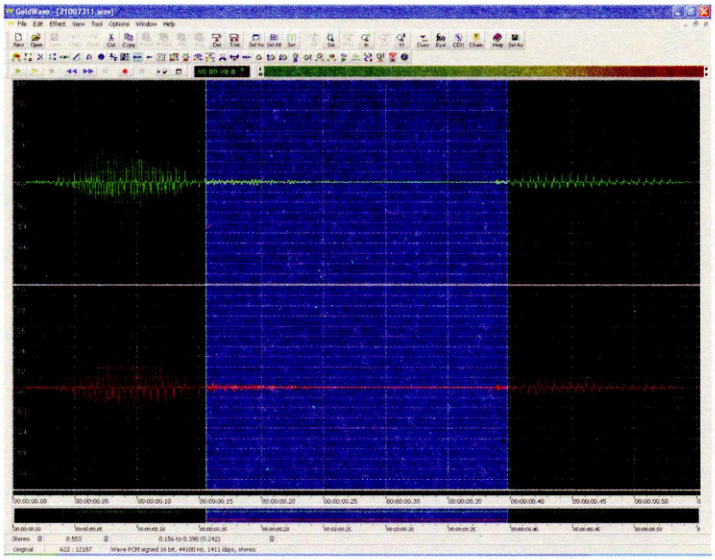


Fig. 8

Oscillogramme pour *cotto* [kóttu] ("cuit"); en bleu [tt] d'une durée de 240 ms (enquête D. Kattenbusch, Laconi, Sardaigne 1999)

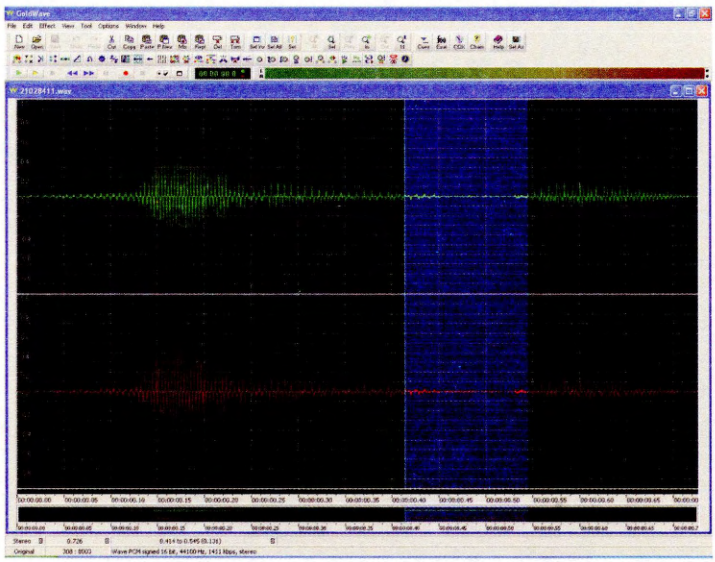


Fig. 9

Oscillogramme pour *venti* [bínti] "vingt"; en bleu [t] d'une durée de 130 ms (enquête D. Kattenbusch, Laconi, Sardaigne 1999)

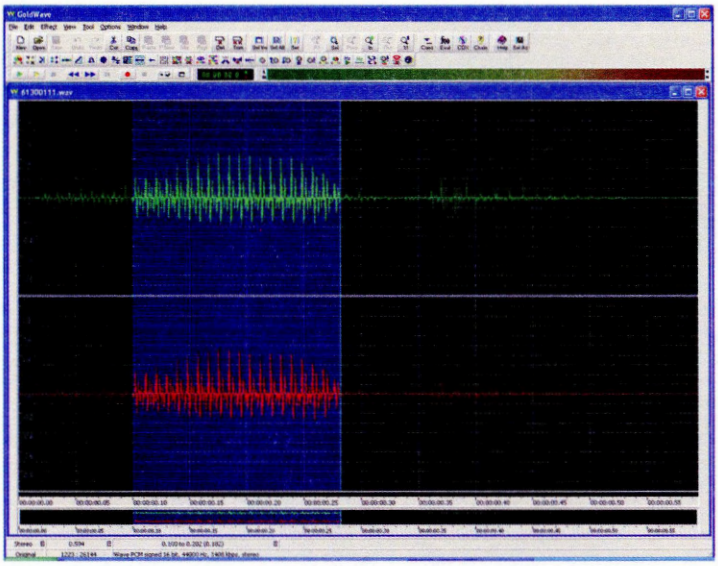


Fig. 10

Oscillogramme pour *l'acqua* [l ega] "l'eau"; en bleu [e] d'une durée de 180 ms
(enquête D. Kattenbusch, San Martin de Tor, Ladinia, Tyrol du Sud 1999)

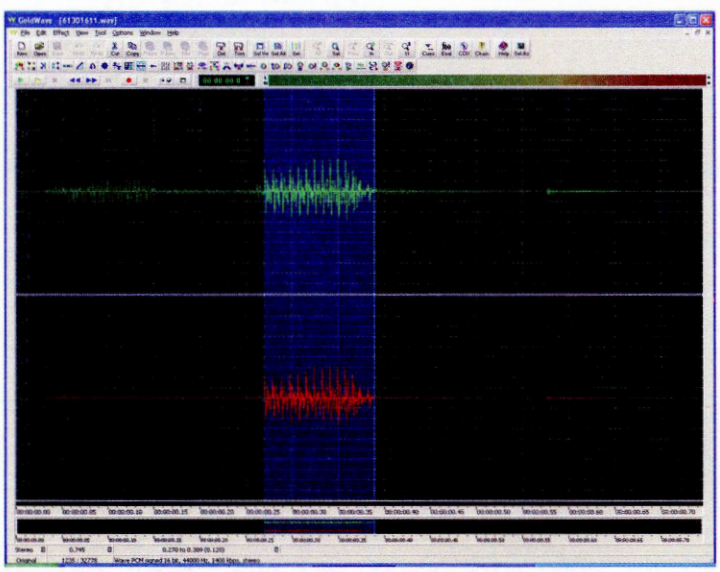


Fig. 11

Oscillogramme pour *il becco* [l bek] "le bec"; en bleu [e] d'une durée de 120 ms
(enquête D. Kattenbusch, San Martin de Tor, Ladinia, Tyrol du Sud 2006)

D'autre part, l'oscillogramme facilite la détermination des voyelles longues et brèves. La fig. 10 représente, par exemple, la courbe qui correspond à la prononciation de *l'acqua* ("l'eau") dans le dialecte ladin de San Martin de Tor au Tyrol du Sud, y compris (toujours soulignée en bleu) la voyelle [e] longue d'une durée de 180 ms environ. Sur la fig. 11 on en voit une variante plus brève, à savoir un [e] de 120 ms environ contenu dans la variante ladine de *becco* ("bec").

Pour ce qui est des paramètres techniques de nos fichiers acoustiques, on distingue entre une version <wav> de base avec 1.411 kbps (Codec 16-bit PCM Audio non comprimé, stéréo, 44.100 Hz) et une version plus "légère", pour ainsi dire, destinée à la publication sur le web et réduite à 96 kbps (Codec MPEG 1 Audio, Layer 3 [MP3], stéréo, 44.100 Hz).

Parallèlement à la segmentation du son, on procède à l'entrée des transcriptions phonétiques dans un fichier *Excel*, qui servira de base pour la création d'une banque de données *SQL* (*"Structured Query Language"*) (voir fig. 12). Pour la saisie des transcriptions, on se sert d'un système de codage alphanumérique développé, il y a environ 20 ans, au sein de l'*Atlas linguistique ladin* (cf. Bauer et al. 1988: 31-42). C'est ainsi que le code <ag\2d\6j\1t> (mis en évidence, sur la fig. 12, par un soulignement rouge) correspond à [a], [g], [ǫ́], [ʂ], [t], donc à la prononciation [agǫ́ʂt] pour désigner le mois d'*août*. Dans ce cas là, il s'agit d'un dialecte piémontais (voir la légende de la fig. 12).

À l'intérieur du fichier *Excel*, on dispose de plusieurs programmes qu'on appelle VBA (*"Visual Basic for Applications"*) et qui garantissent, entre autres, la transposition des codes alphanumériques dans les polices de caractères phonétiques appropriées.

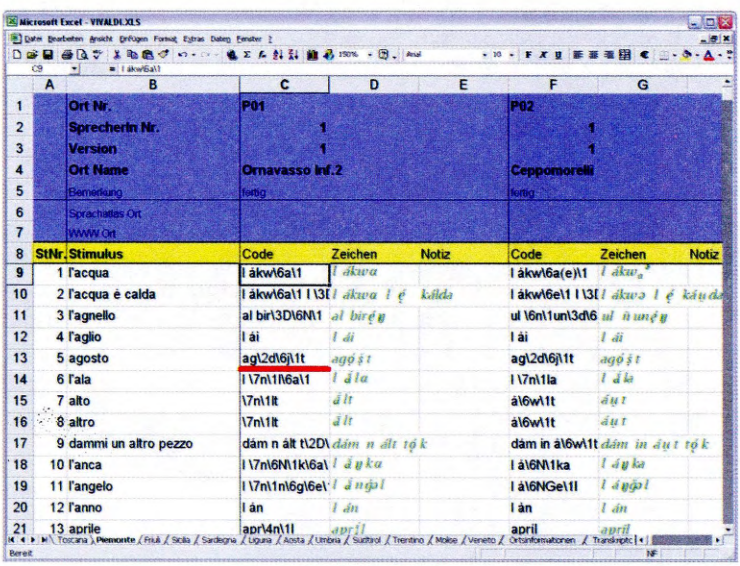


Fig. 12

Extrait d'un fichier *Excel* de VIVALDI
(enquêtes D. Kattenbusch, Ornavasso et Ceppomorelli, Piémont 2007)

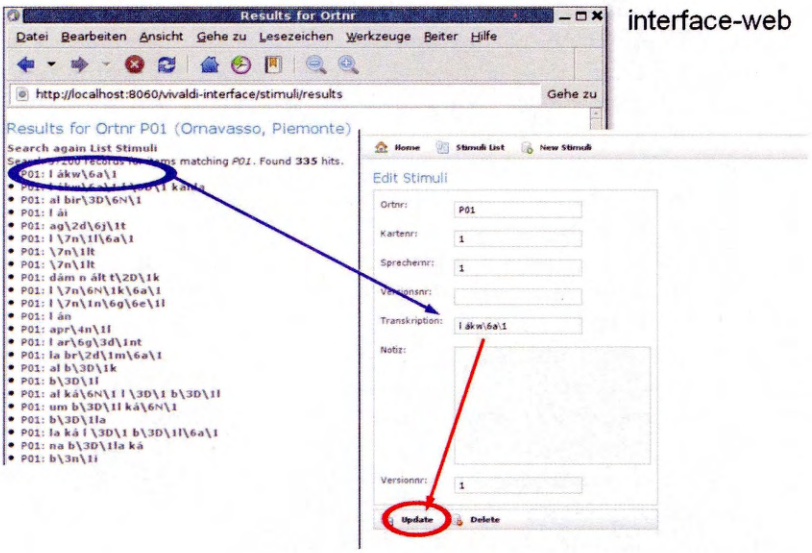


Fig. 13

Extrait du module de correction de VIVALDI
(enquête D. Kattenbusch, Ornavasso, Piémont 2007)

Un deuxième VBA est responsable de l'exportation des tableaux pour les besoins de la banque de données. En dernier lieu, on y peut exporter des fichiers <html> (*"HyperText Markup Language"*) pour les futures pages-web, puis encore des fichiers de contrôle en format <rtf> (*"Rich Text Format"*) ainsi que les fichiers <gif> (*"Graphics Interchange Format"*) qui contiennent les transcriptions phonétiques complètes sous forme de photo.

S'il faut apporter des corrections aux données phonétiques saisies et transmises à la banque de données, on ne recourt plus au système *Excel*. Dans ce cas là, on se sert d'une interface basée sur le web, qui n'est disponible qu'au niveau interne. On y choisit d'abord le numéro ou bien le nom d'une localité pour obtenir une liste de toutes les transcriptions saisies pour le point d'enquête en question. La fig. 13 montre une partie des résultats pour la localité piémontaise no. 1, Ornavasso. C'est en cliquant sur une des transcriptions que l'on reçoit une nouvelle fenêtre permettant l'édition et donc la correction des données saisies. En appuyant sur le bouton <update>, toutes les corrections apportées seront, à leur tour, stockées dans la banque de données centrale.

6. VIVALDI en ligne

La version en ligne de VIVALDI nécessite toute une série de logiciels, tous gratuits et libres, qui fonctionnent en parallèle, à savoir:

1. un serveur web ou bien http *Apache* (en version 2.*),

2. un serveur de base de données *MySQL* (en version 5.*),
3. le module *GD* de *PHP* (qui garantit la gestion des transcriptions comme images),[15]
4. le système *Java* (version 1.4 ou supérieure) pour le chargement d'un applet dans le navigateur utilisé,
5. plusieurs bibliothèques *Java* qui s'occupent du décodage des fichiers acoustiques mémorisés en <mp3>.[16]

En plus, on se sert de quelques programmes réalisés à l'intérieur du groupe VIVALDI, parmi lesquels on retrouve:

6. des scripts en *PHP* (pour la visualisation correcte des informations recherchées dans la banque de données),
7. un applet en *Java* (pour l'écoute des transcriptions visualisées sur la carte) et
8. une interface de contact pour accéder au monde de *Google Maps* (cf. infra).

En ce qui concerne les transcriptions elles-mêmes, on se sert des polices de caractères <ttf> (*"True Type Font"*) développées à Salzbourg au sein de l'*Atlas linguistique ladin*, auxquelles on a apporté de petites modifications (voir fig. 14). Les images des transcriptions, visualisées sur le web, sont stockées en format <gif> avec une résolution de 72 × 72 points par pouce et en 256 couleurs.

Fig. 14

Police de caractères <ald_6.ttf> pour la représentation du consonantisme

Passons au fonctionnement et à la consultation de VIVALDI en ligne. Après avoir choisi une région, le système passe d'abord à la visualisation des informations relatives aux localités explorées et relatives aux enquêtes. La fig. 15 reprend les données qui se réfèrent à trois enquêtes siciliennes menées, entre décembre 1993 et juillet 2006, à Giarratana, à Malfa et à Palerme. A part le nom de la localité, on y trouve: la province et la région d'appartenance, les numéros des points de référence contenus dans les atlas linguistiques italiens nationaux ALI et AIS, des liens à *Wikipedia* et à *Google Maps*, quelques données biographiques de l'informateur (comme le sexe,

[15] <http://www.php.net/pd>.

[16] <jl1.0.jar>, <tritonus_share.jar>, <mp3spi1.9.jar> (cf. <www.javazoom.net>).

l'année de naissance et la profession) ainsi que la date de l'enquête et le nom de l'enquêteur.

Giarratana

Località	Provincia	Ragusa (RG)
	Regione	Sicilia
	Atlante linguistico/Punto	AIS: 896
	Dialetto/Lingua	Siciliano
	Internet	http://it.wikipedia.org/wiki/Giarratana
	Ulteriori Informazioni	Wikipedia 'Giarratana', Google Maps 'Giarratana'
Informatore	Genere	m
	Anno di nascita	1942
	Professione	Muratore
Registrazione	Data	12/1993
	Esploratore	Dieter Kattenbusch

Malfa

Località	Provincia	Messina (ME)
	Regione	Sicilia
	Atlante linguistico/Punto	ALI: 1000
	Dialetto/Lingua	Siciliano
	Internet	http://it.wikipedia.org/wiki/Malfa
	Ulteriori Informazioni	Wikipedia 'Malfa', Google Maps 'Malfa'
Informatore	Genere	m
	Anno di nascita	1935
	Professione	Ragioniere
Registrazione	Data	07/2006
	Esploratore	Roland Bauer

Palermo

Località	Provincia	Palermo (PA)
	Regione	Sicilia
	Atlante linguistico/Punto	AIS: 803, ALI: 1004
	Dialetto/Lingua	Siciliano
	Internet	http://it.wikipedia.org/wiki/Palermo
	Internet	http://www.comune.palermo.it/
	Ulteriori Informazioni	Wikipedia 'Palermo', Google Maps 'Palermo'
Informatore	Genere	m
	Anno di nascita	1939
	Professione	s.i.
Registrazione	Data	11/1999
	Esploratore	Roland Bauer

Fig. 15

Informations relatives à trois enquêtes-VIVALDI menées en Sicile

En deuxième lieu, l'utilisateur de VIVALDI en ligne choisit une partie du questionnaire (par ex. la phonétique) et puis une question spécifique (par ex. *l'agnello* "l'agneau"). C'est ainsi que l'on active un applet écrit en *Java* qui conduit à la visualisation d'une carte interactive de la région désirée (par ex. de la Sicile). Il suffit alors de cliquer avec la souris sur une des localités figurant sur cette carte pour provoquer l'émission parallèle du son ainsi que de la transcription. Au-dessous de la carte elle-même, on retrouve la liste de toutes les réponses basilectales munies, chacune d'entre elles, d'un bouton d'écoute (voir fig. 16).

A part l'accès direct sur la carte et dans la liste, notre logiciel offre une troisième possibilité d'écoute. A ce propos, il faut d'abord choisir l'option *Possibilità di scegliere singolarmente paesi e stimoli* ("choix de localités et de questions") du menu central. Ensuite, on peut sélectionner les combinaisons préférées. Dans notre exemple (voir fig. 17), il s'agit de quatre localités ligures (Airole, Calizzano, Castelnuovo Magra et Gênes) et des signifiés *l'acqua* ("l'eau") et *l'agnello* ("l'agneau"). Comme

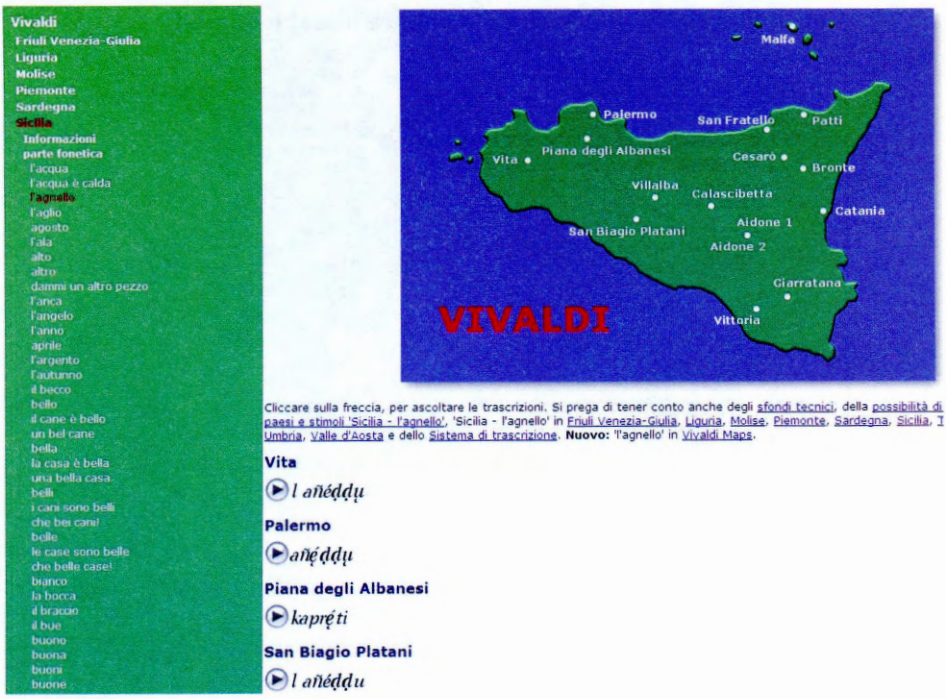


Fig. 16

Ecran-VIVALDI, réseau de la Sicile, choix du corpus phonétique, choix de la question *l'agnello* "l'agneau", en bas: transcriptions et boutons d'écoute

résultat on obtient des listes composées du nom de la localité, de la transcription phonétique y ayant trait et d'un bouton d'écoute. Le nom du point d'enquête est en même temps un lien, dont l'activation conduit à la carte interactive de la région en question.

Une possibilité plutôt récente pour accéder aux données VIVALDI est représentée par le module *VIVALDI Maps*, qui s'appuie sur la plateforme bien connue de *Google Maps*. Après avoir agrandi la zone désirée, l'utilisateur de *VIVALDI Maps* choisit la question en haut de l'écran. L'écran représenté sur la fig. 18 correspond au choix de la phrase *Se l'avessi saputo sarei venuto* ("Si je l'avais su, je serais venu"), qui nous informe, entre autres, sur l'utilisation du subjonctif, de l'indicatif ou bien du conditionnel dans les dialectes de l'Italie. Dans notre réponse exemplaire [se ɪ avíssə sapútu avíssə vənútu] on note, au sujet du dialecte sicilien de Malfa, le double emploi du subjonctif passé, soit dans la proposition principale soit dans sa subordonnée. De l'autre côté, on y remarque, contrairement au standard italien, l'emploi de l'auxiliaire *avere* ("avoir") avec le verbe de mouvement *venire* ("venir").

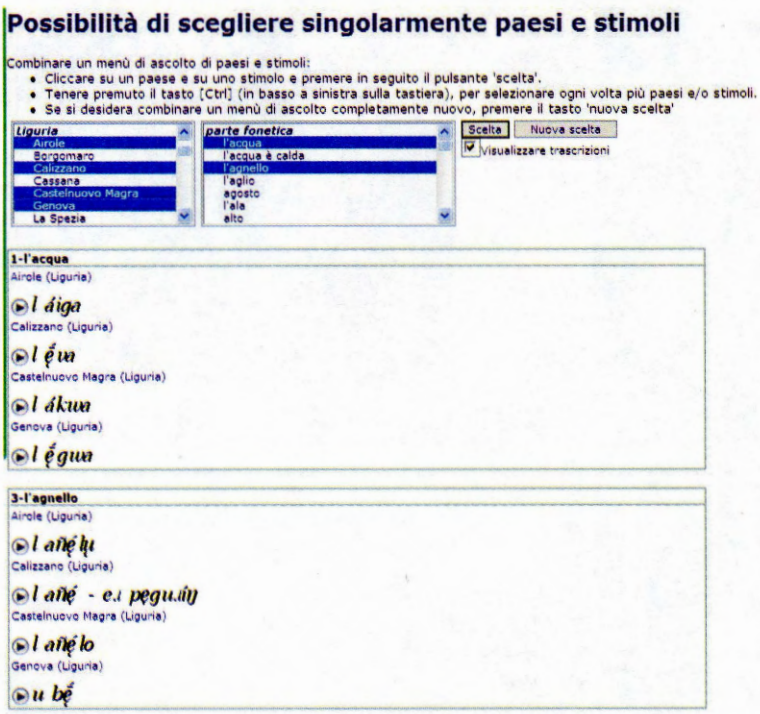


Fig. 17

Ecran-VIVALDI, choix des localités ligures Airole, Calizzano (enquêtes C. Köhler 2001), Castelnuovo Magra et Gênes (enquêtes D. Kattenbusch 2001 et 2002), choix des questions *l'acqua* "l'eau" et *l'agnello* "l'agneau"

7. VIVALDI hors ligne

Encore deux mots sur la version hors ligne de VIVALDI sur DVD. Le disque, réalisé à l'aide du logiciel gratuit *Server2Go*,[17] contient toutes les ressources présentées ci-dessus, sauf *VIVALDI-Maps*. On y trouve tous les fichiers nécessaires, comme par exemple les classes *Java*, les transcriptions phonétiques en format <gif>, les pages en <html>, les cartes de base, toujours en <gif>, et, bien évidemment, tous les fichiers audio en <mp3>.

Après avoir inséré le DVD dans le lecteur, le programme démarre automatiquement tout en simulant un serveur web et une banque de données à niveau local. Ainsi, on ne notera pas de différences entre la version web décrite ci-dessus et la version hors ligne.[18]

---

[17] <www.server2go-web.de/>.

[18] Un exemplaire gratuit du DVD peut être commandé à l'adresse suivante: <dieter.kattenbusch@romanistik.hu.berlin.de>.

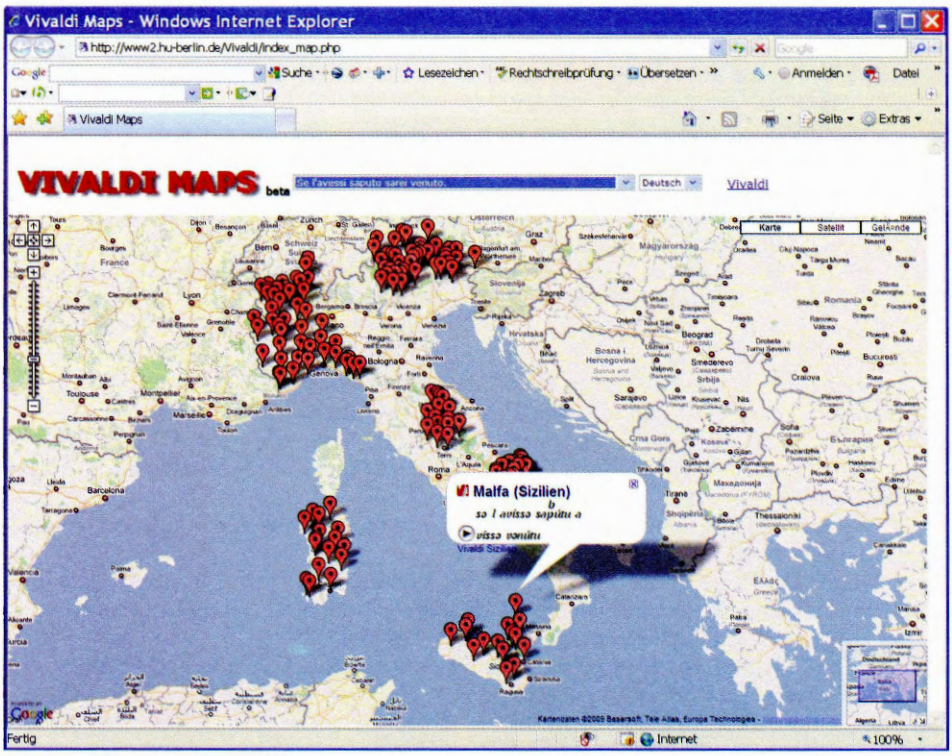


Fig. 18

Ecran-*VIVALDI-Maps* <http://www2.hu-berlin.de/Vivaldi/index_map.php>, choix de la question *Se l'avessi saputo sarei venuto* ("Si je l'avais su, je serais venu"), localité Malfa, Ile de Salina, Sicile (enquête R. Bauer 2006)

Références

AIS: cf. Jaberg & Jud (1928-1940).
ALD-I: cf. Goebl, Bauer & Haimerl (1998).
ALI: cf. Bartoli et al. (1995-).
Bartoli, M. et al. (eds.), 1995-, *Atlante linguistico italiano*, Istituto Poligrafico e Zecca dello Stato, Roma.
Bauer, R., 1991, *ALD-I-CD: 98 campioni fonici per l'ALD-I / 98 Tonproben zum ALD-I,* Institut für Romanistik, Salzburg.
—, 1995, «VIVALDI-Sicilia. Documentazione sonora dei dialetti siciliani», in G. Ruffino (ed.), *Percorsi di geografia linguistica. Idee per un atlante siciliano della cultura dialettale e dell'italiano regionale*, Centro di Studi Filologici e Linguistici Siciliani, Palermo, 543-550.
— & H. Goebl, 2005, «L'atlante ladino sonoro. Presentazione del modulo acustico dell'ALD-I (con alcune istruzioni per l'installazione e per l'uso del DVD allegato)», *Mondo ladino* 29, 37-66.

— et al., 1988, «Arbeitsbericht 3 zum ALD-I / Relazione di lavoro 3 per l'ALD-I», *Ladinia* XII, 17-56.
— et —, 1990, «Arbeitsbericht 5 zum ALD-I / Relazione di lavoro 5 per l'ALD-I», *Ladinia* XIV, 259-304.
Birken-Silverman, G., 1989, *Phonetische, morphosyntaktische und lexikalische Varianten in den palermitanischen Mundarten und im Sikuloalbanischen von Piana degli Albanesi,* Egert, Wilhelmsfeld.
Fanciullo, F., 1995, «Sulla posizione dialettale delle Eolie», in S. Todesco (ed.), *Atlante dei Beni Etno-Antropologici eoliani,* Regione Siciliana, Palermo, 101-113.
Goebl, H. & R. Bauer, 2005, «Der "Sprechende" Ladinienatlas. Vorstellung des akustischen Moduls des ALD-I samt Hinweisen zur Installation und Benützung der beiliegenden DVD», *Ladinia* XXIX, 125-154.
— & Bauer, R. & E. Haimerl (eds.), 1998, *ALD-I: Sprachatlas des Dolomitenladinischen und angrenzender Dialekte I / Atlant linguistic dl ladin dolomitich y di dialec vejins I / Atlante linguistico del ladino dolomitico e dei dialetti limitrofi I,* Reichert, Wiesbaden, 7 voll.
Jaberg, K. & J. Jud (eds.), 1928, *Der Sprachatlas als Forschungsinstrument. Kritische Grundlegung und Einführung in den Sprach- und Sachatlas Italiens und der Südschweiz,* Niemeyer, Halle/Saale.
— & — (eds.), 1928-1940, *Sprach- und Sachatlas Italiens und der Südschweiz,* Ringier, Zofingen, 8 voll.
Kattenbusch, D., 1995, «Atlas parlant de l'Italie par régions: Vivaldi», in N. N., *Estudis de lingüística i filologia oferts a Antoni M. Badia i Margarit.* Universitat de Barcelona: Departament de Filologia Catalana, Barcelona, volum I, 443-455.
—, 2003, «ALD-I und Vivaldi und die Segnungen der akustischen Sprachgeographie», *Quo vadis Romania?* 22, 22-30.
—, 2004, «Akustischer Sprachatlas Siziliens», in W. Dahmen et al. (eds.), *Romanistik und neue Medien. Romanistisches Kolloquium XVI,* Narr, Tübingen, 243-248.
—, 2005, «Diatopische Variation im Aostatal und ihre sprachgeographische Dokumentation», in P. Cichon et al. (eds.), *Entgrenzungen. Für eine Soziologie der Kommunikation. Festschrift für Georg Kremnitz zum 60. Geburtstag,* Edition Praesens, Wien, 279-284.
—, 2008, «Akustische Wirklichkeit und auditive Täuschungen. Wie realistisch kann eine Transkription sein?», in G. Blaikner-Hohenwart et al. (eds.), *Ladinometria. Festschrift für Hans Goebl zum 65. Geburtstag, vol. 2,* Fachbereich Romanistik et al., Salzburg et al., 179-187.
— & C. Köhler, 2004, «La Sardegna nel progetto Vivaldi», in L. Grimaldi & G. Mensching (eds.), *Su sardu. Limba de Sardigna e limba de Europa,* CUEC, Cagliari, 193-203.
Müller, M. L., 2008, «Digitale Sprachatlanten am Beispiel von Vivaldi und Ald-I. Interoperabilität durch die "Geolinguistic Document Architecture (GDA)"», in G. Blaikner-Hohenwart et al. (eds.), *Ladinometria. Festschrift für Hans Goebl zum 65. Geburtstag, vol. 1,* Fachbereich Romanistik et al., Salzburg et al., 291-305.
—, Köhler, C. & D. Kattenbusch, 2001, «Vivaldi: ein sprechender Sprachatlas im Internet als Beispiel für die automatisierte, computergestützte Sprachatlasgenerierung und-präsentation», *Dialectologia et Geolinguistica* 9, 55-68.
N. N., 2008, «Versione on line della lingua dialettale torese», *Il Quotidiano del Molise* XI/153, 9.
Raccuglia, S., 2003, *Vocabolario del dialetto galloitalico di Aidone,* Centro di Studi Filologici e Linguistici Siciliani, Palermo.
Rohlfs, G., 1966, *Grammatica storica della lingua italiana e dei suoi dialetti. Fonetica,* Einaudi, Torino.
Trovato, S. C., 1998, «I dialetti galloitalici della Sicilia», in G. Holtus & M. Metzeltin & C. Schmitt (eds.), *Lexikon der Romanistischen Linguistik. Vol. VII: Kontakt, Migration und Kunstsprachen. Kontrastivität, Klassifikation und Typologie,* Niemeyer, Tübingen, 538-559.

LE THESAURUS OCCITAN: UNE BASE DE DONNÉES MULTIMEDIA CONSACRÉE AUX DIALECTES OCCITANS

Guylaine Brun-Trigaud

Laboratoire BCL, CNRS UMR 6039
Université Nice Sophia-Antipolis, MSH de Nice (France)

Abstract

The Thesaurus Occitan *(abbreviated THESOC) is a multimedia database, which contains, among other things: linguistic and linguistic-related data from field works: maps and survey notebooks from the* Atlas linguistiques, *monographies, audio records, pictures; linguistic data coming from former analyses: lemmatisation, morphology, etymology, microtoponymy; bibliographical references; tools for linguistic analyses: maps generator, instruments for diachronic analyses, comparative cartography procedures, morphological analysis instruments; a Morpho-Syntax Module (MMS), detailed by Pierre-Aurélien Georges.*

Centralised in Nice (France) within the laboratory UMR 6039 «Bases, Corpus, Langage» (attached to the CNRS), this inter-university program associates different teams, upon the direction of Pr. Jean-Philippe Dalbera.

One can say the THESOC is a variable geometry database which considers all kinds of exploitations thanks to specific menus, which integrates all kinds of documents. The advantage of such a database lies also in the fact that it can evolve and be updated permanently to satisfy users' needs.

Key words: *Database, mapping, multimedia, dialectology, occitan languages.*

Fig. 1

Logo Thesoc

Le THESAURUS OCCITAN (ou THESOC en abrégé) est une base de données informatique destinée à l'étude des dialectes occitans.

Elle est développée depuis 1992 et centralisée à Nice dans le cadre de l'UMR 6039 du CNRS «Bases, Corpus, Langage», sous la direction de Jean-Philippe Dalbera, il s'agit d'un programme inter-universitaire qui associe différentes équipes.

Le THESOC contient notamment:

— des données linguistiques et péri-linguistiques issues d'enquêtes de terrain: cartes et carnets d'enquêtes des Atlas linguistiques,[1] monographies, enregistrements sonores, documents iconographiques;
— des données linguistiques procédant d'analyses déjà réalisées: lemmatisations, morphologie, étymologie, microtoponymie;
— des données bibliographiques;
— des outils d'analyse: représentations cartographiques, instruments d'analyse diachronique, procédures de cartographie comparative, instruments d'analyse morphologique.

Il s'agit d'un objet à géométrie variable qui envisage différents types d'exploitations grâce à des menus spécifiques, qui intègrent toutes sortes de documents, de telle sorte que le THESOC se présente comme un outil offrant à la fois, mais toujours séparément, des données linguistiques quasi brutes, des données ayant fait l'objet d'analyses et de traitements et des outils d'investigation. L'intérêt d'un tel outil réside également dans le fait qu'il est en permanence amené à évoluer selon les besoins des utilisateurs.

Pour pouvoir figurer dans la base, les données linguistiques brutes doivent satisfaire aux deux critères suivants: d'une part, elles doivent être de nature orale et sont donc saisies dans la base avec leur transcription phonétique en API, pour assurer le partage des données avec les autres chercheurs. De plus, comme le THESOC offre la possibilité de faire entendre les sons enregistrés au cours des enquêtes —lorsque nous les avons—, cela permet à l'utilisateur de contrôler la transcription proposée. D'autre part, les données doivent aussi être précisément localisées, ce qui constitue évidemment une condition essentielle pour l'étude de la variation diatopique.

Actuellement la base comprend plus d'un million de fiches-réponses, plus de mille extraits sonores et plus de 500 documents visuels (photos, vidéos, et dessins) accompagnent et illustrent les données linguistiques, ce qui fait du THESOC un véritable recueil multimédia, qui peut s'inscrire tout autant comme outil de recherche pour les linguistes que comme outil pédagogique pour le grand public.

Une partie des données est consultable en ligne sur Internet à l'adresse http://thesaurus.unice.fr

1. Les Glossaires

La partie lexicale constitue le «cœur historique» du THESOC, elle peut être consultée à partir d'une interface intitulée «Tableau de bord» (voir fig. 2) qui comporte différents modules:

[1] *Atlas Linguistiques de la France par régions*, Editions du CNRS.

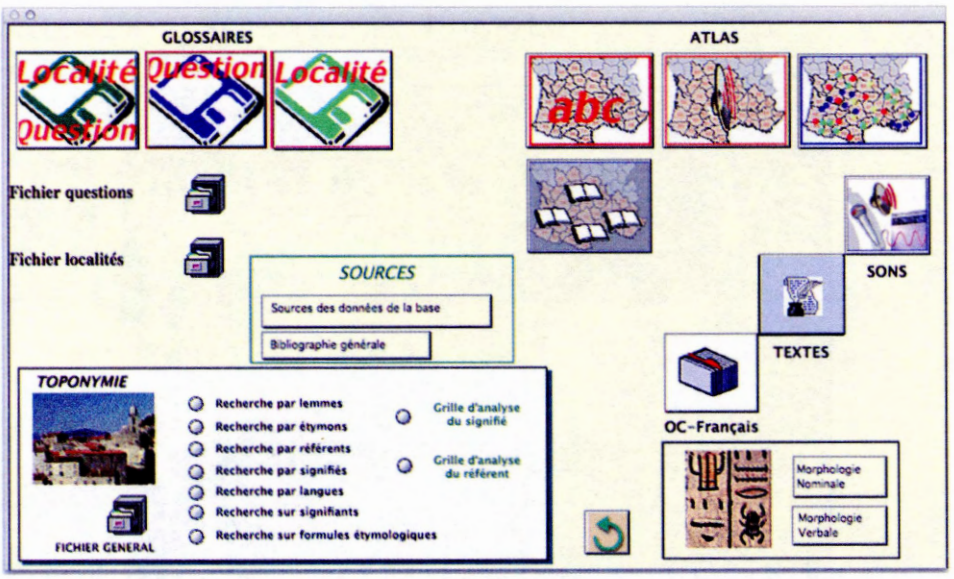


Fig. 2

Tableau de bord

1.1. Les fichiers localités et questions

À l'intérieur de la base de données, chaque forme est identifiée par le couple LOCALITÉ-QUESTION, dont voici les deux fichiers principaux: d'une part, le fichier des localités qui contient 831 entrées recouvrant tout le domaine occitan.

La consultation d'une fiche localité (voir fig. 3) permet d'avoir accès à la liste des enquêteurs et des informateurs associés aux différentes enquêtes qui se sont succédées dans cette localité.

D'autre part, le fichier des questions ou «responsaire» (cf. Oliviéri 2004) qui est le résultat de la somme des cartes et listes publiées par les différents atlas linguistiques régionaux du domaine occitan, ainsi que les éléments relevés dans des monographies ou des résultats d'enquêtes non publiées.

Il comporte 8.338 questions (voir fig. 4), qui sont regroupées suivant les principaux thèmes traités dans les atlas linguistiques régionaux, comme l'élevage, la nature, l'espace, le temps, l'habitat et la vie quotidienne, etc.

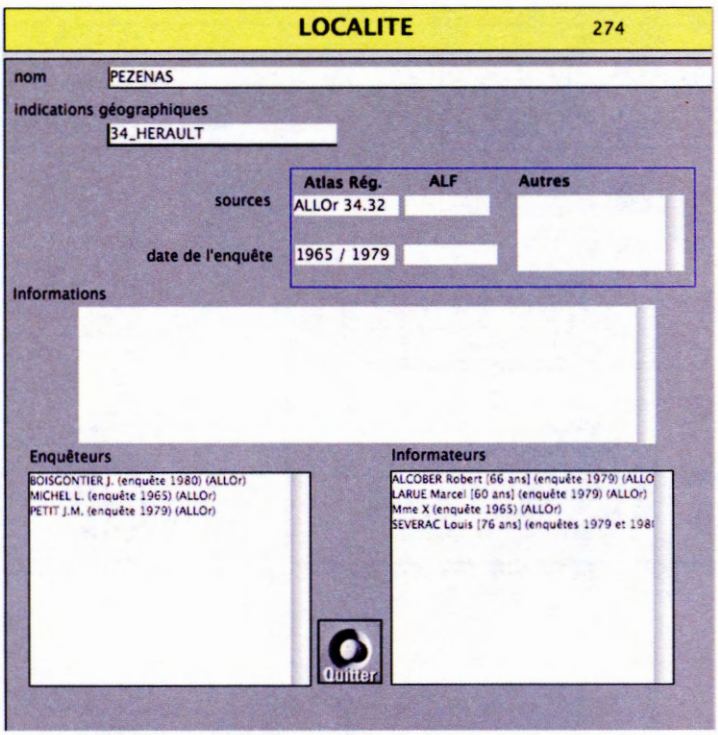


Fig. 3

Exemple de fiche-localité: Pezenas (Hérault)

Liste des questions 8338 fiches

N°	Intitulé	Scient.	Entrée d'index	Thème	Sous-thème
1	abandonner le nid		abandonner	non spécifique	
3	abasourdi		abasourdi	HOMME	Caractère, sentiments, jugement
4	abat-foin		abat-foin	CULTURES	Prés, foins, plantes textiles
5	abattre un arbre		abattre	NATURE	Forêt
6	abcès		abcès	HOMME	Maladies, affections
7	abeille		abeille	ELEVAGE	Ruches
8	abeille sauvage		abeille	NATURE	Animaux sauvages, insectes
9	abîmer		abîmer	non spécifique	
10	ablette		ablette	NATURE	Animaux sauvages, insectes
11	aboyer		aboyer	ELEVAGE	Chiens, chats

Fig. 4

Liste des fiches-questions

La consultation d'une fiche question (voir fig. 5) permet d'avoir accès à la liste des cartes publiées et des carnets d'enquêtes, concernant cette question. Le cas échéant, d'autres sources éventuelles peuvent également y être consignées.

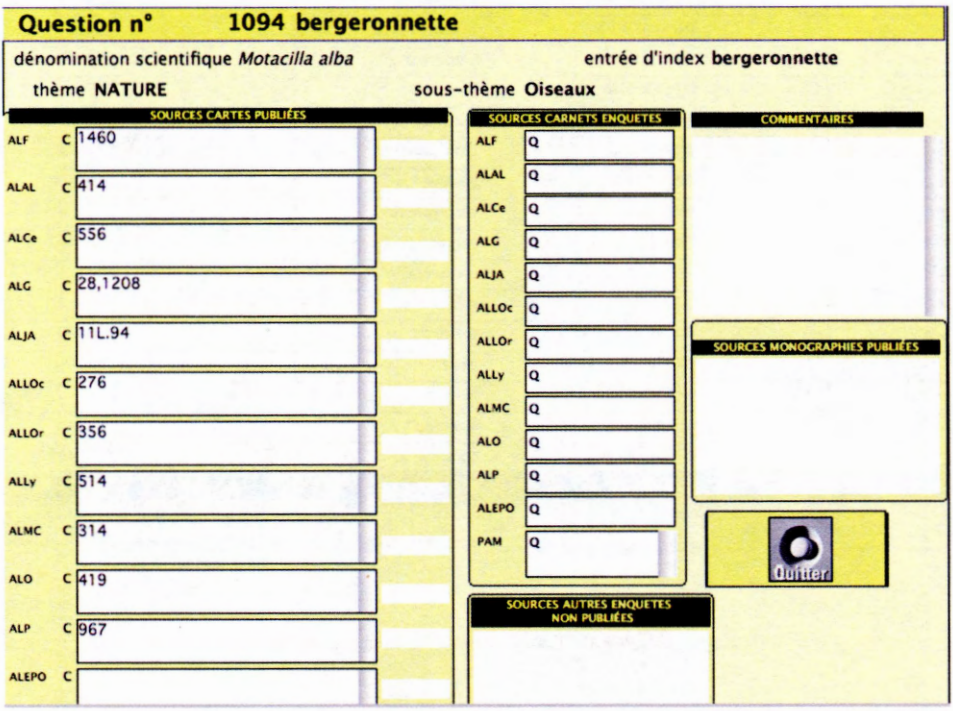


Fig. 5

Exemple de fiche-question: bergeronnette

Les entrées lexicales de la base sont organisées de la manière suivante: chaque entrée lexicale, ou fiche-réponse, est associée à une fiche-question et à une localité donnée.

1.2. Les différents modes de consultation

Il existe différentes possibilités d'interrogation de la base pour consulter les entrées lexicales: on peut soit rechercher toutes les fiches-réponses associées à une fiche-question précise, de manière onomasiologique, pour étudier la variation diatopique (voir fig. 6).

Chaque fiche question peut-être associée à une ou plusieurs illustrations (voir fig. 7), ce qui peut s'avérer très utile dans certains cas: par exemple, pour distinguer les différentes variétés, lorsqu'un terme quelconque peut correspondre à plusieurs objets.

FICHES RÉPONSES : 697 INTITULE : bergeronnette

localité	commentaire	forme phonique	lemme	graphie phonologisante	étymon
SAINT-PIERRE		bɛʀdʒejr'etɔ	bergereta°°		VĒRVĒCARIUS
BANON		pastʀ'esœ	pastressa°		PASTOR
MEZEL		buvejr'ete	bovaireta°°		BŌS
METHAMIS		baʀdʒejʀ'ɛtɔ	bergeroneta°°		VĒRVĒCARIUS
ANNOT		byjeʀ'etœ	bovaireta°°		BŌS
FORCALQUIER		ʀus'ɛtœ	rosseta		RŪSSUS
GORDES		pastʀ'ɛsɔ	pastressa°		PASTOR
MENTON		balaɾ'ina	balarina°°		BALLĀRE
BOULBON		bɛʀʒɛʀon'ɛtɔ	bergeroneta°°		VĒRVĒCARIUS
CAVAILLON		bɛʀdʒir'etɔ	bergereta°°		VĒRVĒCARIUS
MAILLANE		bɛʀdzɛʀ'etɔ	bergereta°°		VĒRVĒCARIUS
GREOUX		g'iɲɔkw'ɔ	guinha-coa°°		WINGJAN*
NICE		balaʀ'ina n'egʀa	balarina°°		BALLĀRE
EYGALIERE		pastʀ'ɛsɔ	pastressa°		PASTOR
CADENET		baʀdʒ'elą	bergièra°		VĒRVĒCARIUS
VILLELAURE		giɲ'yzɔ	guinhusa°°		WINGJAN*
EYGUIERES		g'iɲɔ kw'a	guinha-coa°°		WINGJAN*
SAINT PAUL LES DUR		g'iɲɔkw'ɔ	guinha-coa°°		WINGJAN*

TRIER PAR : Numéros de localités | Lemmes | Illustrations | Impression | Quitter

Fig. 6

Exemple de liste de réponses par question: bergeronnette

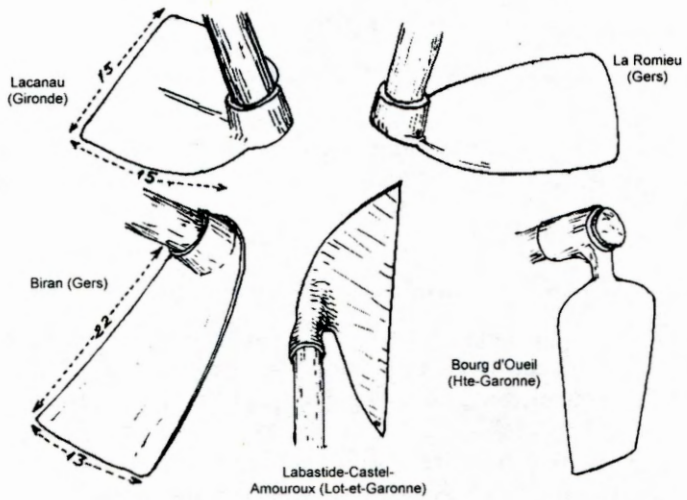
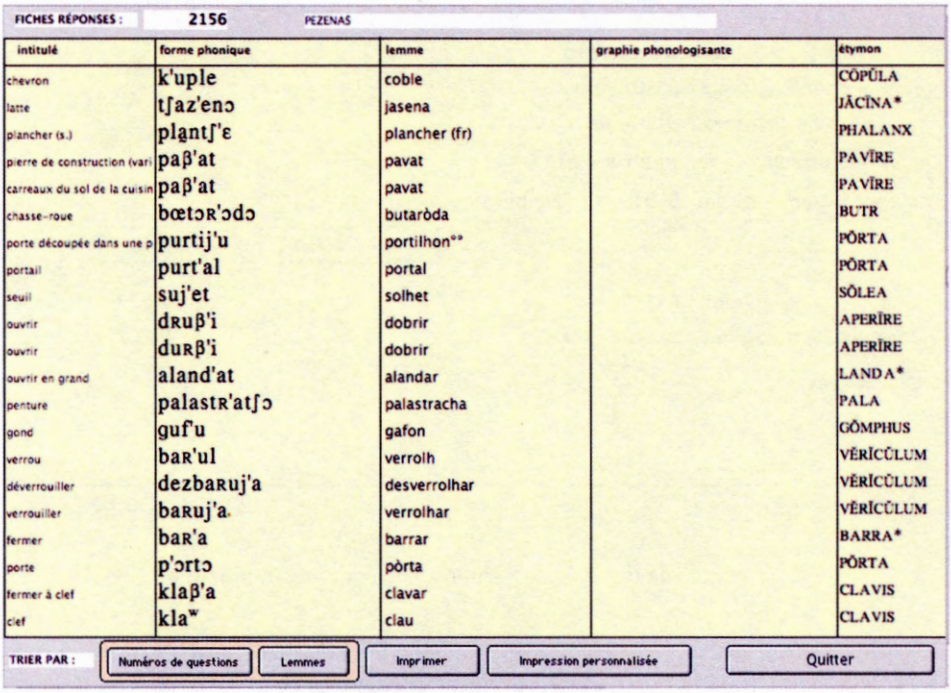


Fig. 7

Exemple d'illustrations: houes

On peut également consulter toutes les fiches-réponses associées à une localité donnée, pour établir une monographie (voir fig. 8).



FICHES RÉPONSES : 2156 PEZENAS

intitulé	forme phonique	lemme	graphie phonologisante	étymon
chevron	k'uple	coble		CŌPŬLA
latte	tʃaz'enɔ	jasena		JĂCĪNA*
plancher (s.)	plɑ̃ntʃ'ɛ	plancher (fr)		PHALANX
pierre de construction (var.	paβ'at	pavat		PAVĪRE
carreaux du sol de la cuisin	paβ'at	pavat		PAVĪRE
chasse-roue	bœtɔʀ'ɔdɔ	butaròda		BUTR
porte découpée dans une p	purtij'u	portilhon**		PŌRTA
portail	purt'al	portal		PŌRTA
seuil	suj'et	solhet		SŌLEA
ouvrir	dʀuβ'i	dobrir		APERĪRE
ouvrir	duʀβ'i	dobrir		APERĪRE
ouvrir en grand	aland'at	alandar		LANDA*
penture	palastʀ'atʃɔ	palastracha		PALA
gond	guf'u	gafon		GŌMPHUS
verrou	baʀ'ul	verrolh		VĒRĪCŬLUM
déverrouiller	dezbaʀuj'a	desverrolhar		VĒRĪCŬLUM
verrouiller	baʀuj'a	verrolhar		VĒRĪCŬLUM
fermer	baʀ'a	barrar		BARRA*
porte	p'ɔrtɔ	pòrta		PŌRTA
fermer à clef	klaβ'a	clavar		CLAVIS
clef	klaʷ	clau		CLAVIS

TRIER PAR : Numéros de questions | Lemmes | Imprimer | Impression personnalisée | Quitter

Fig. 8

Exemple de liste de réponse par localité: Pezenas (Hérault)

On peut enfin rechercher spécifiquement un certain couple localité-question. Par exemple, quel est le terme employé à Pezenas pour désigner "la bergeronnette" (voir fig. 9).

Le détail d'une fiche-réponse fait apparaître un certain nombre de champs:

— le numéro de question dans la base et son intitulé (1.094/bergeronnette), ainsi que son numéro de saisie (n.º 3.387);
— le numéro de localité dans la base et son nom (274/Pezenas) ainsi que son numéro dans l'Atlas Linguistique du Languedoc Oriental (ALLOr 34.32);
— la transcription phonétique en API (forme phonique);
— la transcription graphique: il s'agit d'une transcription «phonologisante», c'est-à-dire une sorte de forme intermédiaire entre la transcription phonétique et la graphie standardisante d'Alibert, cette forme intermédiaire adopte pour l'essentiel les principes du *Tresor dóu Felibrige* de F. Mistral. Elle est automatiquement générée par un algorithme, appelé *transcripteur*. Celui-ci est basé sur un ensemble de règles de réécriture qui peuvent être configurées par l'utilisateur et qui peuvent varier d'une localité à l'autre pour prendre en compte les systèmes phonologiques des différents dialectes occitans;

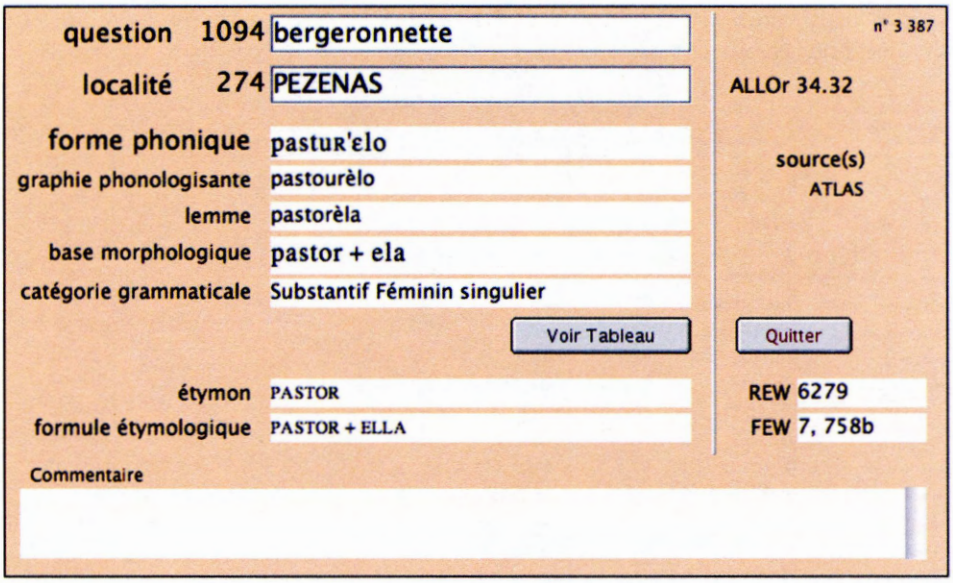


Fig. 9

Exemple de fiche localité-réponse: Pezenas/bergeronnette

— le lemme, qui est conçu comme forme de référence (conventionnelle); il sous-tend tout le faisceau de variantes consignées dans la base. Le choix du lemme est effectué en s'appuyant sur le *Dictionnaire occitan-français* d'Alibert; sa notation respecte donc les principes de la graphie alibertine;
— la base morphologique indique si le mot est un composé ou non ((préfixe) + base + (suffixe));
— la catégorie grammaticale;
— l'étymologie, avec des références aux principaux dictionnaires étymologiques *(Französisches Etymologisches Wörterbuch* (FEW) et *Romanisches Etymologisches Wörterbuch* (REW));
— enfin, en cas de besoin, des commentaires peuvent être adjoints.

Dans le cas des fiches contenant un verbe, en cliquant sur le bouton «Voir Tableau», on peut également consulter le paradigme de conjugaison verbale, lorsque celui-ci a été renseigné dans la base (voir fig. 10).

Lorsque le paradigme morphologique d'un adjectif ou d'un article possède des variations phonétiques contextuelles, un tableau supplémentaire permet de visualiser les différentes formes attestées, suivant le contexte phonétique qui suit et/ou qui précède ce terme, le tout étant illustré d'une série d'exemples (voir fig. 11).

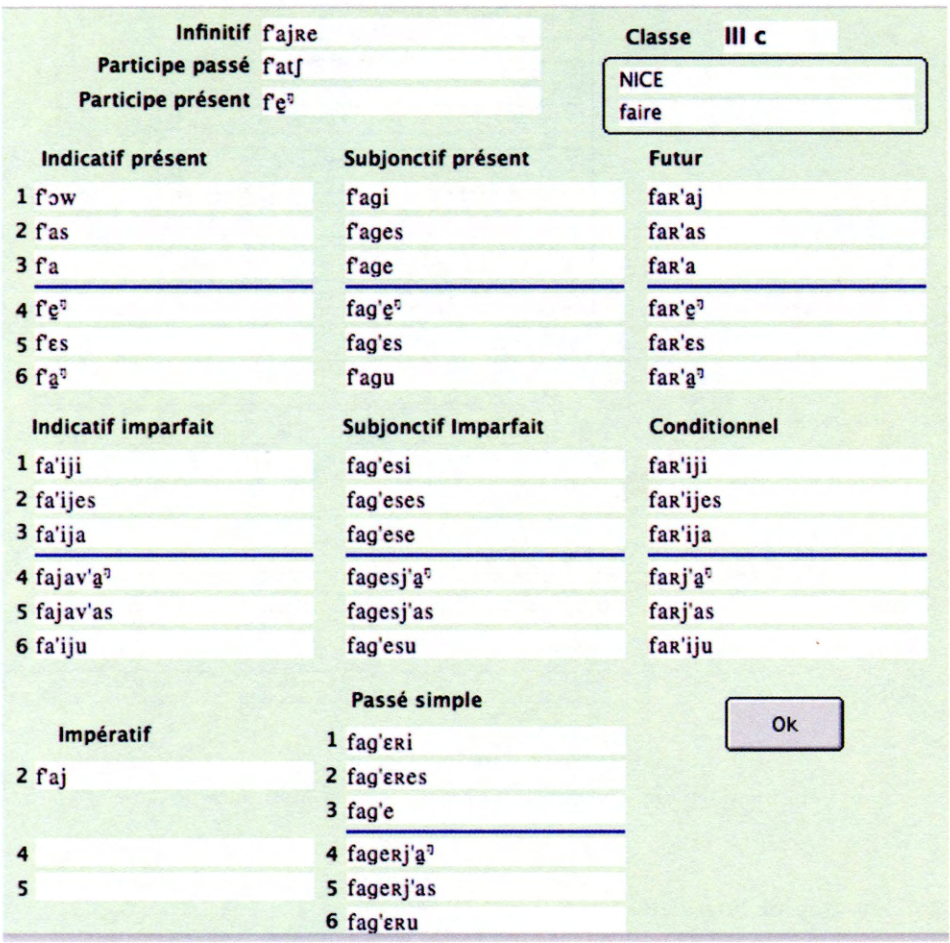


Fig. 10

Exemple de paradigme verbal: Nice (Alpes-Maritimes)/faire

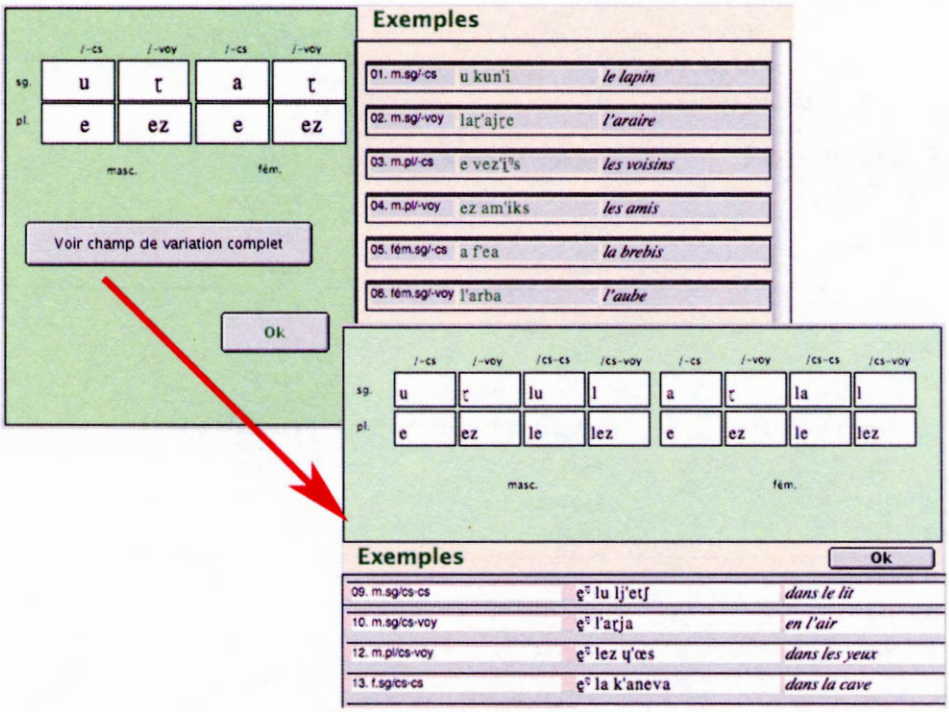


Fig. 11

Exemple de variations contextuelles: Castillon (Alpes-Maritimes)/le (article)

2. Cartographie interactive

La consultation de la base ne permet pas seulement d'obtenir des glossaires, elle permet également de cartographier des faits lexicaux, de différentes manières.

Deux types de cartes sont disponibles dans le THESOC: d'une part, des cartes présentant les faits bruts, et d'autre part, des cartes de synthèse. Aucune carte n'est cependant stockée dans la base de données: elles sont toutes générées dynamiquement à partir des données linguistiques présentes dans la base, en fonction des requêtes demandées par l'utilisateur.

Chaque fois que l'utilisateur modifie sa requête, une nouvelle carte est générée en temps réel. C'est en ce sens que l'on peut dire qu'il s'agit d'une cartographie interactive.

2.1. Les cartes de faits bruts

Comme il ne serait pas lisible d'afficher sur une carte de l'Occitanie toute entière l'ensemble des transcriptions phonétiques attestées dans les différentes localités, les cartes présentant les faits bruts sont disponibles à deux échelles, avec un système de

zoom: en premier lieu, au niveau de l'Occitanie toute entière, un simple point rouge signale les localités pour lesquelles la base contient une réponse à la question qui est cartographiée (voir fig. 12).

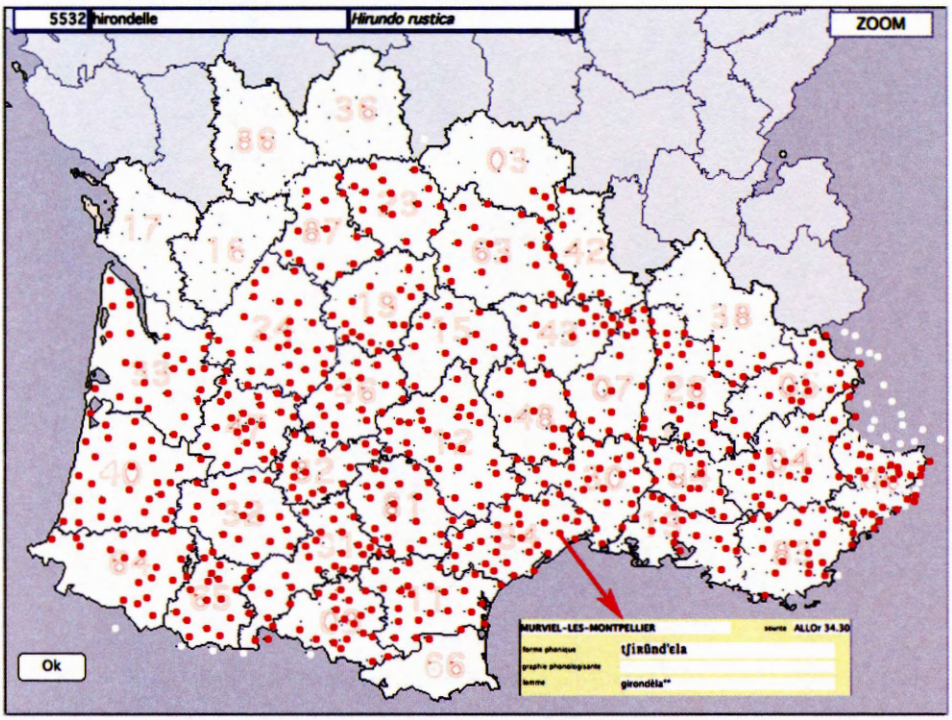


Fig. 12

Carte à symboles: hirondelle

Un clic sur l'un des points permet de visualiser sa ou ses transcriptions phonétiques dans une petite fenêtre.

Puis dans un deuxième temps, il est possible de zoomer à l'échelle d'un département (voir fig. 13).

La carte détaillée affiche la transcription phonétique associée à chaque réponse, à côté du point de la localité concernée, comme dans les atlas linguistiques. Un point rouge indique que l'on peut écouter l'enregistrement sonore associé.

Un module spécifique du Tableau de Bord intitulé «Sons» permet d'écouter les enregistrements regroupés par localités.

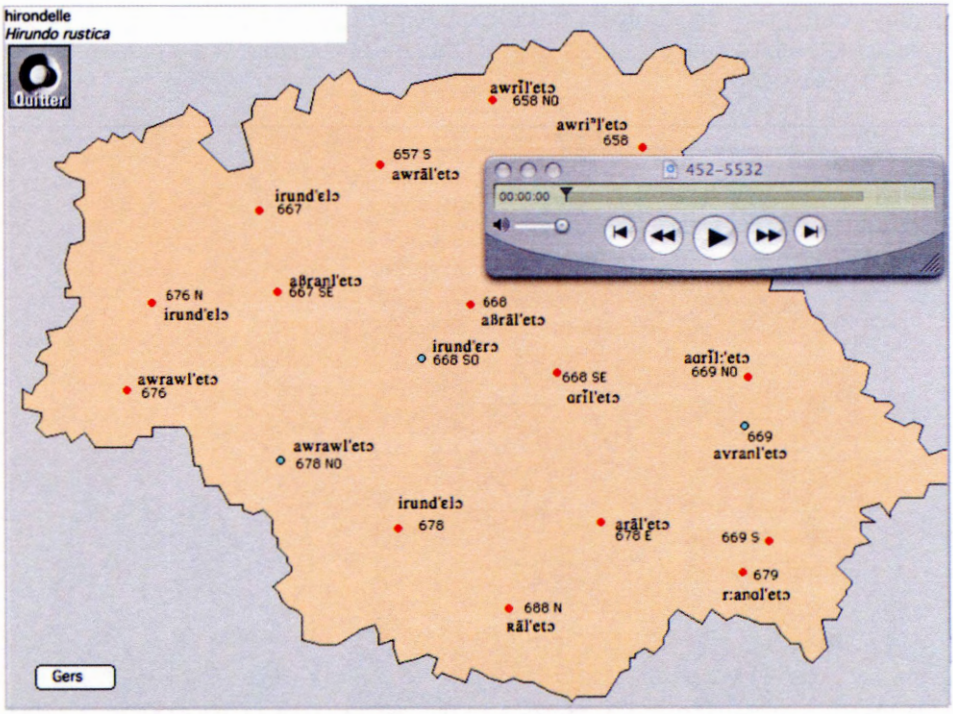


Fig. 13

Carte départementale des réponses en phonétique: hirondelle/Gers

2.2. Les cartes de synthèse

Voyons à présent à partir d'un exemple, le processus d'élaboration d'une carte de synthèse, concernant la répartition géographique des différents types lexicaux désignant le verbe "scier".

Le logiciel affiche à l'écran la liste des lemmes répertoriés dans les différentes fiches réponses de la base concernant ce terme (voir fig. 14). On peut alors effectuer des groupes contenant un ou plusieurs de ces lemmes, selon nos souhaits, et affecter une couleur à chacun de ces groupes. Un simple appui sur le bouton «symboliser» déclenche la construction et l'affichage quasi-instantanés de la carte ainsi demandée (voir fig. 15).

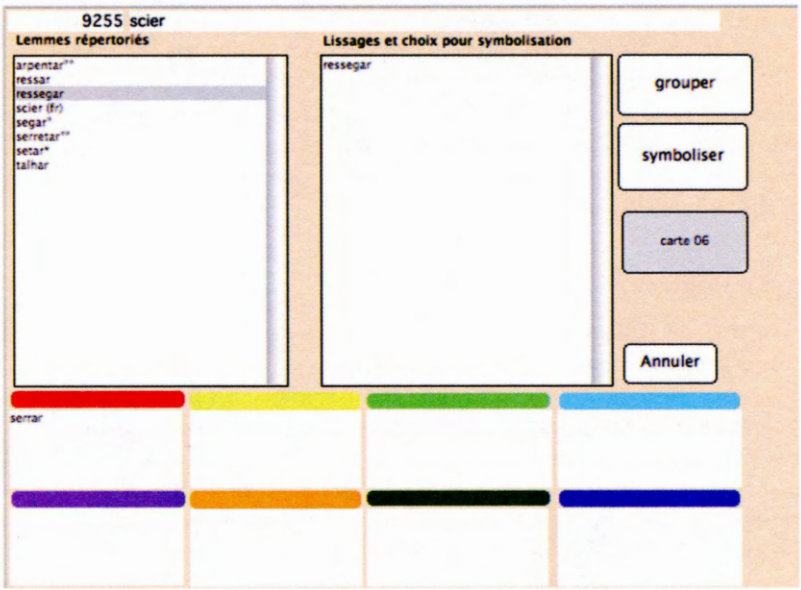


Fig. 14

Tableau de répartition des lemmes pour "scier"

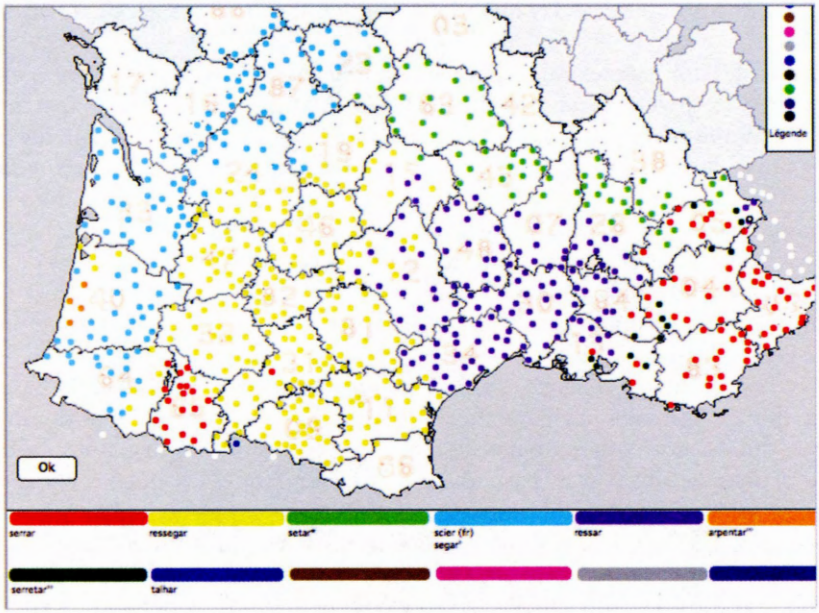


Fig. 15

Carte de synthèse: carte à symboles ("scier")

Sur cette carte aréale à symboles, on voit très bien se dessiner les différentes zones lexicales très distinctes formées par les termes désignant le verbe «scier» dans l'ensemble des parlers occitans.

L'utilisateur peut ainsi modifier les critères et les regroupements à volonté pour générer autant de cartes qu'il le souhaite. Signalons enfin que ces cartes peuvent être exportées au format vectoriel pour être intégrées dans un document PDF ou pour servir de document de travail.

3. Autres ressources

Le THESOC ne se limite pas aux données purement lexicales, mais il permet également d'autres recherches, à divers niveaux.

Nous avons déjà évoqué le module «Sons», mais il faut signaler en outre le module «Morphologie verbale et nominale» qui donne la possibilité de voir les différents paradigmes verbaux et nominaux, comme nous l'avons vu plus haut, le module «Cartes des Atlas» qui permet de visualiser les domaines des différents atlas linguistiques dont les données figurent dans la base, ainsi que le module «Sources» qui donne les références des ouvrages concernés par les travaux du THESOC.

Nous avons également un module spécifiquement dédié à la syntaxe avec de nombreuses fonctions qui seront détaillées par Pierre-Aurélien Georges ici-même.

3.1. Le module Oc-Français ou Dictionnaire inversé

Nous nous attarderons davantage sur le module Oc-Français ou Dictionnaire inversé.

À partir des données lexicales présentes dans la base, le THESOC propose également un dictionnaire inversé occitan / français. Ce module permet une démarche tour à tour onomasiologique et sémasiologique, sur la base des lemmes (voir fig. 16). Il permet ainsi de trouver les différentes notions correspondant à un terme occitan donné. Ce type de requête est particulièrement utile pour des recherches en sémantique lexicale et reconstruction étymologique (cf. Dalbera 2006).

Par exemple, si on choisit comme point de départ, *barbo-*, dans le cadre de droite, s'affichent tous les lemmes commençant par cette séquence. On entre alors dans le tableau principal du module en choisissant dans la liste un lemme occitan donné, par exemple *barbòta* (voir fig. 17).

Dans cette nouvelle fenêtre, s'affichent tous les sens relevés dans la base pour ce mot: lorsqu'ils n'ont pas de signe, c'est que le sens existe dans le dictionnaire d'Alibert qui est notre dictionnaire de référence, et lorsqu'ils sont accompagné d'un signe, c'est que le sens n'avait pas été relevé. On remarque, ici, que les enquêtes des Atlas linguistiques régionaux donnent des d'indications supplémentaires.

En cliquant sur l'une des lignes, toutes les localités qui utilisent ce terme pour désigner le référent correspondant apparaissent dans la fenêtre inférieure (le mot *barbòta* est utilisé à Peyrat-de-Bellac (Haute-Vienne) pour désigner «la couleuvre»).

Tandis qu'en sélectionnant un signifié (le mot français, par exemple «couleuvre»), on accède à la liste de tous les autres termes dialectaux (les signifiants) renvoyant à la même notion (*bòba, cinglant^{oo}, cingla^{o},* etc.).

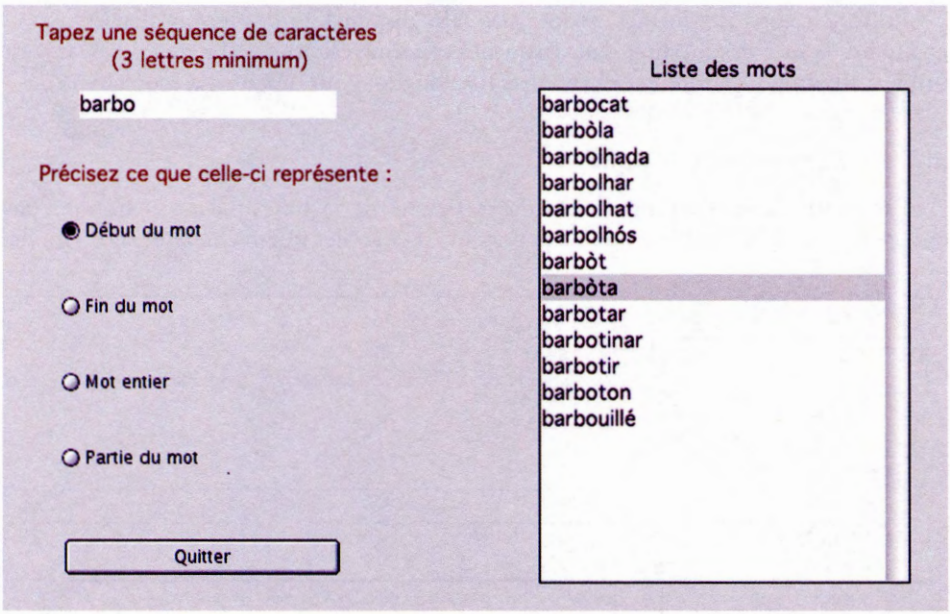


Fig. 16

Tableau de choix du dictionnaire inversé (*barbo-*)

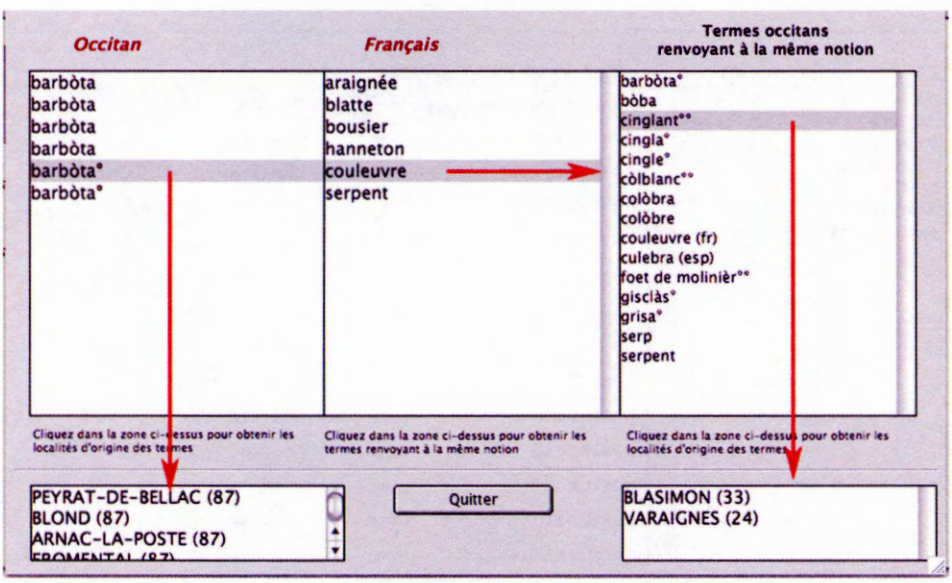


Fig. 17

Tableau de consultation du dictionnaire inversé (*barbòta*)

Enfin, chacune des formes est localisée. Un clic sur l'un des termes dialectaux fait apparaître la liste des localités dans lesquelles ce terme est attesté (Le mot *cinglant*^{oo} est utilisé à Blasimon (Gironde) et Varaignes (Dordogne) pour désigner la «couleuvre»).

3.2. Recherches étymologiques

Parmi les autres fonctionnalités de recherche de la base que nous n'avons pas encore abordées, il faut noter qu'il est possible d'effectuer une recherche par étymons:

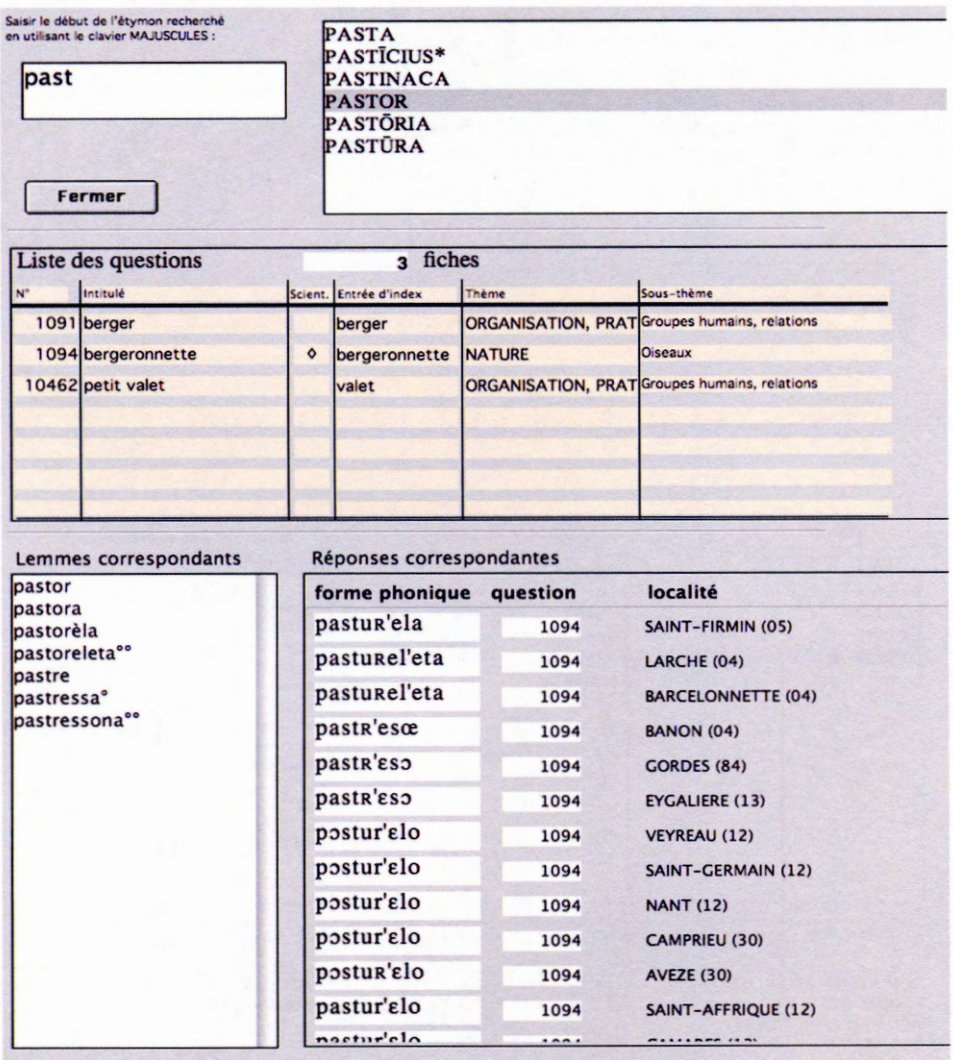


Fig. 18

Tableau de consultation des données étymologiques (PASTOR)

Si l'on saisit par exemple l'étymon latin *PASTOR*, apparaît alors à l'écran la liste des questions en rapport avec cet étymon (voir fig. 18), c'est-à-dire toutes les questions pour lesquelles la base contient au moins une fiche réponse qui possède l'étymon demandé (dans l'exemple les questions «berger» (shepherd), «bergeronnette» (wagtail) et «petit valet» (little farm hand), puis au-dessous la liste des lemmes associés (*pastor, pastora, pastorèla,* etc.), ainsi que dans la fenêtre de droite, la liste des réponses correspondantes pour chaque lemme dans les localités concernées.

3.3. Le volet de microtoponymie

Enfin le THESOC comporte aussi un volet de toponymie qui consigne les différents micro-toponymes recueillis lors des enquêtes.

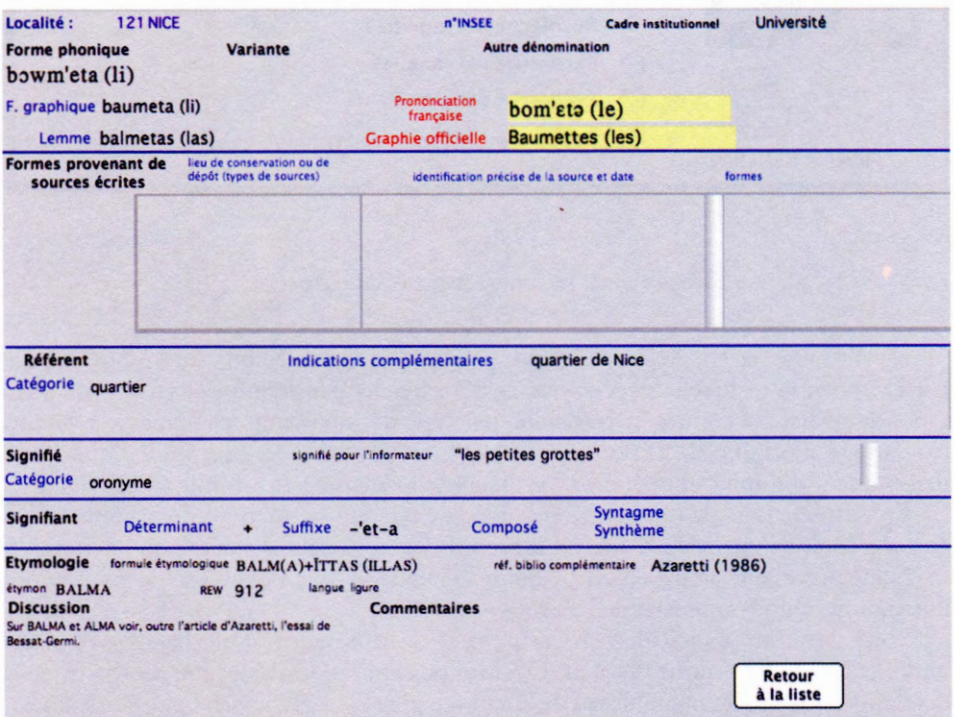


Fig. 19

Exemple de fiche toponymique (*Las Balmetas*/Nice (Alpes-Maritimes))

Une fiche toponyme (voir fig. 19) contient généralement les informations suivantes: la localité dans laquelle le micro-toponyme a été recueilli, les formes graphique et phonétique du toponyme, le signifié associé à ce terme dialectal, les formes graphique et phonétique de sa traduction en français, le type de toponyme (oronyme, hydronyme, etc.), le type de référent (quartier de la commune, chemin,

cours d'eau, forêt, etc.), l'étymon et les références étymologiques du type *REW,* des commentaires, et d'éventuelles références bibliographiques complémentaires.

Lorsque ce toponyme fait par ailleurs l'objet d'attestations écrites, il est également possible de consigner les formes provenant des différentes sources écrites, avec une identification précise de ces sources et de leur date.

À l'instar de la partie lexicale du THESOC, diverses fonctionnalités de recherche et d'analyse proposées par la base permettent ensuite de consulter ces données (voir fig. 20):

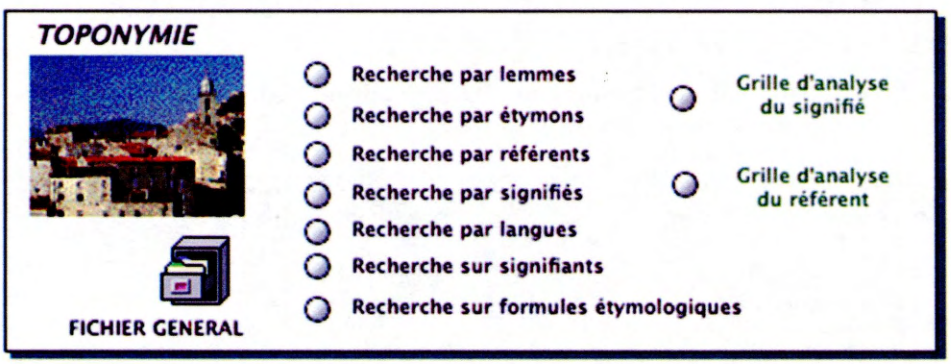


Fig. 20

Toponymie: les différents types de recherches

On retrouve là encore la recherche par localité, pour établir une monographie par exemple, la recherche par lemme, et la recherche par étymon; ainsi que d'autres types de recherche comme la recherche par type de toponyme, par type de référent, par langue d'origine de l'étymon, par type de construction morphologique (avec déterminant ou non, mot composé, syntagme) ou encore par formule étymologique.

En outre, le type de toponyme et le type de référent permettent de proposer deux grilles d'analyse: une grille d'analyse du signifié, et une grille d'analyse du référent.

Nous prévoyons actuellement d'ajouter dans la base la possibilité de joindre des illustrations à une ou plusieurs fiches toponyme.

Voici donc un aperçu de l'ensemble des fonctionnalités de la base de données multimédias du Thesaurus Occitan. D'autres possibilités devraient être prochainement développées, notamment au niveau de la cartographie et de la mise en ligne des données.

References

Alibert, L., 1966, *Dictionnaire occitan-français*, Institut d'Etudes Occitanes, Toulouse.
Dalbera, J.-P., 2006, *Des dialectes au langage: Une archéologie du sens*, Champion, Paris.
Meyer-Lübke, W., 1935, *Romanisches Etymologisches Wörterbuch*, Winter, Heidelberg.
Mistral, F., 1878, *Lou Tresor dou Felibrige*, Raphèle-lès-Arles, Paris.
Oliviéri, M., 2004, «Le responsaire du THESOC», *Actes du colloque 8ème Colloque de dialectologie et littérature du domaine d'oïl occidental, Avignon, 12-13 juin 2002*, P. Brasseur (éd.), pp. 23-34.
Wartburg (von), W., 1922-..., *Französisches Etymologisches Wörterbuch*, Schroeder, Bonn.

THE THESAURUS OCCITAN: A MULTIMEDIA DATABASE DEDICATED TO OCCITAN DIALECTS. PRESENTATION OF ITS MORPHOSYNTAX MODULE

Pierre-Aurélien Georges

Laboratoire BCL, CNRS UMR 6039
Université Nice Sophia-Antipolis, MSH de Nice (France)

Abstract

The Module MorphoSyntaxique *(abbreviated MMS) is a computer tool especially designed for syntactic and morpho-syntactic analysis of Occitan dialects. It is part of the* Thesaurus Occitan *multimedia database (of which a general presentation can be found in these proceedings in another article by Guylaine Brun-Trigaud).*

Following the THESOC's general guidelines (i.e. localised and oral data only), this module contains both oral texts (including ethnotexts*) and single sentences, such as answers to morphosyntactic questionnaires.*

The "oral data" criteria can be somewhat flexed: even if this module was originally conceived for oral data processing, its part-of-speech tagger and syntactic parser are still able to process written texts so far as they are written in a familiar or popular style, close to oral register. The locations where all these texts and sentences have been harvested are stored in the database, thus enabling on the long term a comparison between different dialects on a morphosyntactical or syntactical basis, thus opening new perspectives for dialectology.

Keywords: *Comparative syntax; morphosyntax; Occitan dialects;* ethnotexts *corpus; database*

1. General Presentation

The *THESAURUS OCCITAN* (abbreviated THESOC) is a multimedia database which encompasses oral dialectal data from the whole Occitan domain. Aside from its lexical and

microtoponymy corpuses, the THESOC also contains a module dedicated to morphosyntax and syntax analysis of Occitan dialects.[1] This *Module MorphoSyntaxique* (MMS)[2] database is aimed at studying oral syntax and morphosyntax in a comparative way.

1.1. Conditions

As for the rest of the THESOC, raw data to be included in this MMS module must match the two following criterions:

— *Location condition:* linguistic data must be precisely located. This constitutes an essential condition for diatopic variation studies. In particular, this condition will enable dynamic and automated generation of linguistic maps on demand in the future.

— *Orality condition:* linguistic data should come from oral sources. Indeed, our philosophy here is to integrate linguistic facts collected under oral form (with IPA transcription), which guarantees the reality of these facts under consideration. Moreover, the THESOC allows hearing of the audio tracks recorded during the field works, thus giving to the user the possibility to control or to check the proposed transcription.

1.2. Different types of data

On the one hand, the database contains a collection of single sentences or isolated sentences: answers to morphosyntactic questionnaires such as PAM,[3] sentences found in unpublished survey notebooks of the *Atlas linguistiques,*[4] or sentences published in some of these atlases, such as the ALMC.[5]

On the other hand, the database also contains a collection of texts, such as *ethnotexts* collected on the field, or radio broadcastings.

Similarly to the other parts of the THESOC, multimedia documents such as pictures, sound files, or video files, can be attached to a text or a single sentence.

1.3. Text types and orality condition

Although this module was originally conceived for treating oral data only, its lemmatiser and syntactic parser presented in sections 2.2. and 2.3. are even capable of treating written texts as long as they are written in a familiar or popular style, close to oral register; thus dimming somewhat the orality condition required in section 1.1. This orality condition may then be eventually reformulated as the following: linguistic data must contain oral syntax or popular/close-to-oral syntax.

---

[1] For a general presentation of all the other aspects of the THESOC, please refer to the article by Guylaine Brun-Trigaud also included in these proceedings, and/or (Georges, not yet published, a).

[2] It was formerly known as «base TEXTES» in (Georges 2009) because it originally started has a corpus of texts and *ethnotexts*, as explained in (Oliviéri 2003).

[3] Parler des Alpes Maritimes, supervised by (Dalbera 1994).

[4] *Atlas Linguistiques de la France par régions*, éditions du C.N.R.S, as presented in (Séguy 1973).

[5] Atlas Linguistique du Massif Central (Nauton 1957-1963).

This possibility has allowed us to extend the corpus with other types of texts: some theater plays, articles from popular press, etc. However, each text record from the database contains a field that informs the user about its type/gender. As shown in Figure 1, *Ethnotexts* are thus easily identifiable by the "genre: Ethnotexte" field content, whereas other types of texts have another tag in this field, such as "Chanson" (song lyrics) or "Presse" (press article) for example. Thanks to this field, it's always possible to filter out written texts and to focus only on *ethnotexts* and/or some other types of oral texts: search queries can be configured to show results from the whole corpus or from only certain types of texts specified by the user. This way, linguists can decide whether to stay on strictly oral data, or to also include some type of written texts in order to get a broader corpus.

Fig. 1

Example of an *ethnotext* record

2. Data processing

2.1. Adding new data to the corpus

Data can be added to the database through its XML import/export functionalities[6] or by typing new records directly within the user interface.

The graphical transcription field of a new record can have two different origins, depending on the nature of this record:

— If it's a text (written in a familiar or popular style, close to oral register), for which a graphical transcription is, by definition, already available, it is directly

[6] Thus, TEI import / export can also be achieved by using an XSLT filter.

stored in the database, whatever writing conventions or writing system have been used by its author, since there is no unique official spelling norm (or "*orthographe*") for the Occitan dialects as is the case for French or English, but rather several writing systems that competes.[7] The database also contains an algorithm which tries to automatically detect which writing system has been used within a text,[8] for users' information.

— If it's based on a sound track recorded on the field, for which the user doesn't already posses a graphical transcription (as is typically the case with *ethnotexts*), it's possible to easily generate a phonological graphical transcription: the user simply presses on a button (button #2 shown in Figure 2 below) to call an automated transcriber which is available in MMS (as in the rest of the THESOC), that generates a graphical transcription (#3) directly from the IPA transcription (#1), which must therefore be typed or imported in the database first. This automated transcription is based on conversion rules that can be modified by the user and adapted to different phonological systems among dialects, so that the transcriber will automatically use different rule sets to transcribe different dialectal areas, depending on the information given by the locality field of the record to be processed. The result is a phonological graphical transcription close to (Mistral 1979)'s writing system known as "*graphie mis-*

Fig. 2

Processing a text record

[7] The three most common writing systems used are "*graphie alibertine*", as defined by (Alibert 1966), "*graphie mistralienne*" as in (Mistral 1979), and in a far smaller proportion, the Eastern part of Occitania sometime uses "*graphie italianisante*", which is inspired from Italian's writing system.

[8] This is based on statistical occurrences of some sequences of letters, and these rules are user-customizable, therefore, any new writing system can be added to this automated recognition feature. Should this algorithm ever fail, on a very short text for example, where the situation is unclear, it is always possible to override it and to manually specify the correct writing system used by the text.

tralienne", which provides users a more readable way to access the content of the text, and allows a first level of abstraction that "smoothes" or "hides" phonetic variation.[9]

In both cases, all further linguistic treatments proposed by the different tools available in MMS are based on this graphical transcription. Thus, while these tools were originally conceived for oral data processing, they are even able to process written texts so far as they are written in a familiar or popular style, close to oral register. This is how it is even technically possible to introduce "oral-style" written texts in the database.

Among these tools, the part-of-speech tagger and the syntactical parser automate a great amount of the annotation work, thus simplifying processing of new data, as shown in following sections.

2.2. Lemmatisation process

The next step after importing or typing new data in the database is the lemmatisation process (button #4). The part-of-speech tagger identifies each individual lexical element of a sentence or a text by using a reference dictionary embedded in the database.

In order to manage efficiently the variation in all its aspects (graphical, dialectal, or inflectional variation), this dictionary is structured into two hierarchical levels: the *variantes,* which are in fact the lexical occurrences found in the corpus, are grouped under *lemmas.* This allows performing searches or linguistic treatments either on a particular form (e.g. such inflection, with such graphical transcription, in such dialect), or on all forms associated to a given lemma. Figure 3 below illustrates this two-levels dictionary structure:

Lemmas		
lo	determiner	«the»
li	pers. pronoun	«him»
fòl	adjective	«crazy»
bèl	adjective	«beautiful»

Variantes of a lemma		
fuala	fem. sing.	Nice
fuale	masc. plur.	Nice
fuali	fem. plur.	Nice
fòlha	fem. sing.	Toulouse

Variantes of a lemma		
bèl	masc. sing.	Nice
beu	masc. sing.	Nice
bela	fem. sing.	Nice
bella	fem. sing.	Nice

Fig. 3

Database's dictionary structure

[9] Since the objective of MMS is to study syntax and morphosyntax, it's not a major issue here to disregard phonetic variations in order to simplify the following linguistic treatments performed.

This schematic illustrates the different types of variation handled:

— phonological variation, between *bèl* (preceding a vowel) et *beu* (preceding a consonant) in the dialect of Nice, called *"Nissart"*
— dialectal variation, between *fòlha* in Toulouse's dialect and *fuala* in Nissart,
— graphical variation, in Nissart, between *bela* and *bella*,
— and inflectional variation, also in Nissart, between *fuala*, *fuali* and *fuale*.

As reference forms, lemmas' graphical forms are based on entries from (Alibert 1966) when available[10] whereas the *variantes* part of the dictionary is currently populated by entries coming both from the lexical database of the THESOC and from a set of several paper dictionaries, such as (Eynaudi 1932), that have been digitalized and integrated in the database for this purpose. Moreover, the process of manual lemmatisation of unidentified lexical items sometimes adds new entries to the dictionary, as detailed below in this section, as well as the syntactical-tree annotation of sentences. Thus, dictionary's content is constantly improved by the lemmatisation process of new texts and/or sentences and new entries coming from THESOC's lexical database.

As illustrated in Figure 4 below, the output of the lemmatisation process (#5) shows the following information about each lexical item identified in the text or sentence processed: its lemma, morphosyntactic category, and inflection. Some items may remain unidentified after the lemmatisation process, either because there is no matching entry in database's dictionary or because there are several potential candidates.

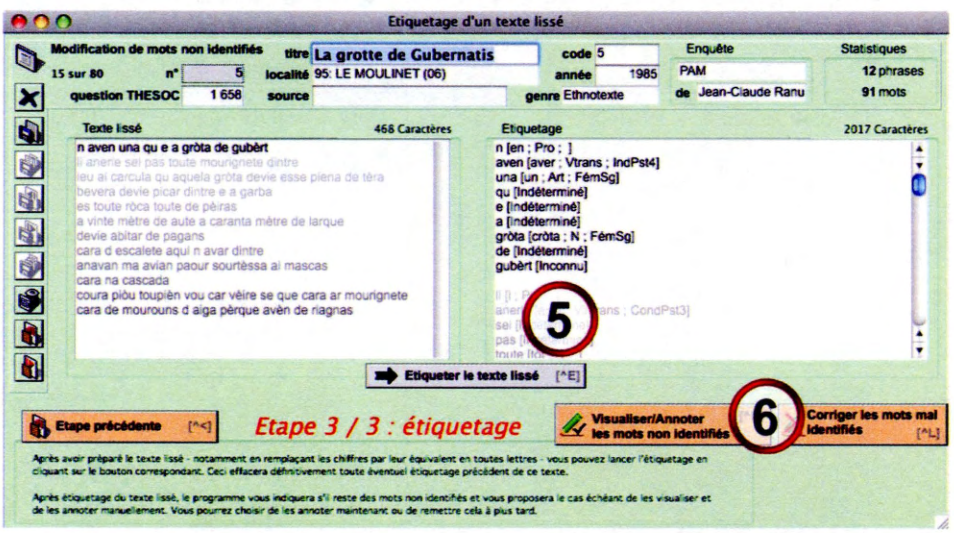


Fig. 4

Lemmatisation process

[10] When there is no corresponding entries in (Alibert 1966), an asterisked form is proposed in Alibert's writing system, known as "*graphie alibertine*".

As can be seen in Figure 4, unknown items are tagged "*Inconnu*" (as is the case of proper noun "*Gubèrt*") whereas ambiguous items are tagged "*Indéterminé*": the graphical form *"e"* can match either with an inflected form of the verb *èstre* "to be" (3rd Pers. Sing.) or with the conjunction coordination "and"; similarly, *"a"* can match either the definite determiner (Fem. Sing.) or an inflected form or the verb *aver* "to have" (3rd Pers. Sing.).

When a lexical item was not correctly identified or not identified at all by the lemmatisation process, the two buttons labelled #6 in Figure 4 allows to manually identify this lexical item. In the case of ambiguous item, with several potential matching candidates in database's dictionary, the user can choose the right entry within a list of these potential candidates as shown by #7 in Figure 5. If the unidentified lexical item has no corresponding entry in the dictionary, it's also possible to add a new entry in the dictionary (#8) and to identify the lexical item with this new entry (these two points are realized at a glance, as one single operation). This unique user interface also shows the unidentified lexical item in context (bottom of Figure 5) to ease its manual identification by the user.

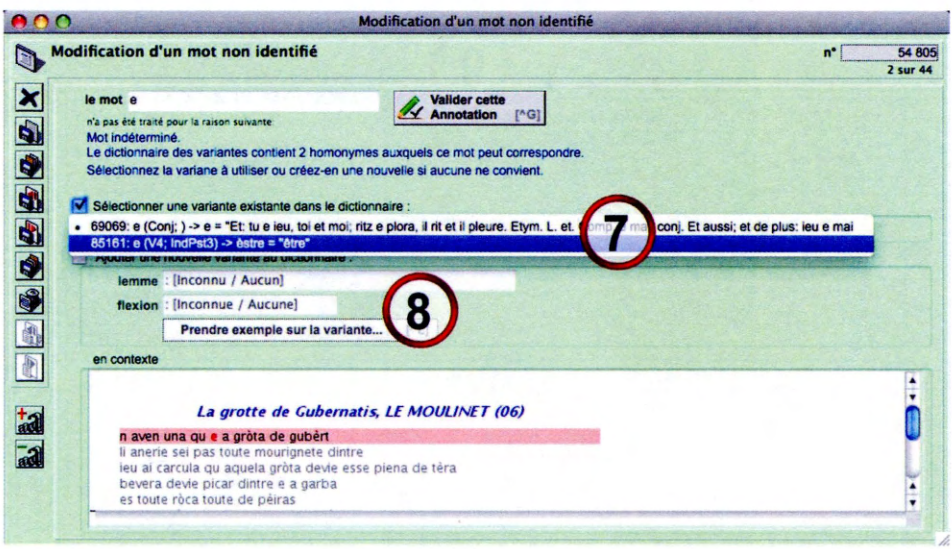


Fig. 5

Annotation of ambiguous items

After lemmatisation process has been performed, the next step is syntactical-tree tagging, eased by MMS' syntactical parser. It is not necessary here to lemmatise each single lexical item of a sentence or a text before using MMS' syntactical parser: if 80% or 90% of the lexical items have been correctly lemmatised, this is enough to switch to next step, as will be explained below.

2.3. Syntactical tree tagging

MMS' syntactical parser uses both data generated by the lemmatisation process and the embedded dictionary of the database in order to propose one or several possible syntactic structures for each sentence,[11] thus automating in some proportion the syntactical-tree tagging process. The syntactic trees are generated in a generativist theoretical framework, but the syntactic rules on which the parser relies are user-customizable.[12]

Figure 6 below gives an example of this process. On the left pane, the results of the lemmatisation process are shown. By clicking on the button #9, the user calls the syntactical parser, which analyses the sentence and output some possible syntactic structures in a list shown just under this button. The user must then browse among these candidates and chose the correct hierarchical syntactic structure(s) that really correspond to this sentence: by clicking on a candidate structure within the list, the hierarchical representation of the selected structure is shown in the right pane.[13]

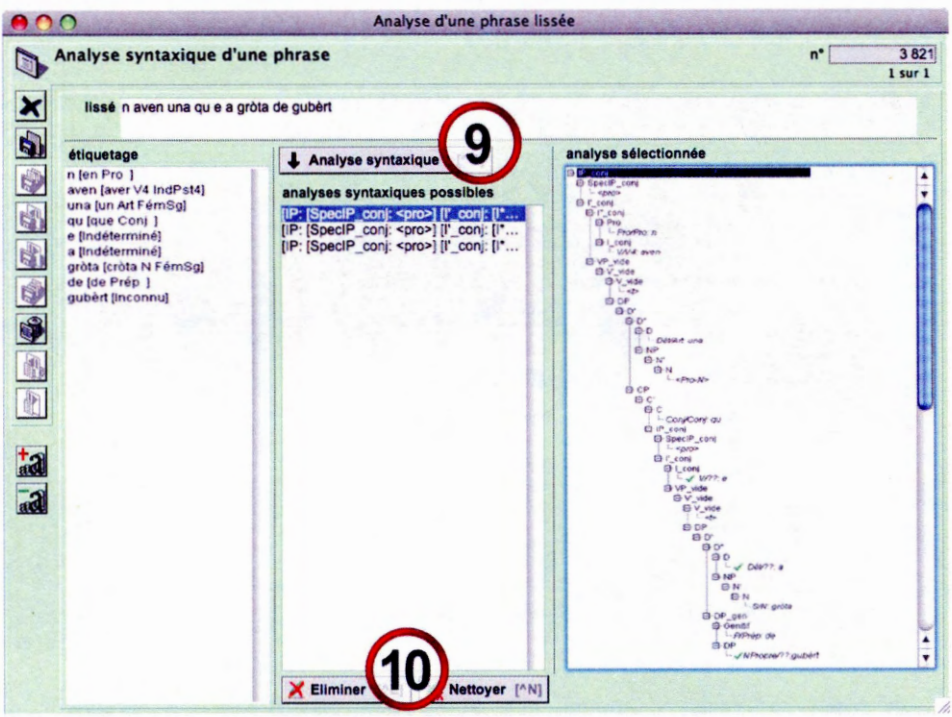


Fig. 6

Syntactical structure annotation

[11] Wether it's an isolated, single sentence, or a sentence from a text.

[12] The only main constraint is that generated syntactical trees must have binary branching nodes.

[13] Please note here that the syntactical tree structure (in the right pane) is displayed vertically instead of the traditional horizontal presentation. This is due to some technical limitations we are currently working on.

Since the candidates list is sorted by probability (most probable candidates are located on top of the list, whereas least probable ones are at the bottom), users therefore typically only need to look at the two or three first propositions to find the correct one(s).

Thanks to the two buttons under #10, false candidate structures are then eliminated in order to keep only the correct one, or the two or three right ones (when there's an ambiguity in the sentence that can not be resolved, for example if the author of this sentence has deliberately made a play on words). If the appropriate structure(s) is/are not present in the list generated by the syntactical parser, the user can choose any proposed candidate and edit this structure in order to manually build the correct hierarchical tree(s).

The use of MMS' syntactical parser thus simplifies and speeds-up the syntactical-tree tagging task by providing the user a semi-automated way of generating hierarchical structures. Moreover, it also simplifies and speeds-up the lemmatisation process: Figure 6 shows for example that the unknown proper noun "*Gubèrt*", that had been neither correctly identified by the automated lemmatiser nor manually tagged by the user, has nevertheless been correctly identified as a proper noun by the syntactical parser. Similarly, the ambiguous elements "*e*" and "*a*" have been successfully identified as respectively an inflected verb and a determiner. Therefore, the syntactical parser can automatically "guess" some lexical items that remained unidentified after the lemmatisation process, and disambiguate some ambiguous lexical items, thus also simplifying the lemmatisation task, since it's no longer mandatory to tag 100% of all the lexical items of a sentence.

After these 10 steps have been performed, annotation process is now complete: each lexical item of each sentence has been lemmatized, and each sentence is tagged with one or several hierarchical syntactic structure(s). Data are then ready to be exploited through different work features offered by the database.

3. Work features

3.1. Search engine

Once the sentences and texts from the database have been annotated with the help of these tools mentioned in section 2., MMS provides several search options to select data with different criterions. For example, it's possible to search all occurrences:

— of a given *variante;*
— of all *variantes* associated to a given lemma;
— of all *variantes* with one or several given morphosyntactic categories;
— of a given sequence of part-of-speech categories;[14]
— of a given location;[15]
— having a graphical similarity with a given lemma or *variante.*

It's also possible to perform searches based on syntactical tree tagging previously processed: one can search for all sentences containing a particular syntactic struc-

[14] Some wildcard options are available, such as "Beginning of sentence only" / "End of sentence only".
[15] This is, all occurrences of all texts associated to a particular given location.

ture or syntactical tree fragment, such as all sentences containing a DP within another DP for example. When an ambiguous sentence has several syntactic structures associated to it, such a search query will show this sentence in the search results if it matches any of its associated structures.

As illustrated in Figure 7, the results are shown in context: the list of matching occurrences is presented in column, and each matching occurrence is displayed within the full original sentence where it is coming from.

3.2. Automated work corpus generation

Whichever search query is formulated, a work corpus can then be automatically generated from these search results.[16] As presented in Figure 7, it is possible, for example, to search for all occurrences with morphosyntactic category «personal pronoun», then to select some particularly interesting ones in the results list displayed,[17] and to automatically generate a work corpus in which these selected occurrences are highlighted in their original context: text title,[18] location, full sentence or full text where the term occurs. In this example, we chose to generate a "concise" work corpus, i.e. only matching sentences are being extracted and displayed gathered, but it is also possible to generate a "full" work corpus, which gathers both matching isolated sentences and the full texts that contains at least one sentence containing one occurrence matching the search query.

In the same way, one can generate a work corpus from search results based on syntactical tree tagging previously processed. These work corpuses can be exported and saved as RTF or Microsoft Word format for further exploitation.

3.3. User-customizable taggings

In 2009, a new feature has been added to MMS: users now have the ability to attribute some user-defined "tags" to some sentences of the database.[19] Then, it's possible to perform cross searches based on these tags within the database. This allows, among other things, to search for presence or absence of a correlation between two or several linguistic parameters, confirming or invalidating user's hypotheses.[20]

It's possible, for example, to search within the database for texts that contain at least one sentence with a given tag but do not contain any sentences with another given tag (evaluating a hypothetical positive correlation between two parameters within the same speaker, therefore the same idiolect). Another use would be to search within the database for locations for which there is at least one sentence with

---

[16] Except if the search query concerns all occurrences of a given location, because this would generate a full list of all texts and isolated sentences coming from this location, and thus would have merely no interest here since it is already possible to get this information from elsewhere in the database.

[17] Of course, it's also possible to select all occurrences from the results list in order to generate an exhaustive work corpus from the database.

[18] If the sentence comes from a text.

[19] Whether isolated sentences or sentences from a text.

[20] For some interesting applications of this feature and concrete cases, cf. (Georges, not yet published, b).

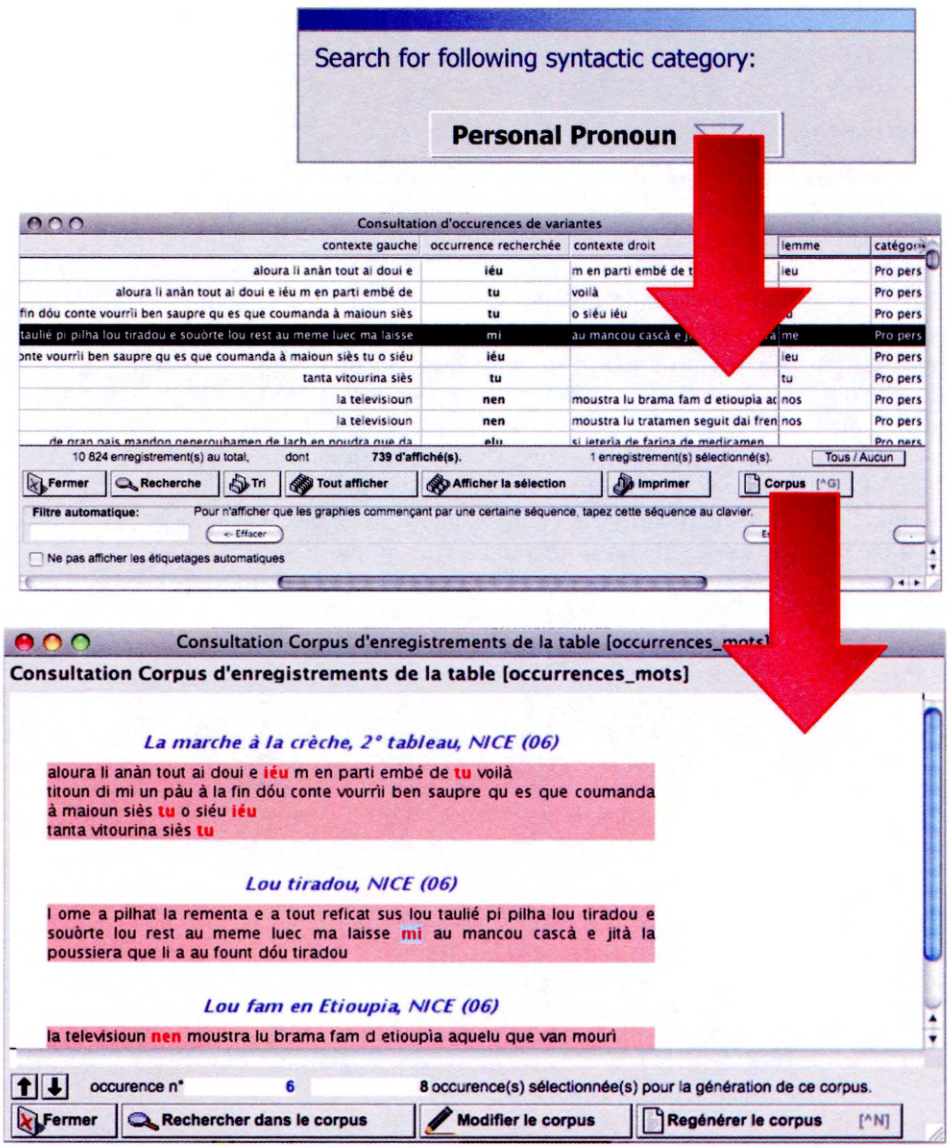


Fig. 7

Search by syntactic category and work corpus generation features

a given parameter (no lexically realised subject pronoun, for example) and also at least one sentence with another given parameter in the same location (evaluating an hypothetical negative correlation between two parameters within a dialect).

In a generative theoretical framework, one could for example use this feature to try to check if (Rizzi 1982)'s correlation between *pro drop* and *free inversion* of sub-

ject and verb can be verified (or invalidated) on the field by dialectal data, such as Occitan dialects.[21]

Conclusion

The MMS module of the THESOC contains flexible tools and features that can be used of several maners and under different theoretical frameworks. Therefore it can fully play its role of a computer tool aiding syntactical and morphosyntactical research on Occitan dialects on a comparative basis.

Currently, we are working on adding cartographic functionalities to MMS, that would fosters the use of MMS for diatopic variation study, giving the users the opportunity to visualize variation of some syntactic or morphosyntactic features among the occitan domain and the possibility to generate maps on demand to illustrate and to support their hypothesis.

Aside from the diatopic perspective, the data entries in the database also contain a "date" field to indicate for each text and each single sentence, at which date they have been collected on the field. So, if enough data is available, it would be theoretically possible to study variation also from a diachronic point of view.

References

Alibert, L., 1966, *Dictionnaire occitan-français d'après les parlers languedociens*, Institut d'Etudes Occitanes, Toulouse.

Dalbera, J.-P., 1994, *Les parlers des Alpes-Maritimes, étude comparative, essai de reconstruction*, Association Internationale d'Etudes Occitanes.

Eynaudi, J., 1932, *Dictionnaire de la langue niçoise*, Imprimerie de «l'Eclaireur de Nice», Nice.

Georges, P.-A., 2009, «Présentation de la base Textes associée au THESOC», in *Actes du colloque La dialectologie hier et aujourd'hui (1906-2006)*, Lyon, 2006, B. Horiot (ed.), pp. 81-94.

— (not yet published), a, «Le THESOC: bases de données et outils d'analyse consacrés à l'étude des dialectes occitans», in *Actes du colloque Bases de données, Méthodes, Modèles de description: de nouvelles perspectives pour la recherche sur les langues régionales et minoritaires?*, Tübingen, 2008, Stauffenburg Verlag (DeLingulis).

— (not yet published), b, «Les chaînes de clitiques: l'outil informatique au service de l'analyse comparative», in *Actes du colloque Mémoires du terrain: enquêtes, matériaux, traitement des données* (Lyon, 2009).

Mistral, F., 1979, *Lou Tresor Dóu Felibrige ou dictionnaire provençal-français*, Culture Provençale et Méridionale, Raphèle-lès-Arles.

Nauton, P., 1957-1963, *Atlas Linguistique et ethnographique du Massif Central*, Editions du CNRS, Paris.

Oliviéri, M., 2003, *Constitution d'une base de textes occitans*, 36ème colloque de la Societas Linguistica Europaea: Linguistique et Corpus, Lyon, 4-7 septembre 2003.

Rizzi, L., 1982, *Issues in Italian Syntax*, Foris, Dordrecht.

Séguy, J., 1973, «Les atlas linguistiques de la France par régions», *Langue Française* 18/1, 65-90.

[21] Dialects from the northern part of Occitania show overt pronoun subjects, whereas the majority of the Occitan dialects are *pro drop*. Comparing these Occitan dialects, which are all closely related, within a microvariational approach, could therefore shed new light on this hypothetic correlation.

NEW METHODS FOR THE STUDY OF GRAMMATICAL VARIATION AND THE *AUDIBLE CORPUS OF SPOKEN RURAL SPANISH*

Inés Fernández-Ordóñez
Universidad Autónoma de Madrid

Abstract

The Audible Corpus of Spoken Rural Spanish *(or COSER after its Spanish abbreviation) is a corpus of oral interviews which aims to study dialect grammar in the Iberian Peninsula. In this paper COSER characteristics and methodology are described and compared to atlases regards the research of dialect grammar. Thanks to COSER, a number of Spanish dialect syntax issues which were partially known or fully ignored have been researched, the geographical distribution of these features has been sometimes considerably broadened, and traditional explanations have been replaced by new ones based on a better knowledge of the data. Thus, the index of grammar phenomena deserving further research has been enlarged. In addition, grammar variation phenomena have showed new areal configurations in Spanish dialectology, and moreover, the study of dialect grammar has also revealed itself as an important source for a better understanding of many cross-linguistic principles.*

Key words: *Corpora of oral interviews vs atlases, dialect grammar*

Until recently, the study of dialectal variation of Spanish in the Iberian peninsula has been based on various regional atlases and those scarce dialectal monographs which devoted particular attention to Spanish (in contrast to the more numerous ones focused on the Asturian-Leonese and Aragonese linguistic domains). Both in atlases and monographs, dialectologists pay more attention to phonetic and lexical variation than to grammatical variation and data have usually been collected by means of a questionnaire. The *Audible Corpus of Spoken Rural Spanish* (referred hereafter as the Spanish abbreviation COSER, *Corpus Oral y Sonoro del Español Rural*) is a corpus made up by recordings of rural speech which started to be compiled in 1990 to supplement those traditional sources and it has been growing since then.

1. An overview description

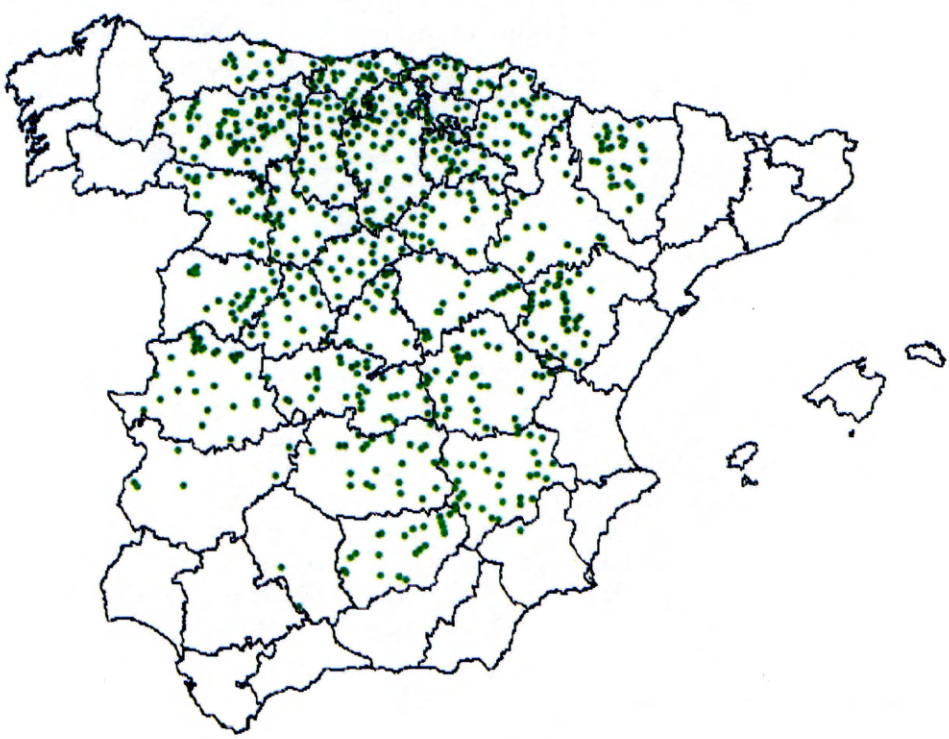
COSER is a corpus of oral interviews restricted to the speech of informants who were deemed interesting for traditional dialectology: rural speakers, elder if poss-

ible, of low education and natives of the place where they were interviewed. Actually, COSER has the same type of informants as linguistic atlases and many dialectal monographs, although its methodology and objectives are different. So far (i. e. the year 2009), 1,408 informants have been recorded, among whom 44% were men and 55,9%, women. The average age of the informants was 72.9 years, being slightly higher in men (73,8) than in women (72).

Regarding the number of informants of each location, in general one single person has preferably been thoroughly interviewed in COSER, either a man or a woman. Nevertheless, recording conditions have sometimes not allowed to avoid interruptions from other individuals (generally members of the family or acquaintances who, drawn by such an extraordinary event as the interview, cannot resist the temptation to take part in the interview by giving their own testimony). Thus, although up to 1,408 informants have been recorded in COSER, most of the times only one informant per location has actually been thoroughly surveyed as desired (almost the half). But sometimes we have made more than one interview per location.

Interviews have been carried out so far in 754 rural enclaves of the Iberian Pensinsula, mainly in the Centre and the North. As shown in the following map, the point density is comparable to that of regional atlases, or even denser.

COSER consists presently of 940 hours of recording, but this number increases every year thanks to new survey campaigns. The final objective is to obtain record-



Geographic distribution of COSER locations (2009)

ings of Spanish spoken in rural areas of the whole Iberian Peninsula. As can be seen in the map, the South is the main area still needed to be interviewed.

The average duration of the recordings is one hour and fifteen minutes (75 minutes) per location, although it may range from just half an hour up to more than two hours and a half. The quality of the data recorded is not directly proportional to the duration, since there are excellent and very informative recordings of just half an hour, whose results are comparable to those obtained in a longer session.

2. Methodology

The methodology used in COSER has consisted in oral interviews, aimed by part of the interviewers at some subjects of traditional country life. The fact that the interview is focussed on such specific subjects does not prevent that, after some time and having gained the informant's confidence, interest is aimed at other subjects, such as education, personal hopes and experiences, life or family, depending on the level of easiness and spontaneity shown by the informant. The decision of focusing the interview on specific subjects related to rural life "of former times" has much to do with the fact that, in order to accept to be interviewed, potential informants need to recognise themselves as experts on a way of life in decline. This knowledge is a product of their own personal experience and age and gives them informative "authority" in front of the urban interviewer. Informants accept the interview as they realize that we are interested to document a way of life in decline about which very few have hardly any memory at all and which they know they are expert on. We think that the informants' spontaneous cooperation would be much more difficult if they would be required at first to be interviewed on personal views or experiences, linguistic matters or other aspects beyond rural life. The fact that the interviewing team has insisted on their specific interest in the strictly local tradition, in contrast to that of other rural locations, as well as in the exclusive informant's condition as recipient of such tradition, has been on many occasions a decisive factor for accepting the interview.

Informants are always randomly contacted, with no previous actions, among the local inhabitants fulfilling the above mentioned requirements. Due to the experience, not much gratifying, of some interviews on account of the informants' low communication ability (people not much willing to speak, who answered with very short sentences or just in monosyllables) it was decided to add subsequently the condition of loquacity (" informants who like talking") to the informants' selection protocol.

Interviewers are recruited among the students of Dialectology at the Autonomous University Autónoma of Madrid, after a methodological preparation. The students work in groups of four people, as most, to collect the data, but just one person acts as interviewer each time while the rest write down all external information about the informant's characteristics that could be relevant for linguistic analysis, and a linguistic profile of the phenomena documented. These notes are kept together with the recording.

Regards the selection of survey points, it has always intended to cover the communication network of the surveyed area —denser or loser, depending on the number of existing locations—. At the same time, geography and population have

been taken into account. In general, small villages have been preferred, and all the geographical major divisions of a certain area have been surveyed.

3. Aims

COSER is a corpus that aims to study grammatical variation. The study of dialect grammar of Peninsular Spanish has been almost non existent until recently. Linguists have not paid much attention to it, mainly because there were no sources available to tackle these issues. Morphosyntax is traditionally an aspect hardly represented in dialectal monographs and in questionnaires of linguistic atlases. Therefore, it was generally presumed an apparent uniformity in Peninsular Spanish syntax together with the overall existence of certain non-standard uses, which were just mentioned to be avoided.

In this respect, COSER has proved especially useful since it provides the study of non-standard grammatical solutions, which are usually systematically avoided in written language and in the speech of sociocultural groups of higher education. Standard languages seem to have a lower tolerance towards grammatical variation. Thus, this type of variables is frequently subject to a sociolinguistic filtration which may alter the linguistic principles that explain their original function. For that reason, Chambers (1995) has proposed, as a sociolinguistic universal, the qualitative character (presence/absence) of grammatical variables in the social scale, in contrast to the quantitative character of phonetic variables. This is the case, for instance, of the non-standard uses of the unstressed pronouns known as *leísmo, laísmo y loísmo.*[1] Thanks to the sociolinguistic interviews of Klein-Andreu (1979, 1981, 2000) and COSER (see Fernández-Ordóñez 1994, 1999), we can know nowadays that what grammarians considered as deviated uses of the regular pronominal use are in fact partial manifestations of alternative pronominal paradigms in which pronouns are selected according to linguistic principles different to those applied in Standard Spanish. Some of these paradigms, like the Castilian referential paradigm, are only fully present in the speech of sociocultural groups of lower status. As the social status becomes higher, most of the characteristic uses of these paradigms (*leísmo* meant for inanimate objects, *laísmo* and *loísmo*) are discarded. This sociolinguistic distribution has traditionally confused its correct interpretation, since most scholars have drawn their hypotheses on this matter exclusively on the partial data offered by the written and cultivated language (in which *leísmo* for a masculine person is accepted whereas the other *-ismos* are normally rejected). COSER data have allowed thus to understand grammatical variables whose linguistic rules became confused as they hardly entered into the standard language or did not enter at all.

In addition, the development of the interview enables to research the use of any grammatical phenomenon in a real context of use: instead of the isolated, out of context and unnatural sentences typical of a questionnaire, the interview collects sen-

[1] *Leísmo* is the use of the dative pronoun *le* instead of the accusative pronouns *lo* and *la* as direct objects. *Laísmo* is the use of the accusative pronouns *la* and *las* instead of the dative pronouns *le* and *les* as indirect objects, and *loísmo* is the use of the accusative pronouns *lo* and *los* instead of the dative pronouns *le* and *les* as indirect objects.

tences uttered in a real speech, in which it is possible to investigate contrastive values, and pragmatic motivations and inferences related to a specific structure. Thus, for instance, data from COSER enable to understand better a structure which existed in Old Spanish and is only found nowadays in some specific rural varieties with a clear focal value: the use of the article followed by the possessive adjective (*el mi hijo-the my son*), which in these varieties is used alternately with the regular emphatic possessive structure in Spanish (*el hijo mío-the son of mine*). The focal character of the structure explains that both are preferably applied with possessives of the first and second persons, relating to the speaker and listener, and with objects highlighting the relationship between possessor and possessed, aspects which may be difficult to record in sentences isolated from speech such as in atlas questionnaires or those sporadically quoted in dialectal monographs.

Moreover, dialect grammar phenomena need a large amount of sentences (and possible variants of sentences) to be studied in order to get a fine analysis of the contrasting uses and to have the possibility of quantifying data: given a specific linguistic variable, the interview enables to quantify the variants in a specific location as well as distinguishing contexts of occurrence.

Instead, in traditional atlases this quantifying is not usually possible since one single answer is normally given for each location and because very few questions related to one specific variable are included. As a result, minority variants of one variable seldom appear in atlases. This conclusion is drawn for instance by the study of a grammatical use found in the Central and Northern area of the Iberian Peninsula, i.e., the use of the conditional indicative (*-ría*) instead of the imperfect subjunctive (*-ra /-se*), a use extended to all type of syntactic contexts accepting the imperfect subjunctive in Spanish (Pato 2004). This use had been recorded in atlases, although quite insufficiently, as they omitted the fact that the imperfect subjunctive is not only replaced by the conditional indicative *-ría* (majority variant), but also by the imperfect indicative *-ba* (minority variant). Although both variants exist, their proportion of use is not equivalent, which accounts for the fact that the minority variant was hardly recorded by atlases: when the imperfect subjunctive is replaced by these forms of indicative, *-ría* was prevalent in 96% of the cases, whereas *-ba* appeared just in 4% of the cases.

Data quantifying is not impossible from data obtained by atlases, but it is statistically more reliable if data come from a corpus like COSER. First of all, because the phenomenon is sometimes recorded in contexts that were unexpected when atlas questionnaires were designed. This was indeed the case for *leísmo, laísmo* and *loísmo*.[2]

---

[2] Thus, the atlas for all the Romance languages in the Iberian Peninsula, *ALPI* (see Heap 2003) devotes five questions to personal *leísmo* (350 *A Miguel le cogieron preso (Michael was held prisoner*, 351 *Le llevaron a la cárcel (He was sent to prison)*, 352 *Al padre le vieron llorando (The father was seen crying)*, 353 *A los niños les socorrieron los vecinos (The children were helped by neighbours)*, 355 *Al enfermo hay que cuidarle (The sick person must be looked after)*): apart from the high number of questions devoted to record the same phenomenon, the standard character of masculine personal *leísmo* is shown by the fact that the questions of the questionnaire are expressed according to a *leísmo* solution. In contrast, those devoted to *loísmo* (356 *Al niño le pusieron un vestido (The child was dressed in a dress)*, 357 *Tráete los candiles para echarles aceite (Bring the oil lamps in order to add some oil to them)* and to *laísmo* (359 *A la madre no le dieron la limosna (the mother was not given any alms)*, 360 *Aquella desgracia le costó a ella la vida*

This problem also happens in the recording of the use of *-ría / -ba* instead of *-ra / -se*, since atlases had planned to record this use preferably in the protasis of conditional sentences and in desiderative sentences using *ojalá (I wish, I hope)*,[3] while in fact the phenomenon appears in complement, relative, final, concessive, causal clauses, etc: i.e., in any subordinate clause where the imperfect subjunctive is likely to be found in Spanish. In the case of both pronominal and verbal uses, the atlas questionnaire records as partial deviations of the general use what is actually an alternative use controlled by different linguistic principles and which takes place in a significantly wider range of contexts.

Secondly, the number of records regarding the phenomenon obtained in any interview is always necessarily higher than that provided by an atlas questionnaire, even if all syntactic contexts likely to show this phenomenon had hypothetically been included. It is this significant number of records what enables to detect the presence of minority variants, which are in fact concealed in atlases. Therefore, in statistical terms, data quantifying from a corpus like COSER enables to draw conclusions far closer to reality as regards linguistic uses. The quantity of data makes it possible to identify focal areas for a linguistic phenomenon, which is not always possible with data coming from atlases, and to apply statistical tests like logistic regression, en-

---

(That misfortune cost her her life), 361 *A las hermanas les enviaron unas cartas (Some letters were sent to the sisters)*, 362 *A la yegua le cansa el trabajo (The mare gets tired working)*), are expressed with the regular solutions of the pronominal case. No questions related to masculine non-personal *leísmo* were planned. Nevertheless, questions 312 and 313, intended to record the conjugation of the verb *vaciar* (*to empty*), might also allow to research non-personal *leísmo* (312 *¿Dónde vacían el cántaro? (Where is the jug emptied?)*, 313 *No lo vacíes en la calle (Do not empty it in the street)*). Regional atlases do not improve much ALPI questionnaire. *ALEANR* devotes less entries of its questionnaire to such uses and besides, most of them are exact to some of those included in the *ALPI* questionnaire (it reproduces thus those numbered 350-351, 353, 356, 359, 362 corresponding to maps 1708-1711). There are no questions which enable to record non-personal *leísmo*, although there is one question which enables to record femenine personal *leísmo* (*A la madre la vio en la calle (The mother was seen in the street)*, map 1713). Only *ALECant* and *ALCyL* include new questions aimed at non-personal *leísmo* (with animate antecedents, *Al lobo lo vimos (We saw the wolf)*, maps 1194 and 118, respectively, and inanimate, *El libro lo olvidé en casa (I forgot the book at home)*, *ALECant* 1195, *El paquete lo olvidé (I forgot the parcel)*, *ALCyL* 116). These two regional atlases also reproduce questions 350, 352-353, 356, 359 and 362 of *ALPI* (*ALECant*, 1243, 1245-1247, 1192, 1197; *ALCyL*, 111-114, 117, 120) and 1713 of *ALEANR*. In *ALECMan* the questions 350-353, 356, 359 y 362 from *ALPI* are included, and new questions are added to research *laísmo*: *A las niñas no (les/las) gusta estudiar (The girls do not like studying)*, *La torre desde aquí se (le/la) ve (The tower, from here it is seen)*. None of the atlases enables to notice the absence of *leísmo* when the antecedent is a masculine mass object (like *pan (bread)*, *vino (wine)*, *trigo (wheat)*, etc.) or the use of *lo* to refer to femenine mass objects (*agua (water)*, *miel (honey)*, *manteca (butter)*, etc), not even *ALECant*, in spite of the fact that Cantabria is a region where the existence of the mass neuter was well-described.

[3] Four relevant questions were included in *ALPI* (386 *Si tuviera dinero lo compraría (If I had money, I would buy it)*, 387 *Si estudiase aprendería (If I studied, I would learn)*, 388 *Si pudiera la mataría (I would kill her if I could)*, 390 *Ojalá lloviese (If only it would rain)*), of which the first and last ones were reproduced in *ALEANR* (maps 1704, 1706), in *ALECant* (maps 1216, 1220) and in *ALCyL* (148, 152). *ALEANR* enriched the syntactic contexts by adding an entry which included a noun clause (1705 *Le dijo que trajera un pan (He told him to bring some bread)*), which *ALECant* and *ALCyL* also inherited (maps 1218 and 150, respectively). *ALECant* added in turn a concessive clause to the list (1217 *Aunque pudiera no lo haría (I would not do it, even if I could)*), reproduced in *ALCyL* (map 149). Finally, only the *ALCyL* questionnaire includes a final clause (151 *Esto te lo dije para que fueras bueno (I told you this so that you were a good boy)*).

abling to assess the simultaneous influence of several variables on the phenomenon manifestation. In the case of the use of *-ría* / *-ba* instead of *-ra* /*-se* in the Castilian varieties, it has been proved that the most widespread opinion according to which the protasis of the conditional sentences was considered as the origin of this phenomenon, was not actually correct. Instead, the prevalence of *-ría* and *-ba* over the subjunctive forms *-ra* / *-se* was first found in complement clauses, extending next to the relative and dependent adverbial clauses and finally, to the conditional and final clauses, as well as the rest of syntactic contexts (Pato 2003, 2004).

COSER can measure the differences which may be found in the speech of sociocultural groups with a lower education in rural areas. It therefore complements both linguistic atlases and the different corpora of cultivated and urban speech which have been compiled or are planned to be so in the Spanish-speaking world. The uniformity in the methodology used makes it useful to measure both the linguistic distance which separates different areas (physical distance) and the linguistic distance which separates this social group from others, like for instance, that of speakers with a higher sociocultural level or that of younger speakers (social distance). Although the proportion of men and women interviewed is not identical (55,9% women vs. 44% men), the number of speakers of each gender is statistically representative and also allows to investigate linguistic differences associated with gender.

4. Problems

Although COSER offers new possibilities for the research of dialects, it also has shortcomings and problems. First of all, as it will be obviously well-known to anyone who has ever carried out fieldwork, success is never assured, and an interview starting under the same conditions may be optimum or dreadful. Thus, not all interviews are equally suitable or informative, depending on the informants' willingness, the interviewers' skills as well as the interaction between them; however, no testimony should be disregarded for that reason, since there is always some valuable information. Secondly, COSER methodology cannot avoid the problem of accommodation between the informant and the interviewer, or the challenging representativeness of the informant ramdomly chosen. Nevertheless, our experience suggests that the quantity of the data allows to circumvent these potential problems, since the data always show geographical coherence and make it possible to identify those informants who could be considered anomalous with their area and scarcely representative. Thirdly, the long time elapsed between the first interviews (1990) and the last (2009) and future ones (2010-) does not assure a total intercomparability of the data. However, the planning of the interviews has intended to research areas related to each other (for historical or linguistic reasons). For example, North and Center Castile has been researched between 1990 and 1995; East Castile and Aragón between 2001 and 2008, and so on. The documentation of an issue related to an area has usually finished in a reasonable time. Finally, the comparability of data provided by a questionnaire is rarely obtained with the methodology of the interview, in which researchers may try to obtain some specific data, but without ever being certain if their aim will be successfully achieved. On the other hand, linguistic atlases offer a type of information which is not provided by corpora like COSER. Oral interviews have proved especially productive

to record phenomena of grammatical character but not as far as lexis is concerned. Since the data come from almost open conversations, the words recorded in COSER are not always repeated and no conclusions are drawn comparable to those of an atlas as regards vocabulary. Although the development of sociolinguistics has shown multiple limitations of atlas methodology, it is important to bear also in mind that, since there are no speech recordings of past times which are equivalent to current recordings (and there is no human means to obtain them), atlas data remain thus a precious testimony —however imperfect it may be— for the study of rural speech (as well as the grammar, as proved by works like Heap's, 2000 and Benito 2009). Therefore, COSER aims to supplement the material collected in linguistic atlases as well as in other type of dialectal sources. COSER data are a supplement which opens up enriching prospects for the study of dialectal grammar.

It always has to be born in mind that success and adequacy depend on aims and results considered together with resources. Firstly, as regards aims, it is not COSER aim to study the lexicon or to obtain the elicitation of a precise structure or use in every place —which just could be done by means of a questionnaire—. Instead, we try to register spontaneous speech by following a similar thematic protocol in order to document, identify and study dialect grammar phenomena (known and unknown). Intercomparability is expected to be achieved between areas and not between locations. Secondly, as regards the relationship between aims and resources, COSER recordings started as a modest activity linked with the teaching of Dialectology, which aimed to document and understand grammar phenomena absent in linguistic atlases and traditional dialect monographs. Therefore, it can prevent neither the accomodation problem nor the time elapsed because it was not globally designed and planned as a research project, although COSER has also received support of several research projects. COSER interviews were collected by each year Dialectology students and annual campaigns have surveyed as many locations as possible depending on the availability of number of students, time and funding.

Despite all this, COSER is still a valuable source to study dialect grammar since anything alike exists for rural Spanish. By now only COSER offers a collection of oral interviews for the almost whole rural Spanish speaking territory. Moreover, the interest of COSER recordings reveals itself in the progress made in our knowledge of Spanish dialect grammar since 1990. This is true regarding the study and better understanding of dialect phenomena which were hitherto partially known —or even completely ignored by grammarians and dialectologists up to now—. The index of dialect grammar issues deserving further research has been considerably enlarged too. I will deal with these issues in the following sections.

5. Outcomes

There are hitherto three kinds of results coming out from COSER recordings: specific progress made in our knowledge of dialect grammar issues, general progress made in Spanish dialectology regards areal configuration, and finally theoretical outcomes.

Regarding the specific results, a number of dialect grammar issues so far have been researched and their geographical areas have been defined —confirming or dis-

carding areas already known, or describing areas for the first time—, a better description and understanding of the dialect structures have been achieved, and an explanation has been proposed. I will present three examples. Firstly, the analysis of the data from rural speech recordings has enabled to establish the exact geographic delimitation of the areas where the phenomena traditionally known as *leísmo, laísmo* and *loísmo* are found, while it has also proved that the apparent lack of coherence in their frequency is actually due to the existence of several pronominal paradigms, alternative to the regular paradigm of Spanish. The data from these paradigms were mixed in earlier studies altering thus the interpretations (Fernández-Ordóñez 1994, 1999, 2001). Secondly, COSER interest is enhanced by the fact that it has recorded dialect phenomena completely ignored by grammarians and dialectologists up to now. The best example is mass neuter agreement. This agreement was traditionally known in Central and Eastern Asturias and Cantabria, but went fully unnoticed in Castile. Thanks to COSER recordings, the geographical area with mass neuter agreement has been considerably enlarged to the South and a global explanation has been developed (Fernández-Ordóñez 2007, 2006-2007). Finally, COSER data quantifying has also made possible to demonstrate that the syntactic locus of origin for the replacement of subjunctive by indicative are complement clauses in Northern Spanish, and not the conditional clauses, and that the focal area is located in Castile, and not in the Basque Country, as usually believed (Pato 2003, 2004).

Our general knowledge of Spanish dialectology has improved so far in that new dialect areas of Peninsular Spanish have clearly emerged, namely a Western Spanish vs. an Eastern Spanish both layed out from North to South (Fernández-Ordóñez 2009a). Also contrast between urban speech and rural speech regards grammar have made clear sociolinguistic selection of some grammatical features. On the other hand, contrast between data from old atlases and present oral data has showed the historical development of some dialect aspects, whether maintenance (subjunctive replacement Pato 2004 or inflected infinitives Pato & Heap 2009) or decline and loss (of some pronominal phenomena, see Heap 2006, Pato 2009).

Studies coming out from COSER have confirmed some theoretical generalizations concerning dialectology and linguistic typology. As for dialectology, concepts like mixed lect, fudged lect and scrambled lect, suggested by Trudgill (1986) as typical phonetic solutions of transition areas, have been confirmed to be operating in grammar variation (Fernández-Ordóñez, under review). As for typology, dialect grammar has revealed as an important source for a better understanding of many cross-linguistic principles which opens up new ways to test their validity and to achieve their refinement (see Kortmann 1999, 2004a, 2004b). For instance, out from COSER data and research, a proposal of refinement of the Agreement Hierarchy (Corbett 2006) has been proposed (Fernández-Ordóñez 2006-07), and two different syntactic pathways for the development of gender as a lexical category have been identified (Fernández-Ordóñez 2009a).

6. Prospects

A small sample of COSER materials is now available in the Internet (*www.uam.es/coser*) as audio and text files: the recordings of 32 locations from 8 prov-

inces (c. 40 hours). The current project "Variation and change in Peninsular Spanish Syntax" (FFI2009-10817) (2010-2012), funded by the Spanish National Research Funds, opens new prospects for COSER. Aiming to integrate COSER into the European Dialect Syntax (EDISYN) network, the project envisages two areas of activity, as is the case in the other European projects concerning dialect syntax (for example DYNASAND, see Barbiers 2006).

On the one hand, the project will concentrate on the collection, processing and storage interview recordings which will allow for the study of syntactic variation in Spanish. To that end, our goal will be to digitalize, transcribe and web-publish at least 150 hours of recordings from COSER. So the interviews of 120 localities from 30 provinces —4 per province— will be available in the Internet as audio and texts files. It is also aimed the alignment of sound and text and the morpho-syntactic tagging of those 150 hours. At the end, there will be a searchable database in the Internet with several searching possibilities (by lemma, area or location, sex, age, and tags). The searched data will be retrieved in small listenable paragraphs.

On the other hand, the project will focus on the study of a set of syntactic variation phenomena within the framework of comparative and typological linguistics. This approach combines modern and historical data in order to achieve a unified analysis of the mechanisms which underlie and condition language change. To this end we will endeavour to integrate into our overall explanation current and historical linguistic perspectives with geolinguistic and, to some degree, sociological aspects of a set of phenomena which vary grammatically in Peninsular Spanish and which to date have only been partially investigated. The following grammatical issues will be part of ongoing research for the next three years: quirky dative marking in Eastern Spanish; possesives in Northern and Western Spanish, quirky quantifiers agreement; dative objects (with verbs such *help, call, follow, warn, scold*) and structures alternating dative and accusative; reflexive, passive, impersonal and middle diathesis associated with reflexive verbs and reflexive structures; inflected infinitives; diminutive suffixation related to word classes, semantic proprieties and syntactic functions, as well as to geographic areas; and anteriors in Peninsular Spanish dialects.

References

ALCyL: Alvar, M., 1999, *Atlas lingüistico de Castilla y León*, 3 vols, Junta de Castilla y León, Valladolid.

ALEANR: Alvar, M. et al., 1979-83, *Atlas lingüístico y etnográfico de Aragón, Navarra y La Rioja*, with the collaboration of A. Llorente, T. Buesa and E. Alvar, 12 vols, Institución Fernando el Católico & CSIC, Zaragoza.

ALECant: Alvar, M. et al., 1995, *Atlas lingüístico y etnográfico de Cantabria*, 2 vols, Fundación Marcelo Botín, Madrid.

ALECMan: García Mouton, P. & F. Moreno Fernández (dirs.), *Atlas Lingüístico (y etnográfico) de Castilla-La Mancha*, Universidad de Alcalá, <http://www2.uah.es/alecman>

ALPI: Navarro Tomás, T. (dir.) et al., 1962, *Atlas lingüístico de la Península Ibérica*, vol. 1, *Fonética*, with the collaboration of F. de Borja Moll, A. M. Espinosa [junior], L. F. Lindley Cintra, A. Nobre de Gusmão, A. Otero, L. Rodríguez-Castellano and Manuel Sanchis Guarner, Madrid, CSIC.

Barbiers, S. et al, 2006, *Dynamic Syntactic Atlas of the Dutch dialects* (DynaSAND), Amsterdam, Meertens Institute <*http://www.meertens.knaw.nl/sand/*>.

Benito, C. de, 2009, *Descripción y análisis del pronombre concordado con el sujeto en el Atlas Lingüístico de la Península Ibérica*, Master's degree memory, UAM.

Chambers, J. K., 1995, *Sociolinguistic theory*, Blackwell, Oxford.

Corbett, G. C., 2006, *Agreement*, Cambridge U. P.

COSER: Fernández-Ordóñez, I. (dir.), 2005-, *Corpus oral y sonoro del español rural*, Universidad Autónoma de Madrid, Madrid <*http://www.uam.es/coser*>.

Fernández-Ordóñez, I., 1993, «Isoglosas internas del castellano. El sistema referencial del pronombre átono de tercera persona», *Revista de Filología Española* LXXIV, 71-125.

—, 1999, «Leísmo, laísmo y loísmo», in I. Bosque & V. Demonte (eds.), *Nueva gramática descriptiva de la lengua española* (3 vols.), Espasa-Calpe, Madrid, vol. 1, 1317-1397.

—, 2001, «Hacia una dialectología histórica. Reflexiones sobre la historia del leísmo, el laísmo y el loísmo», *Boletín de la Real Academia Española* LXXXI, 389-464.

—, 2006-07, «Del Cantábrico a Toledo. El «neutro de materia» hispánico en un contexto románico y tipológico», *Revista de Historia de la Lengua Española* 1, 67-118; 2, 29-81.

—, 2007, «El neutro de materia en Asturias y Cantabria. Análisis gramatical y nuevos datos», in I. Delgado Cobos & A. Puigvert Ocal (eds.), *Ex admiratione et amicitia. Homenaje a Ramón Santiago*, Ediciones del Orto, Madrid, 395-434.

—, 2009a, «Los orígenes de la dialectología hispánica y Ramón Menéndez Pidal», in X. Viejo Fernández (ed.), *Cien años de Filoloxía Asturiana (1906-2006)*, Alvízoras & Trabe, Oviedo, 11-41.

—, 2009b, «The development of mass / count distinctions in Indo-European languages», in V. Bubenik, J. Hewson & S. Rose (eds.), *Gramatical Change in Indo-European Languages* (Current Issues in Linguistic Theory 305), John Benjamins, Amsterdam / Philadelphia, 55-68.

— (under review), «Dialect areas and linguistic change: Pronominal paradigms in Ibero-Romance dialects from a cross-linguistic and social typology perspective».

Heap, D., 2000, *La variation grammaticale en géolinguistique: les pronoms sujet en roman central*, Lincom Europa, München.

—, 2003-: *Atlas lingüístico de la Península Ibérica. ALPI searchable database*. University of Western Ontario, London, Ontario <*http://www.alpi.ca*>.

—, 2006, «Secuencias «invertidas» de clíticos: un cambio (¿?) en tiempo real», in J. J. de Bustos & J. L. Girón Alconchel (eds.), *Actas del VI Congreso Internacional de Historia de la Lengua Española* (Madrid, September 29-October 3, 2003), Arco/Libros, Madrid, I, 785-98.

Klein-Andreu, F., 1979, «Factores sociales en algunas diferencias lingüísticas en Castilla la Vieja», *Papers: Revista de Sociología* 11, 46-67.

—, 1981, «Distintos sistemas de empleo de 'le, la, lo': perspectiva sincrónica, diacrónica y sociolingüística», *Thesaurus* XXXVI, 284-304. (Reprinted in O. Fernández Soriano (ed.), *Los pronombres átonos*, Taurus, Madrid, 1993, 337-353).

—, 2000, *Variación actual y evolución histórica: los clíticos le/s, la/s, lo/s*, Lincom Europa, München.

Kortmann, B., 1999, «Typology and dialectology», in B. Caron (ed.), *Proceedings of the 16th international Congress of Linguists*, CD-ROM, Elsevier Science, Amsterdam.

—, 2004a, «Why dialect grammar matters», *The European English Messenger* XIII, 24-29.

— (ed.), 2004b, *Dialectology meets typology. Dialect Grammar from a Cross-Linguistic Perspective*, Mouton de Gruyter, Berlin /New York.

Pato, E., 2003, «Contextos neutralizadores de la oposición modal y relaciones de alomorfismo desde el español medieval: Las formas *cantase*, *cantara* y *cantaría*», *Moenia* 9, 223-252.

—, 2004, *La sustitución de* cantara/ cantase *por* cantaría/ cantaba *(en el castellano septentrional peninsular)*. Universidad Autónoma de Madrid, Madrid <*http://joule.qfa.uam.es/coser/publicaciones/enrique/2_es.pdf*>
—, 2009, «Nivelación lingüística y simplificación: el uso de preposición + *tú* en la historia de la lengua», in E. Montero et al, *Actas del VIII Congreso Internacional de Historia de la Lengua Española* (Santiago de Compostela, 14-18 de septiembre de 2009), in press.
— & D. Heap, 2009, «Plurales anómalos (el morfema verbal *-n*) en los dialectos y en la historia del español», in E. Montero et al, *Actas del VIII Congreso Internacional de Historia de la Lengua Española* (Santiago de Compostela, 14-18 de septiembre de 2009), in press.
Trudgill, P., 1986, *Dialects in contact*, Blackwell, Oxford.

THE APPLICATION OF SPEECH SYNTHESIS AND SPEECH RECOGNITION TECHNIQUES IN DIALECTAL STUDIES

Maria-Pilar Perea
Universitat de Barcelona

Abstract

Speech analysis techniques open new perspectives in the processing of dialectal oral data. Speech synthesis can be useful to create or recreate voices of speakers for extinct languages, to re-edit dialectal material using new technologies or to reconstruct utterances of informants that only were registered in notebooks. Speech recognition, applied to sound dialectal sequences, can make easier automatic transcription of oral texts. In this paper the possibilities of speech analysis techniques in their application to the dialectal studies is described. The presentation is illustrated with the results obtained in different projects.

Key words: *Dialectology, linguistic variation, text-to speech systems, speech recognition, automatic mapping.*

1. Introduction

Given that contemporary dialectology draws heavily on statistical and data processing resources as well as the methods applied in corpus linguistics in the edition and interpretation of its data, it is hardly surprising then that when working with oral texts it should want to take advantage of speech analysis techniques —after all, voice is the raw material obtained by the dialectology researcher from her informant's responses—. Speech analysis techniques —both of synthesis and recognition— are evolving rapidly and are being put to use in many areas of everyday life. Speech synthesis is being used in programs where oral communication is the only means by which information can be received, while speech recognition is facilitating communication between humans and computers, whereby the acoustic voice signals changes in the sequence of words making up a written text.

Speech analysis techniques are opening up new avenues of research in the processing of oral dialectal data. Speech synthesis is useful for creating or recreating the voices of speakers of extinct languages, for re-editing dialectal material using new technologies and reconstructing utterances of informants that have only been recorded in notebooks. Speech recognition, when applied to sound dialectal sequences, can facilitate the automatic transcription of oral texts, so that, first, a phonetic representation can be obtained, and, then, an orthographic representation.

In this paper the possible applications of speech analysis techniques in dialectal studies are described. The discussion is illustrated with results from various finished and on-going research projects.

2. Text-to-speech system (TTS)

In present-day dialectology, notebooks have become archaeological artefacts, replaced by digital and video recording systems. These modern techniques of data collection allow data to be preserved and reproduced as often as required in order to obtain the most faithful transcriptions possible. Earlier data, collected in notebooks, are wonderful dialectal treasures, but lack sound. The development of voice synthesis techniques, which have many useful applications including audiobook production for the disabled, can now provide sound for these dialectal data.

Speech synthesis (Keller 1994) is the process of converting written text into machine-generated synthetic speech. In general, there are three approaches concerning text-to-speech (TTS) systems: *a*) *formant:* this employs a set of rules to synthesise speech using formants, which are the resonance frequencies of the vocal tract; because it does not use human voice the result is a robotic voice; *b*) *concatenative*: this is based on the idea of concatenating pre-recorded human speech units in order to construct the utterance; and *c*) *articulatory*: this tries to model, analogically or digitally, the human articulatory system, i.e. the vocal cords, the vocal tract, etc.

To date, dialectologists have not used speech synthesis resources. In this paper we explain our attempts to put sound to a body of dialectal data for which, due to their age, we have only written documentary records. The dataset is "La flexió verbal in els dialectes catalans" (from now on "Verbal inflection"), by A. M. Alcover and F. de B. Moll, a corpus of almost half a million forms that describes the complete conjugation of 80 verbs from 149 towns of the Catalan linguistic domain. The data were gathered between 1906 and 1928 and published between 1929 and 1932. In 1999 the computerisation process of the materials started: first, a database was created and, later on, methodologies of automatic mapping were applied. Recently, the data were associated with the voice (male or female voices, since the informants belonged to both sexes).

Speech synthesis was possible because the materials do not only include the orthographic answers, but also the corresponding phonetic transcriptions. Without this sort information, the application of these techniques had been impossible.

Next, we explain the steps given in order to obtain the corresponding sounds of the numerous registers of "Verbal inflexion" by means of the speech synthesis techniques, the difficulties surpassed, and the types of syllabic forms that have permitted to obtain a better quality of speech (cf. Perea 2008). The result has been applied to the existing maps, and a sound atlas of Catalan verb morphology with data of the beginning of the 20th century has been obtained.

2.1. The original database and the new computer treatment

In the original edition of *La flexió verbal* ("Verbal inflection"), the sixty-seven verbs studied were classified by conjugation. Different verb tenses (infinitive, gerund, participle, present indicative, past indicative, etc.) of each verb were shown. To

develop the verb paradigm, each verb form was related to a morphological variant, which helps to determine its dialectal scope, and to a phonetic form. Alongside the phonetic variant there was a list of the localities in which this answer was recorded. The localities were represented by numbers (Figure 1).

96 *Segona conjugació*

18. — CREURE

INFINITIU

Creure: krę̯urə 1-31, 34-37, 39, 41-55, 57-58, 60-62, 132, 137-138, 140, 142. krę̯urę 30, 32-33, 36, 38, 40, 58-59, 87, 137. krę̯urę 56. krę̯uri 56, 75, 86, 89. krę̯urę 63-86, 88-107, 109, 111-117. krę̯urę̍r 108, 110. krə́urə 118-121, 123, 126-131, 133-136, 139-141, 143-147. krə́urę 119, 122. krə́urǭ 121, 125. krę̯urō 124. krę̯ura 132. krę̯urę 138-139, 141. krę̯ura 148.

PARTICIPI PASSAT

Cregut: -út 1-3, 5-148. -ę̍t 4. — **Cres:** krę́s 84.

PARTICIPI PRESENT

Creyent: krəi̯ę́n 1-2, 5, 11-13, 15-17, 22-24, 26-29, 33-41, 44-57, 59-62. kri̯ę́n 6. krę̯i̯ę́n 63-67, 71-73, 75-76, 78-80, 82-87, 89, 91, 94. krəi̯ę́nt 133, 141. krai̯ę́nt 148. — **Creguent:** krəgę́n 1, 3-5, 8-12, 14, 16-19, 21-23, 25, 29-32, 39, 41-45, 48-49, 53, 57, 59, 142-147. krę̯gę́n 68, 70, 73, 75, 81, 85, 94. krę̯gę́nt 101-107, 111. krəgę́nt 118, 120, 122, 124-125, 128-131. krəgę́nt 119-121, 123, 125-127, 132-141. — **Creuren:** krę̯urən 7, 18. — **Crevent:** krəbę́n 19-22, 27, 30-32, 40. — **Creent:** krəę́n 58. krę̯ę́n 88, 90, 92-93, 95-100, 109, 112-113, 115. krę̯ę́nt 101, 103, 105, 107-108, 110, 114, 116-117. krəę́nt 133, 135-136, 138, 141. — **Cregüent:** krę̯gu̯ę́n 77.

PRESENT D'INDICATIU

1.ª sg. — **Crech:** krę́k 1-2, 5-14, 16-62, 87, 124, 132, 137-142. krę́k 63-86, 88-117, 148. krə́k 118, 128-131. krək 119-123, 125-127, 133-136, 141, 143-147. — **Creui:** krę̯úi 3, 5, 15-16. — **Cresi:** krę́zi 4.

2.ª sg. — **Creues:** krę̯uəs 1, 9, 11-12, 16-17, 19, 21. — **Creus:** krę̯us 2-3, 5-8, 10, 13-62, 87, 124, 132, 137-142. krę̯us 63-86, 88-117, 148. krə́us 118-123, 125-131, 133-136, 141, 143-147. — **Creses:** krę́zəs 4.

3.ª sg. — **Creu:** krę̯ú 1-62, 87, 124, 132, 137-142. krę̯ú 63-86, 88-117, 148. krə́ú 118-123, 125-131, 133-136, 141, 143-147.

1.ª pl. — **Crèyem:** krəi̯ę́m 1-3, 5-8, 11-15, 22-24, 26-29, 31, 33-39, 41-42, 44-62. krəi̯ę́m 55. krę̯i̯ę́m 63-67, 69, 71-76, 78-80, 82-86, 89, 91, 94. krę̯i̯ę́m 87. krai̯ę́m 148. — **Cresem:** krəzę́m 4. — **Creguem:** krəgę́m 9-10, 14, 16-19, 21, 30, 40, 43. krę̯gę́m 68, 70, 73, 81, 85, 101-106, 108, 111. — **Crevem:** krəbę́m 16-17, 19-23, 25, 27, 30-32, 36, 40, 51. — **Creseveu:** krəzəbę́m 17, 25. — **Creuem:** krę̯uę́m 77. — **Creem:** krę̯ę́m 88, 90, 92-93, 95-100, 104-105, 107-110, 112-117. — **Creym:** krə́i̯m 118-123, 125-131, 133-136, 141, 143-147. krę̯i̯m 124, 132, 137-142.

2.ª pl. — **Creyeu:** -ę̯ú 1-3, 5-8, 11-15, 22-24, 26-29, 31, 33-39, 41-42, 44-62, 87. -ę̯u 55, 63-67, 69, 71-76, 78-80, 82-86, 89, 91, 94, 148. — **Creseu:** -ę̯u 4. — **Cregueu:** -ę̯u 9-10, 14, 16-19, 21, 30, 40, 43. -ę̯u 68, 70, 73, 81, 85, 101-106, 108, 111. — **Creveu:** -ę̯u 16-17, 19-23, 25, 27, 30-32, 36, 40, 51. — **Creseveu:** -ę̯u 17, 25. — **Creueu:**

Figure 1

These data, along with the unpublished material included in the original notebooks, were introduced in a database. Once systematised and completed, they formed a suitable corpus for the creation of a computerised linguistic morphological atlas (Figure 2).

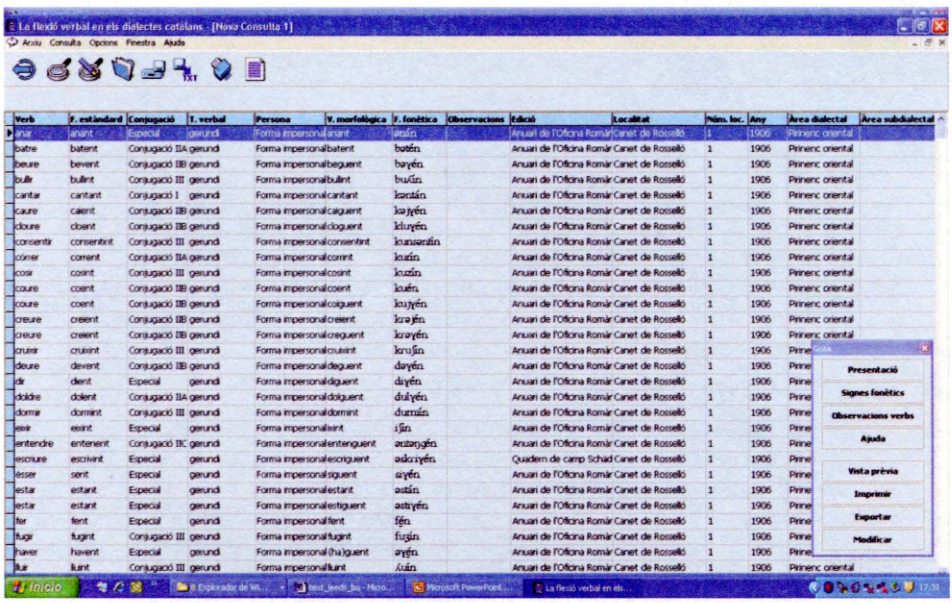


Figure 2

The process of ordering and completing the materials generated a morphological corpus of 470,255 entries (adapted to the IPA alphabet) —cf. Perea (2004)—. In fact, to be synthesised, from the total amount of entries, 1,080 were removed from the verbal database, because, in spite of being included in the notebooks, were forms impossible to be pronounced. They did not adapt to the rules of Catalan pronunciation (i. e. [krúʃɣi], it has to be [krúʒɣi]). Probably, they were written wrongly during the process of data gathering or data transcription. The used registers correspond to 28,161 different single verb forms.

The linguistic atlas (Perea 2005) created by the computer program is a collection of 6,000 potential maps that can be updated as the user wishes. Each map places the phonetically transcribed dates in the various localities surveyed, represented by points. The result (Figure 3) is a linguistic atlas that accomplishes three goals: *a*) it presents a synchronic, morphological and phonetic description; *b*) it shows the formation of different linguistic areas through the distribution of coinciding forms; and *c*) it provides representative material for subsequent study or interpretation of the data.

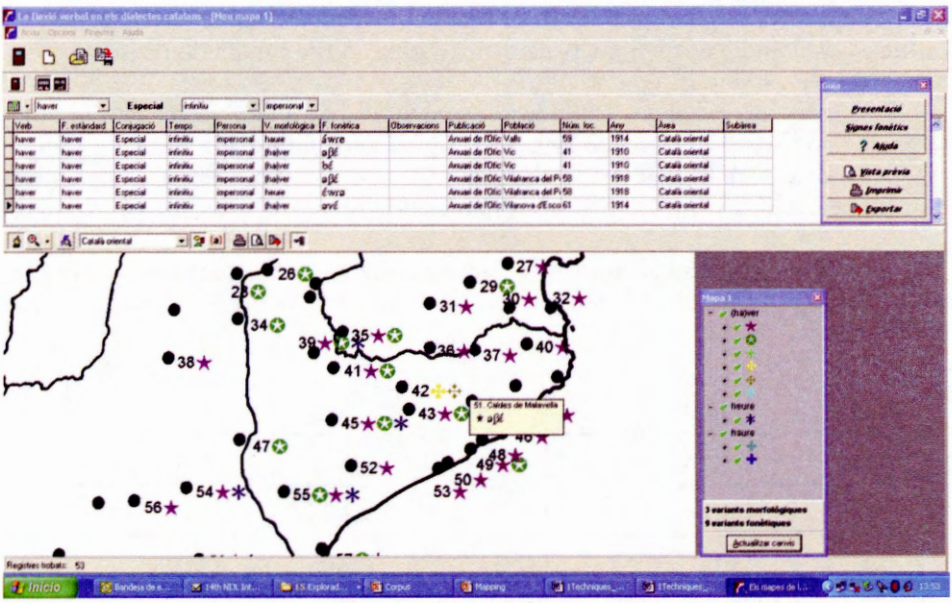


Figure 3

2.2. The stages of speech synthesis

As we have said before, there are three approaches concerning text-to-speech (TTS) systems (*formant, concatenative*, and *articulatory*). Here, we used a concatenative speech synthesis system (Lewis & Tatham 1999) because it gave more natural-sounding output. The system uses a corpus that stores a large inventory of units to be concatenated. To obtain pre-recorded speech units it is necessary to record human voices.

The stages in the process of creating the speech synthesis system were as follows:

b.1) recording human speech;
b.2) developing a speech synthesis system;
b.3) applying the concatenation process;
b.4) editing sound and solving problems.

2.2.1. Recording human speech

The speech corpus was created by recording different real and unreal words pronounced by a number of speakers (male and female). The words contained all the syllables required by "Verbal inflexion" (3,526). Because the original data represent all the Catalan varieties (Eastern Catalan, Western Catalan, Valencian, *Alguerés*, *Rossellonés*, and Balearic), we surveyed different speakers of the four main dialects (Eastern Catalan, Western Catalan, Valencian, and Balearic), especially as regards their idiosyncratic pronunciation. In some cases, and in order to obtain specific sounds, we also surveyed speakers from small towns with particular vowel or consonant pro-

nunciation; for example, speakers from Felanitx or Santa Coloma de Queralt, concerning vowel endings, or from Benavarri, relating to the production of the consonant group [pʎ + vowel].

The survey covered about 1,000 words (monosyllables and polysyllables) in order to gather stressed and unstressed syllabic samples. Each informant pronounced only the part related with his/her own dialect. However, the questionnaire had a main section that coincided with the general pronunciation of Catalan read by two speakers (male and female).

Figure 4 shows a sample of the questionnaire eliciting words that contain syllables with [ɲ] and [ʎ]. On the right we can see the word read; on the left, the syllable obtained.

ɲút	bəɲút
ɲin	báɲin
ɲis	báɲis
ɲi	báɲi
ɲu	báɲu
ɲun	báɲun
ɲus	báɲus
ɲut	báɲut
ɲuk	báɲuk
ɲik	báɲik

ʎən	báʎən
ʎəs	báʎəs
ʎə	káʎə
ʎuk	káʎuk
ʎus	ʎuskám
ʎut	káʎut
ʎui	ʎuiré
ʎúu	
ʎəw	káʎəw
ʎóis	

Figure 4

The speech corpus was recorded at a frequency of 44,100 KHz, using one channel (mono) and 16 bits. The words of the questionnaire were read by the speakers without any specific context in order to obtain a homogeneous pitch (in spite of the differences between speaker voices) and similar frequencies. All the recorded speech that showed emphatic pronunciation, a sort of marked modulation or other anomalies was rejected.

We sought young speakers with similar pitch. This was because at the beginning of the twentieth century, the informants used for Alcover's "Verbal inflection" were young people, both men and women. In the case of female voices the process of homogenising differences was easier; in contrast, the male voices recorded showed a wide variety of low and high pitches.

We used the tools of the free software WavePad in order to homogenise different pitch voices, change the volume of the recordings, improve its quality, and reduce noise, etc. (see Figure 5).

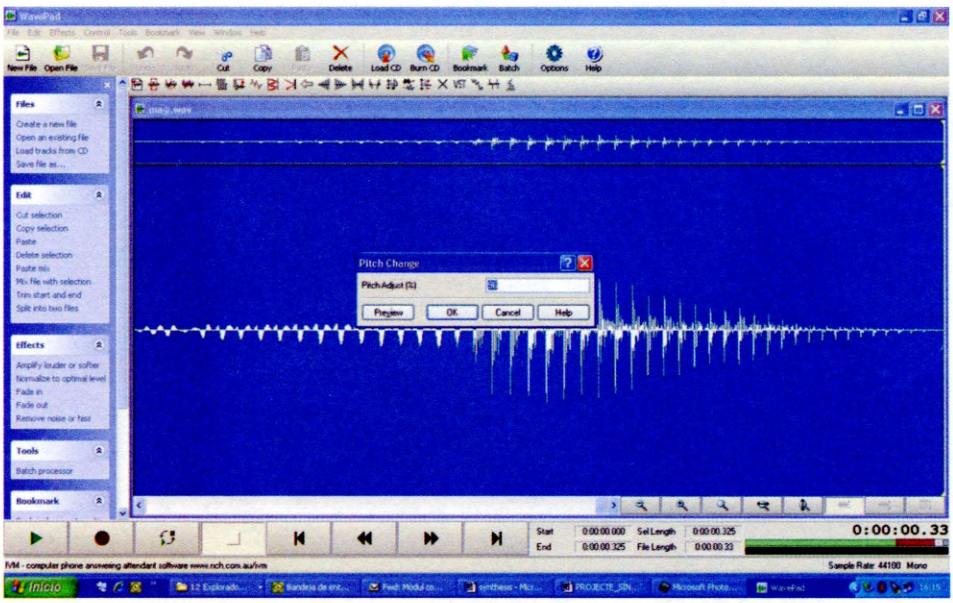


Figure 5

Several different variables may affect the recording of read text, and these need to be taken into account. For instance, the use of different microphones can produce unexpected results related to the pitch.

2.2.2. Developing a speech synthesis system

The speech synthesis system is based on the concatenation of sound units. In our system the syllable was chosen as the main unit for generating synthesised voice. Sounds for which syllables present some problems were used as supplementary units.

We used the syllable as the minimal sound unit, because, after different verifications, we observed that this yielded the best results as regards speech fluidity. Using phonemes as units also requires an intermediate basic unit called a "diphone" (based on the concatenation of each pair of recorded phonemes). In our system, and after applying certain algorisms, use of the syllable was observed to yield —with some exceptions— transitional results that were generally natural enough. Moreover, the syllable is a controlled unit. In Catalan there are about 5,000 different syllables combining consonants and vowels. As regards our corpus, the database of "Verbal inflection" required 3,526 syllables.

Catalan has a lot of monosyllabic words. Concerning stressed syllables most of the words read were real monosyllabic words. The syllable inventory created was stored in a speech database (DBISam database format). In this database each syllabic unit was labelled according to its phonetic representation (Figure 6).

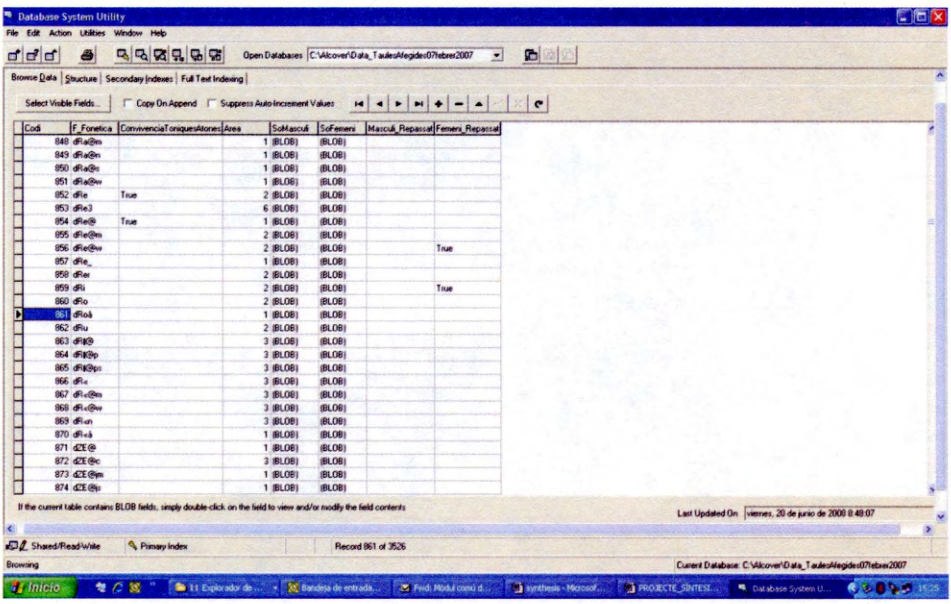


Figure 6

In order to obtain the most natural sounding speech, the corpus contains both stressed and unstressed syllables. The former require a higher pitch and frequency, while the latter require a more muted and constant pitch.

2.2.3. Applying the concatenation process

Firstly, a system was applied to the verbal database so that when we chose a word to be pronounced it automatically divided each word, transcribed phonetically, into syllables according to the specific rules of the Catalan language and taking into account diphthongs, triphthongs and hiatus (Figure 7 shows an example with *acudir* 'to attend').

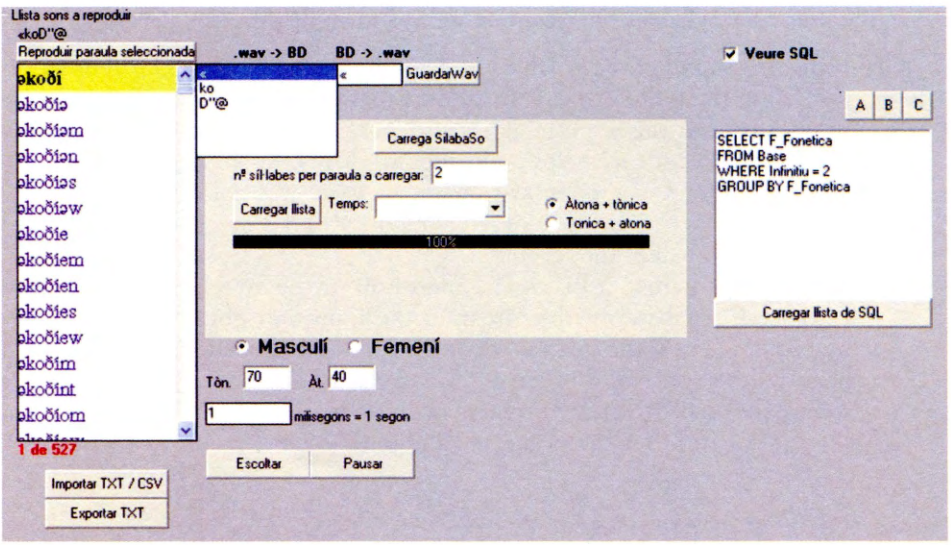


Figure 7

After the syllabic division, the corresponding syllabic sound from the speech database was associated with the phonetic transcription. This sound, comprising a series of numbers corresponding to its wave signal, was stored in a temporary file. Next, the sound of the following syllables (if the word is polysyllabic) was also associated. When a blank space showed the end of the word the program concatenated the different temporary files and the sound of the chosen word was obtained (Figure 8).

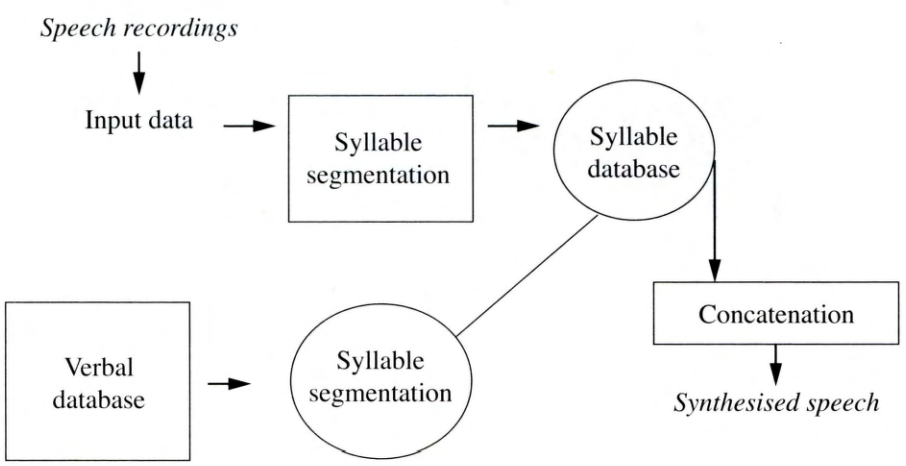


Figure 8

In the process of syllable concatenation we had to apply different algorisms:

c.1) If the first sound of the syllable was a voiceless plosive ([c],[1] [k], [p] or [t]) or if it fell in the middle of the word, the program added a silence of 20 milliseconds (ms) at the beginning, because these sounds are produced in a very short time. The silence helps to distinguish them clearly.

c.2) If the last sound of the last syllable of the word was not a voiceless plosive ([c], [k], [p] or [t]), an effect of reducing the sound progressively was applied to the second half of the syllable.

c.3) If a word contained a hiatus the transition of the vowels was included in the speech database. In this situation the transition produces a more natural sound. Thus, the process adapted to the word [koém] 'we cook' would be as follows: this word originally has two syllables: [ko] + [ém], but a new sound [oé] will be created, which will be associated to [k] and [m], so: [k] + [oé] + [m]. It is thus a sort of "diphone".

2.2.4. Editing the sound and solving problems

We used the WavePad program to edit the different syllables and, in some cases, to modify the sound qualities or to change a part of a syllable (Figure 9). In this case a complete synthesised sound is produced.

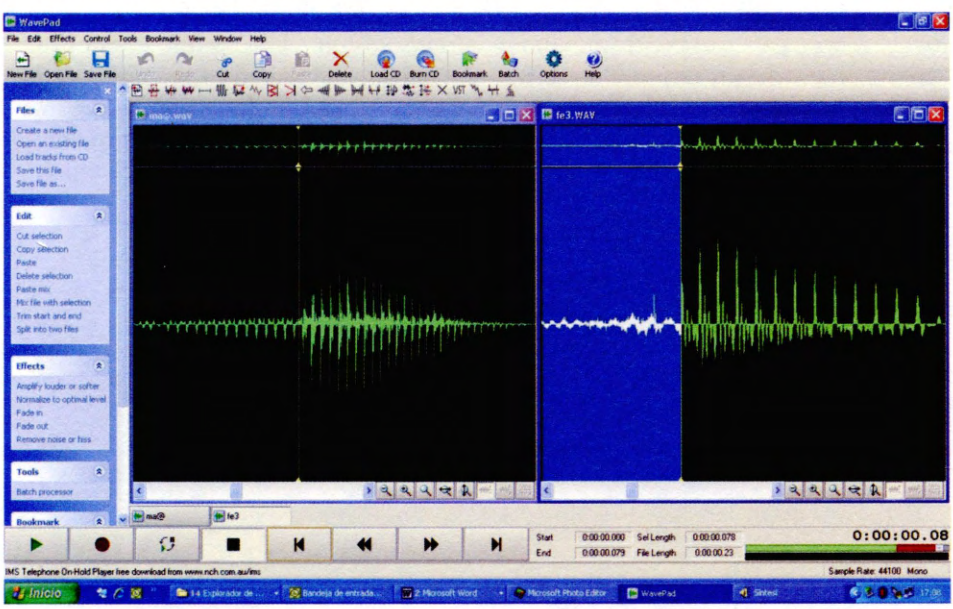


Figure 9

[1] The sound [c] is a velar voiceless plosive palatalised, pronounced in some Majorcan villages. It appears in very predictable contexts.

At times, in the process of concatenation of syllables pronounced by different speakers, we noticed stress and speed differences, even though the pitch was correct. When this happened, we selected the syllables, or a part of them, that presented the best phonetic quality. This process was applied to similar syllables, such as [bíu], [kíu], [díu], [fíu], [gíu], [jíu], [míu], [níu], [ʎíu], [líu], [ɲíu], [ɾíu], [ríu], [síu], [ʃíu], [tíu], [píu]. In this case, [íu] was the most representative part of the syllable and the initial consonant was added to this part.

The most important difficulty in terms of sound concatenation was joining vowels (cf. b.3.c.3) divided into two syllables ([ɾí]+[a] from [kan]+[ta]+[ɾí]+[a] *cantaria* 'I would sing'), because the contact produced a superposition or a sudden change in the pronunciation. To solve this problem we recorded this joining (or transition) independently.

The most conflictive vowels were the stressed [í] and [ú]. The vowel contact with these sounds led to the appearance of "ghost" sounds. Thus, in a form such as *plaïu* [pla-í-u] 'you please', in the coarticulation between [a] and [í], the listener does not hear this form, but rather a non-existent transition consonant: pla[β]íu, pla[ɣ]íu or even pla[ɾ]íu. Here, as happens with other sounds, there is also a perceptual question.

Other problematic sounds to be synthesised were the nasals, [m] and [n] at the beginning of the syllable or [ŋ] and [ɲ] between syllables.

In contrast, the synthesis was easier when the syllables start with the voiceless plosives [c] and [k] or end with [c], [k], [p] and [t]. As regards voiced and voiceless fricatives, in the middle of the word, [ʒ], [s] and [z] there is no problem of synthesis and perception. The same happens when [s] or [ʃ] are at the end of the syllable.

The lateral [ʎ] at the beginning or end of the syllable does not present any problems. However, younger generations tend to lose this sound changing it into [j]. Thus, the chosen speakers had to pronounce this consonant clearly. The sound [l], in the coda position, was sometimes perceived as [w], especially when it was followed by another syllable [val] + [dɾí] + [a] *valdria* 'it would cost'. The sound [w] can be explained by the fact that the lateral alveolar in Catalan has a strong velar secondary articulation.

Rhotic [r] is easier to synthesise than the intervocalic [ɾ].

Other sounds can be synthesised by starting from others. The approximants [ɣ], [β] and [ð] between vowels and between other contexts can be extracted from the second half of the corresponding voice plosives ([g], [b] and [d]).

When the end of a syllable was any strong vowel and the next one was an unstressed [i] or [u] it was necessary to make these weak vowels longer so that they could be clearly perceived. In contrast, when the joining was produced between the strong vowels of the end of a syllable and [j] or [w], the diphthong was distinguished by making the glides shorter and making the pronunciation faster.

A new difficulty has to do with sounds that existed at the beginning of the twentieth century but which are no longer pronounced today, especially among younger generations. This is the case of the sound [á], which can experience changes or a sort of diphthongation in [eá] or [eɛ́]. The diphthongation was pronounced in Son

Servera, a Majorcan village.[2] The transcription of this diphthong in the "Verbal inflection" was [ɐǽ]. Because we have the recordings of the two sounds separately ([ɐ] and [ǽ]) it was possible to create this diphthong joining them.

Finally, because we synthesise only words, prosodic and intonation aspects were avoided. Here the main objective was the intelligibility, quality and naturalness of the utterance of each word pronunciation.

2.3. The new (sound)maps

Application of corpus-based concatenative speech synthesis systems to Alcover's data enabled us to produce a sound atlas, which not only shows the geographical different distribution of Catalan verb morphology but which can also reproduce the phonetic verbal form through a male or female voice (Figure 10).

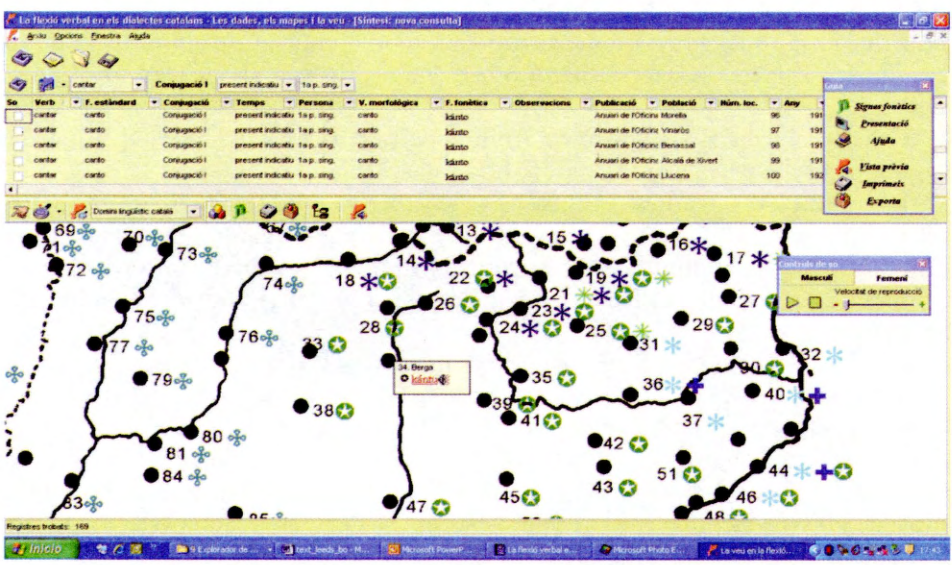


Figure 10

Moreover, the program maps the spatial distribution of the different phonetic and morphologic variants of a verb form, and enables the user to listen to the series of utterances, stopping when desired or repeating a form as often as is required. The speed of the pauses between two utterances can also be changed, making them slower or quicker (Figure 11).

[2] See about this sound, Veny (1983: 100) and Recasens (1991: 91).

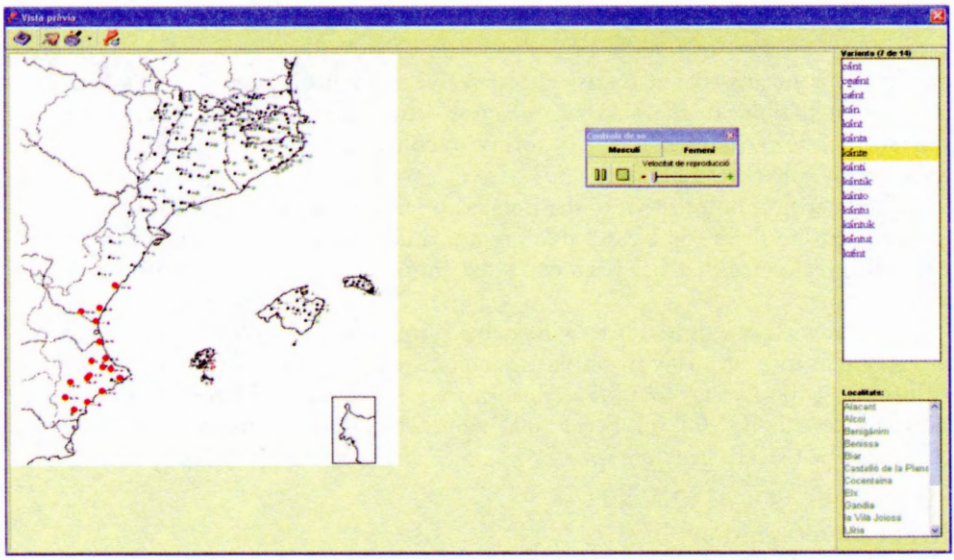


Figure 11

Developing this project has implied a small step in the use of speech synthesis techniques as applied to dialectology. Until now these techniques have been used when voice is the only way of receiving information: for blind people, phone applications (call centres) or in some experimental applications. However, they can also be useful to create voices of speakers for extinct languages, to re-edit dialectal material using new technologies or, as in this case, to recreate the utterances of the informants surveyed by Alcover one hundred years ago. Because at that time it was impossible to record dialectal material, the fact that data are transcribed phonetically helps in using the text-to-speech system.

There are two aspects of the original data which have made the synthesis procedure easier:

a) It is a closed corpus with a limited number of verbal forms — that is to say, it comprises isolated word sequences (in which emotions and prosodic elements have not been considered), and for that reason the speech database was limited. Only in a few cases is there a conjunction of two words in the periphrastic verb forms (*he cantat* 'I have sung' or *vaig cantar* 'I sang'). These utterances were more difficult to produce.

b) Verbal forms are already phonetically transcribed. Thus, we leave out what is called the "front-end" of the text-to-speech system: the input plain text that is transformed in phonetic transcription and the subsequent division into prosodic units.

There is an important element that has sometimes been associated with the intelligibility of an utterance. The rich Catalan verbal variety that was gathered by Alcover at the beginning of the twentieth century is now progressively reduced under the pressure of standard morphology. As regards the subjective evaluation, we

checked that extinct verbal forms or those that are disappearing are more difficult to perceive by speakers of standard Catalan, even though the speech output is of good quality. This means that it is easier to perceive a predictable or known word, as it is more intelligible. In the "Verbal inflection" the dialectal variation presented by a verb such as *obrir* 'to open' is of 919 forms; in contrast, the standard paradigm of this verb, or any other, has only 50.

From a critical point of view, the isolated utterance of a single word in a point on the map can result in some cases that are not much natural; however, the same happens with real voice that has been extracted from a given context, in order to create sound maps.

The knowledge acquired here will enable future research to apply this speech synthesis system to a corpus with similar characteristics. For example, the phonetic transcriptions of the *Diccionari Català-Valencià-Balear*,[3] also by Alcover in collaboration with Francesc de B. Moll (cf. Perea 2004), and which likewise forms a closed corpus, could also be used to generate speech.

3. Speech recognition[4]

Broadly speaking, speech recognition (Zue, Cole & Ward 1997) aims at permitting oral communication between human beings and computers. The challenge is to harmonize information coming from different fields of knowledge (acoustics, phonetics, phonology, lexis, syntax, semantics, and pragmatics) and to obtain, in spite of possible ambiguities and errors, an acceptable interpretation of an acoustic message.

Speech recognition can be applied at different language levels, presented here in increasing order of difficulty:

1. Isolated words.
2. Connected words.
3. Continuous speech.
4. Spontaneous speech.

At each of these levels the speech acoustic signal has to be transformed into the sequence of words from which the written text is configured. In general, the procedure involves the conversion of sound waves into variations in voltage through a microphone device, followed by the sampling of this voltage, so as to obtain, for example, 8,000 samples per second (i.e., each sample is encoded in eight bits; while greater fidelity requires a greater sampling frequency; in our case, 44,100 KHz encoded in 16 bits with which the answers of the informants were recorded). Having obtained the sample by means of the "digital processing of the acoustic signal", the data are analysed in order to extract the prominent information that enables words to be identified. Thus, the intonation of the sentence does not contribute to their identification, but rather it is the frequency of the acoustic signal as a sound is being

[3] http://dcvb.iecat.net/default.asp.

[4] This research is sponsored by the Spanish Ministerio de Educación y Ciencia and the FEDER (research project HUM2007 65531/FILO: "A dialectal oral corpus exploitation: analysis of the linguistic variation and development of computerised applications to automatic transcription" (ECOD))".

emitted that is the prominent information, since it indicates its phonic characteristics (pitch, means of articulation, etc.).

In theory, once the parameters characterizing the sign acoustics have been established, the search can be initiated. But prior to commencement, a dictionary of words has to be generated. The search is then conducted in this dictionary so as to identify the word that most closely resembles the word we wish to identify. However, problems arise when working with continuous speech (without any artificial pauses in the majority of words) and, in particular, when dealing with spontaneous speech, because part of the search process also involves determining the limits of the word, i.e., identifying where words begin and finish. Two types of external knowledge facilitate this search: the *acoustic* and the *language models* (Fosler-Lussier, Byrne & Jurafsky 2005). The former establish the distribution of the acoustic parameters of each phoneme (thus, /b/, for example, is a voiceless bilabial plosive), while language models establish the usual distribution of the words of a specific language (i.e., the sonorous sequences that are most likely to occur). Thus, based on the characteristics of the acoustic signal and the expectations set up by the acoustic and language models respectively, a hypothesis can be formulated regarding what was said.

Automatic speech recognition has many applications in daily life, including machine translation, speech recognition in cars, computers (dictaphones), GPS, *Speech-to-Text* programs (SMS text transcriptions of speech), robotics, etc. Yet, as is the case with voice synthesis, automatic speech recognition systems have not as yet been applied to dialectology.

The aim of one part of the project designed by the *Universitat de Barcelona*, "A dialectal oral corpus exploitation: analysis of the linguistic variation and development of computerised applications to automatic transcription" ((ECOD) [HUM2007 65531/FILO]) is to develop data processing tools that facilitate the automatic transcription —phonetic and subsequently orthographic— of interviews collected in dialectal surveys carried out in the county capitals (or equivalent centres) of the Catalan linguistic domain. We have gathered a total of 258 oral texts ranging in length from between five and ten minutes and recorded in 86 different places.

We are currently seeking to design a tool that transcribes the oral texts of the informants phonetically, regardless of the variety of their Catalan dialect. Subsequently, we wish to develop an application that can convert the phonetic version to the corresponding orthographic transcript by applying rules developed from the generalizations of each dialectal system.

3.1. From "WavSons" to "WavFonemes"

Our speech recognition system has been applied initially to isolated words from eastern Catalonia. Here, as our point of departure comprises oral texts and a phonetic database, we need to segment out data in isolated words. To do this, we developed a sound processing program, "WavSons 1.0", which enables us to segment, process and store a range of syllabic units in files. As with voice synthesis techniques, the unit of analysis is the syllable (cf. Adda-Decker, de Mareüil, Adda & Lamel 2005), although diphones have also to be taken into consideration. Note, however, that we do not work with isolated sounds or complete words.

"WavSons 1.0" parameterises and stores syllables according to their defining acoustic and phonetic properties. The program proposes an initial segmentation of the sequence under analysis, but subsequent modifications have made certain adjustments to this. The empirical base drawn upon is an oral corpus known as the *Corpus Oral Dialectal* (COD), which provides specific syllable samples.

Samples were collected from informants resident in a number of towns in western Catalonia, including Granollers, Terrassa, Mataró, Santa Coloma de Farners, Vic and Berga. Then, the ability of the program to recognise syllables was tested by analysing these samples. The program was found to distinguish plosives and vowels more readily than liquids, nasals and approximants followed by a vowel, and consonant sounds in the same block. As for the elements of analysis, the collocation of syllables comprising a voiceless plosive plus a vowel was given particular emphasis. Next, the sounds of the segmented syllable were associated with their orthographic and phonetic transcription. In this way, we were able to create a syllabic database. Last year, we replaced "WavSons 1.0" with an improved version, "WavFonemes". This program has been used to design a Catalan sound database and subsequently to normalise it using voice patterns. Our voice pattern database has been created by selecting significant groups of sounds obtained from the speech of our informants. These groups correspond to syllables, though they might also be considered to be diphonemes (understood as entities formed from two sounds).

When incorporating these voice patterns to the database we are concerned solely with the characteristics of the sound spectrum. Thus, first it is necessary to transform the voice signal recorded directly with a microphone (time axis) to sound frequencies (frequency axis), since the identifying and invariable information obtained from the vowels and consonants, independent of the informant, is encoded on the frequency axis. A mechanism must be applied, therefore, that can transform the information in the samples collected on the time axis into frequencies. To carry out this transformation, we apply the "Fourier Transform" operation, which works on the principle that each sound signal, even if it is complex, can be split into the sum of its simple periodic functions (sinusoidal and cosinusoidal) of different frequency. In this way we are able to obtain values that determine the importance of the frequencies, and the most significant are those that compose and shape the voice patterns that we wish to create.

3.2. Quality voice patterns

Obtaining quality patterns requires going through two prior stages:

a) Training patterns (pattern design): establishing consistent representations of minimally reliable sound patterns.
b) Comparison of patterns (obtaining speech recognition from the pattern designs): a direct comparison is made between the unknown signal (the one to be recognized) and all the possible patterns acquired in the training stage so as to determine which pattern gives the best fit.

The main difficulty encountered in adopting this method is that of creating a complete and correct speech pattern database. This was particularly true in our case where we worked with isolated words coming from a closed database. An accurate database is clearly not easily obtainable, because of the dimensions involved and the variations in

the signal sound, and given that the spectrum information it contains does not have a recognizable phonetic meaning. This variation in sound can be attributed to:

1. The variation presented by a single informant as she seeks to maintain a constant and consistent pronunciation in her word production.
2. The variation between informants due to length differences in the processing of vowels and consonants, and to differences in accents associated with dialects, etc.
3. The variation in microphones when recording speech in terms of volume, intensity, etc.
4. The variation in the recording settings (general background or specific noises).

3.3. The functions of the "WavFonemes" program

The "WavFonemes" program allows us to:

Segment parts of a wave file and assign to each part its corresponding phonetic syllable. Once segmented, the sequence can be stored in the database together with the file name referring to that segment and the phonetic transcription assigned to it (Figure 12).

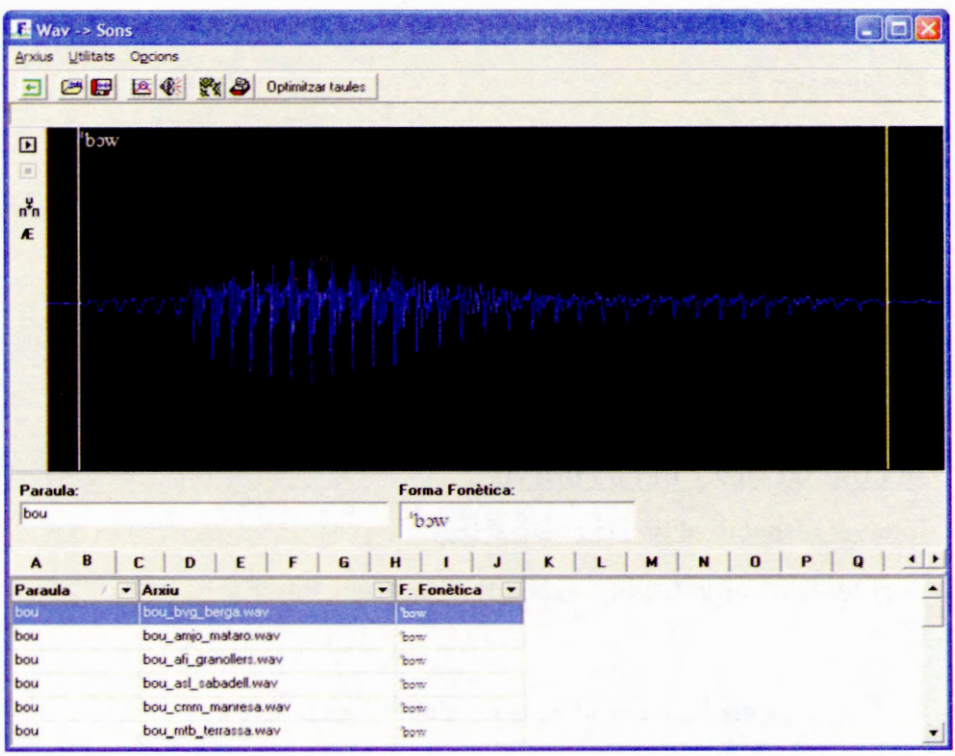


Figure 12

The program has two main tools:

a) Normalising (training stage) the phonetic syllables stored in the database. The aim here is to build a valid phonetic pattern for speech recognition, in which the most representative frequencies of each syllable have a particular importance, independent of the informant. These frequencies are those with the highest values. Various parameters are taken into consideration in obtaining this optimum normalization: first, maximum and minimum frequencies (in general the frequency interval corresponding to the threshold of the human ear, which oscillates between 20 and 10,000 Hz); and second, the average frequency among the higher values, fixed in this case at between 70 and 80%. Note fixing the value at 100% would be incorrect as this would include all sample values, and clearly the values that appear infrequently do not contribute to the construction of a valid pattern (Figure 13). What we seek to do is to generalize and avoid excessive diversification.

b) Recognising (comparison stage) phonetic syllables after they have been normalised. The aim here is to use the wave file to bring up on the screen the di-

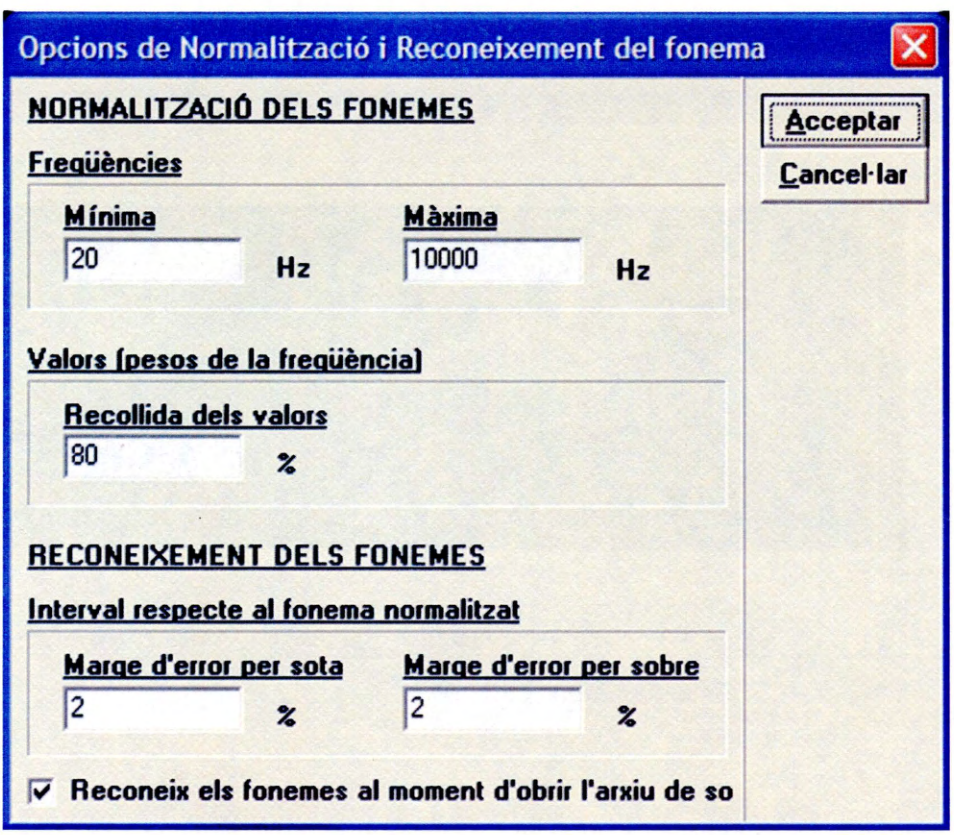


Figure 13

visions in the phonetic syllables and so the corresponding syllable can be identified from the patterns stored in the database.

In the first stage of the voice recognition process, a graph representing the voice signal on the frequency axis is taken as a reference. The samples of the signal are then divided into superimposed sections (windows) of the same size. On each set of samples within each window the "Fourier Transform" formula is applied to detect and group similar frequencies, before each syllable is divided. Once this division is complete, the most typical values are collected, according to the parameters of maximum and minimum frequencies, the average value (the frequency weighting), and the margins of error with respect to the normalized phoneme.

Only the values that coincide are collected. In figure 13, the values have a margin of error of 2%. If these parameters were to be expanded, we would obtain a broader list of candidates, but they would not be as exact. To obtain a progressively more precise technique, the mean values would have to be manipulated. Finally, the frequency values are compared with the previously stored database values and a list of possible candidates drawn up, with the phonetic syllables organised from greatest to smallest degree of accuracy.

3.4. Prospects

a) In the normalisation stage, more statistical parameters might be incorporated in order to condition and perfect the phonetic pattern (median of frequencies, standard deviations of the values that have to be collected, etc.).

b) Other methods of speech recognition might be explored so as to increase the degree of accuracy in the patterns. One such model, capable of providing satisfactory results at the practical level, is the Hidden Markov Model (HMM) (Knill & Young 1997). It should be borne in mind, however, that the model of the pattern can differ depending on how the patterns are obtained and on the particular speech sequence that we wish to recognise. In fact, there is no escaping the marked influence of the informant who recorded the particular speech sequence.

c) The database might be restructured in order to make the processes of normalisation and recognition more agile.

4. Conclusion

Dialectology must take advantage of new computer technologies that can facilitate the processing of dialectical data. To date, the procedures developed in the text-to-speech (TTS) systems and in speech recognition programs have only been used in conjunction with standard language varieties, and so the challenge is now to apply them to the study of linguistic variation.

References

Adda-Decker, M., de Mareüil, P. B., Adda, G. & L. Lamel, 2005, «Investigating syllabic structures and their variation in spontaneous French», *Speech Communication* 46, 2, 119-139. http://dx.doi.org/10.1016/j.specom.2005.03.006.

Fosler-Lussier, E., Byrne, W., & D. Jurafsky (eds.), 2005, Pronunciation Modeling and Lexicon Adaptation. Special Issue. *Speech Communication*, 46, 2.

Jurafsky, D. & J. Martin, 2000, *Speech and Language Processing: an Introduction to Natural Language Processing*, Computational Linguistics and Speech Recognition, Upper Saddle River, New Jersey, Prentice Hall.

Keller, E., 1994, *Fundamentals of Speech Synthesis and Speech Recognition. Basic Concepts, State of the Art and Future Challenges*. Jon Wiley & Sons, Chichester.

Knill, K. & S. Young, 1997, «Hidden Markov Models in Speech and Language Processing», in S. Young & G. Bloothooft (eds.), *Corpus-Based Methods in Language and Speech Processing*. Kluwer Academic Publishers, Dordrecht (Text, Speech and Language Technology, 2), 27-68.

Lewis, E. & M. Tatham, 1999, «Word and syllable concatenation in text-to speech synthesis», in *Proceedings of the European Conference on Speech Communication and Technology* (http://www.cs.bris.ac.uk/Publications/Pa-pers/1000377.pdf).

Perea, M.-P., 2004, «New Techniques and Old Corpora: *La flexió verbal en els dialectes catalans* (Alcover-Moll, 1929-1932). Systematisation and Mapping of a Morphological Corpus», *Dialectologia et Geolinguistica*, 12, 25-45.

—, 2005, *Dades dialectals. Antoni M. Alcover*, Conselleria d'Educació i Cultura. Govern de les Illes Balears, Palma de Mallorca (CD-ROM edition).

— (to be printed), «Retrieving the sound: applying speech synthesis to dialectal data», paper read at METHODS XIII (Thirteenth International conference of Methods in Dialectology), Leeds, July, 2008.

Pols, L., 2001, «Acquiring and implementing phonetic knowledge», in P. Dalsgaard, B. Lindberg & H. Nemmer (eds.), *Eurospeech 2001. Proceedings of the 7th European Conference on Speech Communication and Technology*. September 3-7, 2001, Aalborg, Denmark. Vol 1. pp. K3-K6. In IFA Proceedings (Institute of Phonetic Sciences, University of Amsterdam) 24 (2001): 39-46.

Recasens, D., 1991, *Fonètica descriptiva del català*, Institut d'Estudis Catalans, Barcelona.

Veny, J., 1983, *Els parlars catalans*, Moll, Palma de Mallorca.

Zue, V., Cole, R. & W. Ward, 1997, «Speech Recognition», in R. A. Cole, J. Mariani, H. Uszkoreit, A. Zaenen, & V. Zue (eds.) *Survey of the State of the Art in Human Language Technology*, Cambridge U. P., Cambridge, 4-10.

RELEVANCIA DEL ANÁLISIS LINGÜÍSTICO EN EL TRATAMIENTO CUANTITATIVO DE LA VARIACIÓN DIALECTAL

Esteve Clua
IULA-Universitat Pompeu Fabra

Abstract

In previous studies we have argued in favour of the need for a phonological analysis of dialectal data prior to quantitatively determining linguistic distance. Our idea, based on generative linguistics, has to do with the fact that phonetic coincidences may frequently hide relevant phonological divergences. Thus, we consider that quantitative analyses carried out solely from phonetic data cannot reflect existing differences or similarities among dialectal varieties adequately. So far, our reasoning has always been backed by small-scale laboratory tests performed on small sets of data. In the present study, however, our reasoning in favour of linguistic analysis is corroborated by a tested and global treatment of data from the Contemporary Catalan Oral Dialect Corpus. To this end we have carried out two quantitative studies: one on the basis of phonetic data, and another one using a phonological analysis of the data.

Key words: *dialectometry, quantitative analysis, linguistic distance, phonological analysis, phonetic data.*

1. Introducción[1]

Una de las características distintivas de los trabajos dialectométricos[2] que se han desarrollado alrededor del *Corpus Oral Dialectal* (COD)[3] del catalán contemporáneo, es el hecho de basar los tratamientos cuantitativos para describir la variación lingüística del catalán en un análisis lingüístico previo. Como veremos, en estudios anteriores hemos justificado esta necesidad a partir de pequeños conjuntos de datos extraídos del corpus.

---

[1] Este trabajo forma parte del proyecto de investigación HUM2007-65531/FILO (ECOD "Explotación de un corpus oral dialectal: Análisis de la variación lingüística y desarrollo de aplicaciones informáticas para la trascripción automatizada, 2.ª fase"), financiado por el Ministerio de Ciencia e Innovación y el FEDER. Más información sobre el proyecto se encuentra disponible en http://www.ub.edu/lincat.

[2] *Vid.*, por ejemplo, Viaplana (1999), Clua (1999a y b, 2007), Clua *et alii* (2008 y 2009).

[3] *Vid.* la versión en CD del COD: Viaplana *et alii* (2007).

La finalidad de este trabajo es presentar por primera vez el contraste entre dos análisis cuantitativos globales del COD: uno realizado a partir de los datos fonéticos y otro a partir de los datos analizados fonológicamente. Nuestro objetivo es discernir si las posibles diferencias entre ambos tratamientos son pertinentes y justifican, por tanto, el trabajo de análisis lingüístico de los datos antes de llevar a cabo el tratamiento cuantitativo.

Para ello, en primer lugar (§2), argumentamos la necesidad de llevar a cabo un análisis lingüístico previo de los datos para poder captar todas las similitudes o diferencias que pueden existir entre determinadas variedades lingüísticas en relación a un determinado rasgo. En segundo lugar (§3), describimos el análisis fonológico que aplicamos a los datos del COD. A continuación (§4), a partir de una pequeña muestra de datos, como si se tratase de un pequeño ensayo de laboratorio, presentamos las implicaciones de tal planteamiento para determinar cuantitativamente la distancia lingüística entre variedades. En el apartado siguiente (§5), describimos y analizamos los resultados en contraste de los dos tratamientos cuantitativos aplicados a los datos del COD. Finalmente en (§6) presentamos las conclusiones de nuestro estudio.

2. El porqué del análisis lingüístico previo

Si alguna cosa caracteriza globalmente los diferentes métodos cuantitativos de descripción y clasificación dialectal que han proliferado durante las últimas cuatro décadas en Europa, es el hecho de compartir la adopción del concepto de distancia lingüística como pieza básica para la descripción de la variación lingüística. El concepto de distancia, adoptado del ámbito científico del análisis de datos, se asocia generalmente en nuestro campo a la cuantificación de las similitudes o diferencias que existen entre variedades lingüísticas en relación a un conjunto de datos. Se trata de un punto de vista sobre la variación lingüística que se aparta sustancialmente de la dialectología tradicional, pero que ya se podía vislumbrar en el siglo XIX en las palabras de Durand (1889):

> Et maintenant, qu'est-ce qui constitue le degré de ressemblance qui rapproche deux langues entre elles, et le degré de dissemblance qui les éloigne l'une de l'autre? La ressemblance se mesure à la proportion des caractères communs, la dissemblance à la proportion des caractères particuliers.

Pero ¿cómo determinamos las diferencias o similitudes entre variedades lingüísticas a partir de las cuales establecer el tratamiento cuantitativo? ¿Reflejan las representaciones fonéticas de los ítems de un determinado atlas o corpus todas las diferencias existentes entre las variedades lingüísticas? Vamos a intentar responder a estas preguntas a partir de los ejemplos siguientes en los que contrastamos las formas del numeral *dos* en dos variedades lingüísticas.

(1)	*Variedad 1*	*Variedad 2*
a.	dos ['dos]	dos ['dos]
b.	dos cafès ['doska'fɛs]	dos cafès ['doska'fɛs]
c.	dos bonsais ['dozβon'sajs]	dos bonsais ['dozβon'sajs]
d.	dos animals ['dozani'mals]	dos animals ['dosani'mals]

Si contrastamos la realizaciones fonéticas del numeral aislado (1a) en las dos variedades, vemos que no presentan ninguna diferencia. Tampoco hay diferencias cuando aparece delante de otra palabra empezada por consonante sorda (1b). Cuando precede una palabra empezada por consonante sonora (1c) tampoco encontramos diferencias entre ambas variedades, pero en este caso podemos observar que el sonido sibilante del final de la palabra presenta una realización sonora [z], que contrasta con las realizaciones sordas de este segmento en los contextos anteriores. Finalmente en (1d) si que podemos observar una realización diferente del numeral en las dos variedades contrastadas: mientras que la variedad 1 presenta una realización sonora de la sibilante final, en la variedad 2 encontramos una realización sorda.

Teniendo en cuenta esto podríamos concluir, en principio, que estas dos variedades no presentan ninguna diferencia en la representación fonética del numeral *dos* ['dos], porqué los tres segmentos que integran esta palabra coinciden totalmente; pero la coincidencia no puede ser total si tenemos en cuenta el comportamiento del último segmento, el sibilante [s], que se sonoriza siempre delante de un segmento sonoro [± vocálico] en la variedad 2, mientras que solo lo hace delante de un segmento no vocálico en la variedad 1. Es decir, las dos variedades coinciden totalmente en los segmentos fonéticos que constituyen el numeral *dos*, pero los procesos fonológicos que afectan dichos segmentos son diferentes. Si entendemos que estos procesos fonológicos son básicos para comprender la estructura sonora de las variedades lingüísticas, tendremos que colegir asimismo que no podemos pasarlos por alto al intentar captar las diferencias existentes entre ellas.

Para poder tener en cuenta todos estas diferencias, aplicamos a los datos fonéticos de nuestro corpus un análisis lingüístico basado en el modelo de la fonología generativa clásica, porque permite discriminar las diferencias superficiales o predecibles, que expresan las regularidades de las lenguas (y de las variedades que las componen), de las diferencias subyacentes o impredecibles, que afectan a la estructura léxica o gramatical de las palabras (*vid.* Lloret & Viaplana 1998). Desde nuestro punto de vista, la distinción entre estos dos niveles de análisis es fundamental para determinar la distancia lingüística entre variedades (*vid.* Clua 1999a, b y Viaplana 1999).

3. El análisis aplicado a los datos del COD

Para el análisis de fenómenos fonológicos específicos, hemos seguido mayoritariamente la orientación propia de la fonología generativa derivacional. En cuanto a los aspectos morfológicos, seguimos básicamente el enfoque morfémico clásico (*Item and Arrangement*), aunque también en esta área hemos empezado a avanzar en el campo de la morfología paradigmática (*Word and Paradigm*) para poder explicar los efectos analógicos y de contraste derivados de las relaciones intraparadigmáticas e interparadigmáticas que se establecen entre palabras morfológicamente relacionadas.[4]

[4] Para un análisis derivacional y morfémico de los datos del COD, *vid.*, entre otros, los trabajos citados anteriormente; para un enfoque simultáneo y paradigmático, *vid.*, por ejemplo, Bonet & Lloret (2005) y Lloret (2004).

Podemos ver, a continuación, un caso de diferencias fonológicas (subyacentes) y otro de diferencias fonéticas (superficiales). El catalán presenta distintas terminaciones para la 1.ª persona del singular del presente de subjuntivo: unas variedades presentan -[e] (en la mayor parte del área valenciana), otras -[i] (variedades del catalán oriental) y otras -[a] (en algunas variedades del catalán nordoccidental). Ninguna de estas diferencias, sin embargo, puede ser atribuida a un fenómeno sistemático de la fonología del catalán; es decir, en estas variedades no existe ningún proceso regular por el cual -/e/ se convierta en -[i] o en -[a], o viceversa. Estas diferencias, por tanto, han de ser atribuidas directamente a la estructura morfológica de las palabras. Podemos verlo en los ejemplos de (2), correspondientes al verbo *cantar*.

(2) Diferencias Fonológicas
1.ª persona del singular del Presente de Subjuntivo de *cantar*

a. Variedad 1: [ˈkante] /e/
b. Variedad 2: [ˈkanti] /i/
c. Variedad 3: [ˈkanta] /a/

Otro ejemplo de variación que afecta a las terminaciones verbales lo encontramos en las formas del gerundio. Muchas variedades, como es el caso del catalán de Barcelona, presentan una -[n] final en esta forma verbal; en cambio muchas de las variedades valencianas acaban el gerundio con -[nt]. Si realizamos una cuantificación de las diferencias a partir de los datos fonéticos, estas diferencias son iguales que las de (2). Desde la óptica generativa, sin embargo, la alternancia [n] ~ [nt] que se observa en la variedad 1 (3a) puede ser atribuida a una única forma fonológica /nt/, que se realiza como [n] en posición final de palabra y como [nt] cuando le sigue un clítico, ya que en estas variedades opera un proceso fonológico de simplificación del grupo consonántico [nt] en final de palabra, que no se produce cuando la forma verbal va seguida de un clítico pronominal; en cambio en las variedades de (3b) no se produce ningún tipo de alternancia, ya que estas variedades no conocen el proceso de simplificación del grupo [nt].

(3) Diferencias Fonéticas

a. Variedad 1:

cantant	'cantando'	[kanˈtan]	/nt/
cantant-ho	'cantándolo'	[kanˈtanto]	

b. Variedad 2:

cantant		[kanˈtant]	/nt/
cantant-ho		[kanˈtanto]	

Desde nuestra perspectiva las diferencias observadas en (2) son fonológicas, solo predecibles a partir de la estructura morfológica; mientras que en (3) las diferencias son meramente fonéticas y se pueden predecir por un proceso fonológico. Creemos que se trata de una distinción pertinente que debe tenerse en cuenta en la cuantificación de la distancia lingüística entre variedades, ya que de lo contrario el resultado del análisis cuantitativo puede apartarse considerablemente de la realidad. Así pues,

el análisis lingüístico nos permite, por un lado, captar similitudes o diferencias entre variedades que a simple vista fonética serían imperceptibles, y por el otro, nos permite distinguir entre diferencias estructurales (las que aquí denominamos fonológicas) y diferencias predecibles (fonéticas) a partir de los procesos fonológicos sistemáticos que caracterizan las variedades lingüísticas.

4. Un ensayo de laboratorio

A partir del análisis de los clíticos pronominales, una de las áreas en que el catalán presenta más variación dialectal, hemos justificado en trabajos anteriores (*vid.* Clua y Lloret 2006 y 2007) la pertinencia del análisis lingüístico para realizar un tratamiento cuantitativo adecuado. Se trataba de lo que podríamos catalogar de pequeño ensayo de laboratorio a partir de la variación que presentan estos clíticos en tres variedades lingüísticas del catalán, entre las que medíamos y representábamos la distancia lingüística, primero, a partir de los datos fonéticos y, a continuación, a partir de los datos analizados fonológicamente.

Volveremos aquí a presentar nuestra argumentación, en este caso ciñéndonos al pronombre de 1.ª persona del singular *me* en tres variedades occidentales del catalán. El pronombre aparece en negrita en los ejemplos de (4).

(4) Clíticos pronominales de 1.ª persona del singular

	Variedad 1	*Variedad 2*	*Variedad 3*
a. *em pensaré* "me pensaré"	[**em**pensa'ɾe]	[**me**pensa'ɾe]	[**me**pensa'ɾe]
b. *m'esperaré* "me esperaré"	[**me**spera'ɾe]	[**me**spera'ɾe]	[**me**spera'ɾe]
c. *vol esperart-me* "quiere esperar-me"	[espe'ɾar**me**]	[espe'ɾar**me**]	[espe'ɾar**me**]
d. *esperà'm* "espérame!"	[es'pera**m**]	[es'pera**m**]	[es'pera**me**]

Los ejemplos de (4) muestran que en estas variedades el pronombre *me* se presenta con una forma no silábica [m] y con dos formas silábicas distintas, [em] y [me]. En la variedad 1, [em] aparece delante de un verbo que empieza en consonante (4a), mientras que [me] aparece detrás de un verbo que acaba en consonante (4c). En la variedad 2, [me] aparece en los dos contextos anteriores (4a,c). En la variedad 3, [me] aparece en los contextos anteriores (4a,c) y también cuando el verbo acaba en vocal (4d). Desde el punto de vista tradicional la distancia lingüística entre estas variedades es similar: las tres coinciden en las formas de (4b,c), y mientras las variedades 1 y 3 difieren en dos casos: (4a) y (4d), la variedad 2 se distingue de la variedad 1 en una forma, (4a), y de la variedad 3 en otra forma, (4d).

Si basamos nuestro estudio de la variación dialectal únicamente en los datos fonéticos, la distancia lingüística entre las variedades 1 y 2 tiene el valor 1, ya que solo difieren en una forma: [**em**pensa'ɾe] vs. [**me**pensa'ɾe]. Las variedades 1 y 3, en cambio, presentan una distancia lingüística de valor 2 porque difieren en dos formas:

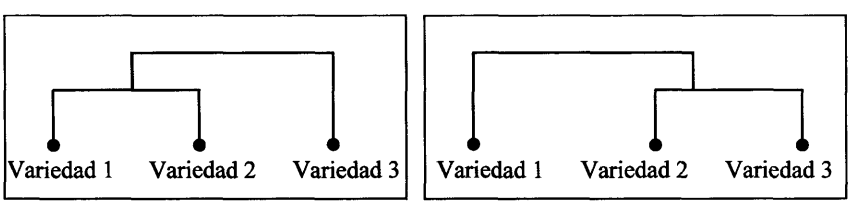
[**em**pensa'ɾe] vs. [**me**pensa'ɾe] i [es'pera**m**] vs. [es'pera**me**]. Por su parte, las variedades 2 y 3 tienen una distancia lingüística también de valor 1, ya que únicamente difieren en la forma: [es'pera**m**] vs. [es'pera**me**]. A continuación presentamos la matriz de distancias que corresponde a estas diferencias fonéticas.

(5) Matriz de distancias a partir de los datos fonéticos

	Variedad 1	Variedad 2	Variedad 3
Variedad 1	0		
Variedad 2	1	0	
Variedad 3	2	1	0

La representación dendrográfica de esta matriz de distancias presenta dos opciones posibles, dependiendo de si agrupamos primero las variedades 1 y 2, o 2 y 3, ya que ambas agrupaciones presentan una distancia mínima de valor 1. En todo caso, la distancia entre las tres variedades es mínima y podríamos decir que hay una distancia muy similar entre ellas.

(6) Representación dendrográfica de la distancia lingüística a partir de los datos fonéticos.



Estos casos de variación puede reanalizarse si se tienen en consideración aspectos relacionados con la estructura silábica y distinguimos entre formas subyacentes y formas predecibles. Así, en las variedades 1 y 2 las distintas formas del pronombre pueden explicarse a través de una única forma subyacente no silábica /m/. En este caso, la vocal [e] (que es la vocal no marcada del sistema vocálico átono de estas variedades: [a], [e], [i], [o], [u]) tiene carácter epentético; es decir, se añade para permitir la silabación adecuada de la secuencia formada por el pronombre y el verbo. La diferencia entre ambas variedades radica en la ubicación de la epéntesis. En la variedad 1, la vocal epentética aparece siempre en la periferia del grupo formado por el verbo y el pronombre: [**em**pensa'ɾe] (7a) vs. [espe'ɾar**me**] (7c); en cambio, en la variedad 2 la epéntesis siempre aparece en un lugar fijo, a la derecha del pronombre: [**me**pensa'ɾe] (7a) y [espe'ɾar**me**] (7c). Desde esta perspectiva, la variedad 3 es completamente distinta. El ejemplo determinante es (3d), [es'pera**me**], en donde la presencia de la vocal final del pronombre no puede justificarse por razones de silabación, pues esta variedad podría presentar perfectamente una forma [es'pera**m**] sin necesidad de recurrir a epéntesis alguna. En este caso, es más coherente postular una forma subyacente distinta, /me/, con una vocal final que se elide en determinados contactos vocálicos

[**me**spera'ɾe], en (7b)), al igual que ocurre en otros casos (cf. *entre amics: entr[a]mics*; *no és tan gran: n[o]s tan gran*).

(7)

	Variedad 1	Variedad 2	Variedad 3
a.	[**em**pensa'ɾe]	[**me**pensa'ɾe]	[**me**pensa'ɾe]
b.	[**me**spera'ɾe]	[**me**spera'ɾe]	[**me**spera'ɾe]
c.	[espe'ɾa**rme**]	[espe'ɾa**rme**]	[espe'ɾa**rme**]
d.	[es'pera**m**]	[es'pera**m**]	[es'pera**me**]
f.	/m/	/m/	/me/

Podríamos decir, en otras palabras, que la variedad 3 ha mantenido la forma originaria del clítico, *te* (coincidente con la forma del latín), aunque que elide la vocal por procesos fonológicos sistemáticos en contacto con determinadas vocales. En cambio, en la variedades 1 y 2 se ha producido una restructuración del sistema pronominal: la mayoría de clíticos pronominales de estas variedades suelen estar constituidos subyacentemente por una consonante (/m/, /t/, /s/...) a la cual añaden una vocal epentética cuando lo exigen las reglas de silabación. Teniendo en cuenta este análisis, los aspectos que presentan variación entre estas tres variedades son los de (8).

(8) Diferencias entre variedades

a. Variedad 1: /t/ y epéntesis en la periferia
b. Variedad 2: /t/ y epéntesis a la derecha del pronombre
c. Variedad 3: /te/ y elisión de vocales en contacto

Ahora la perspectiva de la distancia lingüística entre estas variedades es bastante diferente, mientras que las dos primeras coinciden plenamente en cuanto a la forma subyacente y solo difieren en el tipo de epéntesis utilizado, la tercera variedad se aparta de ellas considerablemente, tanto en el ámbito de las formas subyacentes como en el de los procesos fonológicos.

A continuación presentamos las matrices de distancias obtenidas a partir del análisis fonológico de los clíticos pronominales. En (9a), donde se presentan las diferencias morfológicas o subyacentes, podemos ver cómo mientras las variedades 1 y 2 presentan una distancia de valor cero, ya que ambas variedades coinciden en la forma /m/ del pronombre, la variedad 3 tiene un valor 4 respecto de las otras dos variedades.

(9) Matrices de distancias a partir de los datos analizados fonológicamente

a. Diferencias morfológicas (Variedades 1 y 2 /t/, Variedad 3 /te/)

	Variedad 1	Variedad 2	Variedad 3
Variedad 1	0		
Variedad 2	0	0	
Variedad 3	4	4	0

b. Diferencias en los procesos fonológicos

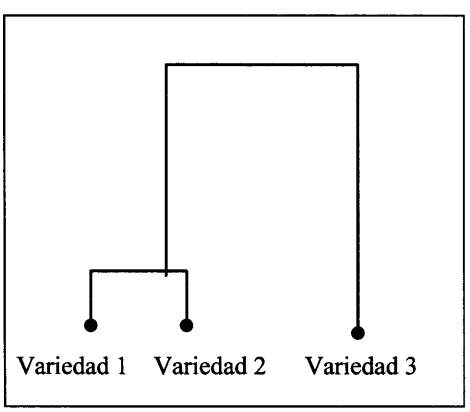
	Variedad 1	Variedad 2	Variedad 3
Variedad 1	0		
Variedad 2	1_a	0	
Variedad 3	2_b	2_c	0

a. Tipo de epéntesis: epéntesis en la periferia / epéntesis a la derecha del pronombre
b. Epéntesis y elisión de vocal
c. Epéntesis y elisión de vocal

En (9b) podemos ver las diferencias relacionadas con los procesos fonológicos. Entre las variedades 1 y 2 solo existe una diferencia derivada del tipo de epéntesis que experimentan: periférica en el caso de la variedad 1/ a la derecha del pronombre en la variedad 2. En cambio, entre estas dos variedades y la variedad 3 encontramos dos diferencias: la elisión de vocal, que se produce en 3 y no en las otras dos variedades; y la epéntesis, que no experimenta la variedad 3, a diferencia de lo que ocurre en las otras dos variedades.

De acuerdo con este análisis, el dendrograma siguiente muestra un grupo muy compacto entre las variedades 1 y 2, mientras que la variedad 3 presenta una distancia considerable en relación con este primer clúster.

(10) Representación dendrográfica de la distancia lingüística a partir de los datos analizados fonológicamente.



Con este ensayo de laboratorio que hemos realizado a partir de un ejemplo de variación extraído del clítico pronominal de 1.ª persona del singular, creemos demostrar claramente la pertinencia de basar el tratamiento cuantitativo de la distancia lingüística en datos analizados lingüísticamente. Sólo de esta manera podemos captar que las diferencias lingüísticas son menores entre las variedades 1 y 2, que la variedad es la que presenta una distancia mayor, a pesar de que fonéticamente parezca que la distancia entre las tres variedades sea similar.

5. Contraste de dos análisis dialectométricos de los datos del COD

A pesar de todo y en contra de la argumentación presentada en el apartado anterior, algunas veces se ha esgrimido que en un tratamiento cuantitativo global, con grandes cantidades de datos lingüísticos como las que pueden ofrecer un atlas o un corpus como el COD, estas diferencias pueden ser totalmente inapreciables e intrascendentes. Por eso, con la finalidad de intentar dilucidar este punto, hemos llevado a cabo un tratamiento cuantitativo contrastado de los datos del COD, que presentamos a continuación. Se trata de contrastar la representación de la distancia lingüística de los datos de dicho corpus previamente analizados fonológicamente (*vid.* Clua et *alii* 2009) con la que hemos obtenido a partir de los mismos datos, pero sin el análisis fonológico previo.

Los datos fonéticos que hemos tratado cuantitativamente coinciden con el subcorpus analizado fonológicamente de Clua et *alii* (2009). Son pues el resultado de aplicar una primera selección al conjunto total de los materiales del COD. Esta selección se basa en escoger la respuesta mayoritaria entre los informantes de un punto de encuesta, considerando que se trata de la opción más representativa del habla de una determinada localidad. A través de este proceso de filtrado, hemos elaborado varias bases de datos —una para cada ámbito morfológico estudiado— que presentan dos ventajas respecto del corpus original: por una parte, reducen la cantidad de datos a comparar, de tal manera que se simplifica notablemente el tratamiento estadístico final; por la otra, recogen los rasgos más representativas de cada uno de los puntos de encuesta y reflejan, así, las características más comunes del habla de sus habitantes.

El cómputo final de registros que han sido objeto de comparación es de 29.364; de estos, la mayor parte (20.500) corresponden a la morfología verbal, mientras que el resto (8.864) pertenecen a las otras seis categorías lingüísticas analizadas, que desglosamos a continuación: artículos, posesivos, clíticos pronominales, pronombres personales, demostrativos neutros y adverbios locativos.

Cabe decir que el cómputo de las diferencias entre variedades no se ha realizado a partir de la comparación de las diferentes formas verbales o nominales, sino a partir de la comparación de los segmentos morfológicos que las constituyen. Los segmentos morfológicos que se han tenido en cuenta en el caso de la flexión verbal han sido los de (11). ('Extensión' es un sufijo post-radical que en catalán determina la subclase verbal; 'TAM' representa las categorías de tiempo, aspecto y modo). En el caso de la flexión nominal hemos contrastado los segmentos: raíz, género ([± femenino]) y número.

(11)

	Raíz	Tema	Extensión	TAM	Número/persona
cantareu	kant	a		ɾé	w
serveixis	serb		éʃ	i	S

En la determinación de la distancia a partir del análisis lingüístico previo de los datos, además de la comparación de los segmentos morfológicos también se tuvieron en cuenta los diferentes procesos fonológicos que los afectan y que permiten explicar su realización fonética. Entre muchos otros, algunos de los procesos que hemos uti-

lizado como base de comparación en la flexión verbal son: desacentuación de la raíz (*canto* [kánto] pero *cantava* [kantáβa]), ensordecimiento de obstruyentes en situación final de palabra (*begui* [béɣi] pero *bec* [bék]), elisión de *-r* en situación final de palabra (*cantar-la* [kantárla] pero *cantar* [kantá]), etc.[5] En la flexión nominal hemos trabajado, entre otros, con los procesos de elisión de vocales y de inserción de epéntesis como los que hemos comentado en el apartado anterior.

Para establecer la comparación entre estos elementos y poder definir la distancia lingüística hemos utilizado el siguiente índice de distancia:

$$(12) \quad dist\ (i,j) = \frac{\sum_{k=1}^{long} dif_k(i,j)}{long} \times 100$$

Es decir, la distancia lingüística entre dos variedades (*i, j*) es igual al sumatorio (Σ) de las diferencias en cuanto a una variable k entre las variedades (*i, j*), dividido por *long*, que es la longitud (número de sonidos) de cada segmento morfológico comparado.

En cuanto al método de representación gráfica, nos hemos servido del *Cluster Análisis* y hemos usado un algoritmo de clasificación basado en el método UPGMA (*Unweighted Pair-Group Method Using Arithmetic Averages*) (*vid.* Sneath & Sokal 1973), que ha sido contrastado ampliamente en aplicaciones de la taxonomía numérica en múltiples disciplinas. Para la evaluación de la distorsión entre las representaciones y la distancia original, aplicamos el coeficiente de correlación cofenética, con unos resultados que corroboran la fidelidad de las representaciones jerárquicas en relación con la distancia lingüística de partida.[6]

5.1. Representaciones dendrográficas de los dos análisis dialectométricos del COD

A continuación presentamos las dos representaciones dendrográficas de la distancia lingüística entre las variedades del COD. En la figura 1 tenemos la distancia lingüística resultante del análisis dialectométrico aplicado a los datos del corpus después de ser analizados fonológicamente; en la figura 2 podemos ver el resultado de aplicar el mismo análisis dialectométrico a los datos fonéticos del COD.

De una primera aproximación al dendrograma realizado a partir del análisis fonológico se desprende que son cuatro las áreas dialectales claramente diferenciadas en el marco de la lengua catalana, en este caso parce imposible vislumbrar agrupaciones de nivel superior. Observamos un primer grupo que está formado por las variedades baleares (mallorquín, menorquín e ibicenco); un segundo bloque que comprende el catalán central y el septentrional; un tercer grupo en el que se integran tanto el catalán norte-occidental como la mayoría de las variedades de la provincia de Castelló, y, finalmente, un cuarto y último conjunto que comprende el resto de

[5] Para una análisis dialectométrico de la morfología verbal del COD, *vid.* Clua (2007).

[6] La definición de la distancia lingüística se ha realizado con el programa Microsoft ® Excel. El análisis de conglomerados y la representación gráfica de la distancia lingüística se ha llevado a cabo con el sistema de análisis multivariante GINGKO (Departament de Biologia Vegetal, Universitat de Barcelona http://biodiver.bio.ub.es/vegana/index.html).

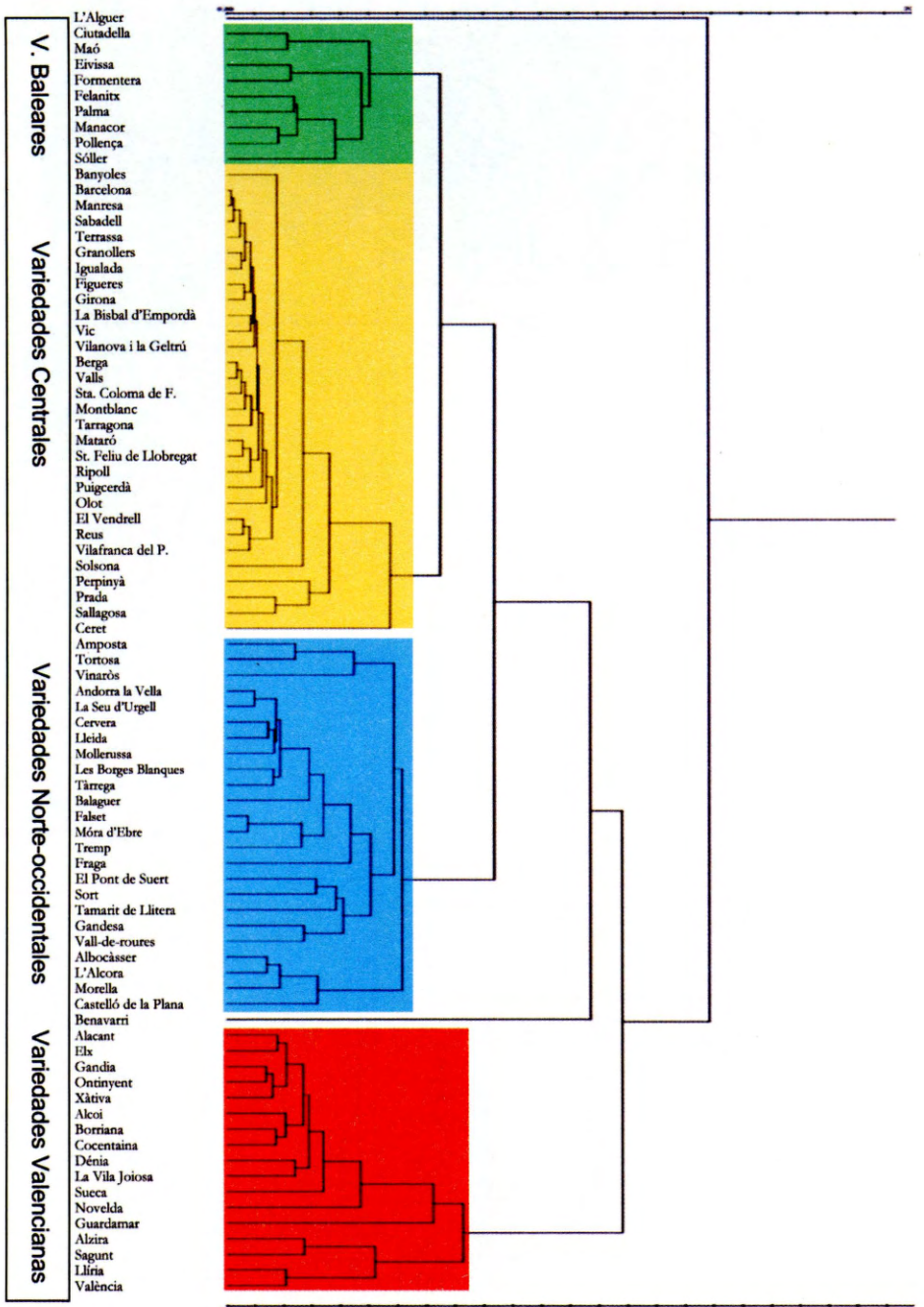


Figura 1

Distancia lingüística a partir de los datos del COD analizados fonológicamente

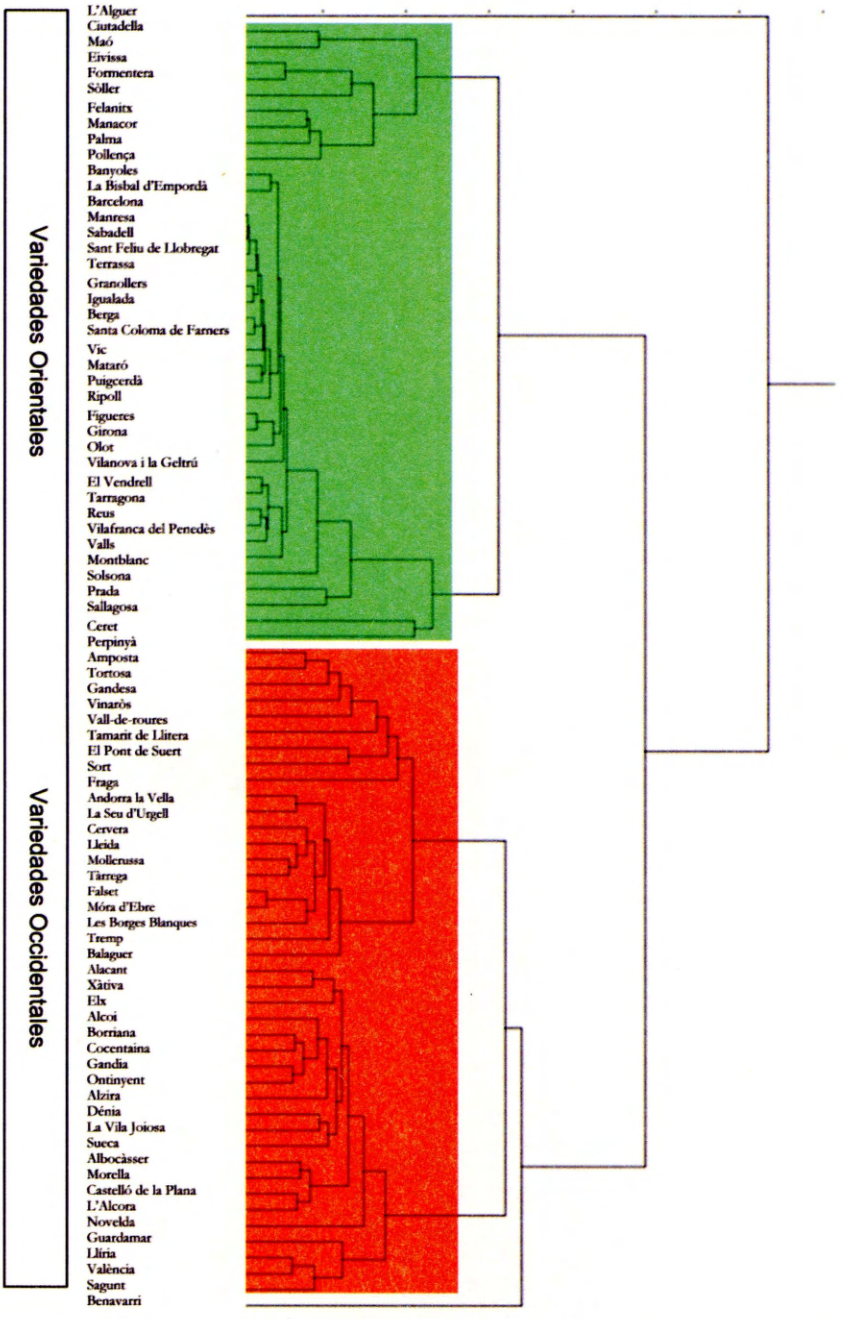


Figura 2

Distancia lingüística a partir de los datos del COD fonéticos

variedades valencianas. Por su parte, las variedades de Benavarri y, en especial, la de la localidad sarda del Alguer, ocupan posiciones notablemente aisladas; el dendrograma refleja, por consiguiente, su carácter marcadamente idiosincrático en el conjunto de la lengua catalana.

En cambio, en la representación arbórea que resulta del tratamiento de los datos fonéticos podemos discernir claramente una agrupación de nivel superior a las anteriores que coincidiría, grosso modo, con la división que ha postulado tradicionalmente la dialectología clásica catalana. Una división dialectal del catalán en variedades orientales y occidentales. Esta clasificación estaba basada en un número limitado de isoglosas y principalmente, se ceñía a los diferentes procesos de reducción vocálica, por cuya acción los sistemas vocálicos tónico y átono de estas variedades difieren sustancialmente (*vid.* Veny 1982: 17-18). Si tenemos en cuenta que cuando se postuló esta clasificación tradicional, los estudios dialectales se basaban únicamente en los datos fonéticos, no es de extrañar que se produzca esta coincidencia. Por lo que respecta a las variedades del Alguer y de Benavarri, la coincidencia con el dendrograma fonológico es casi total; en este caso también aparecen como variedades aisladas del resto de los conglomerados.

En cuanto al resto de variedades agrupadas en la clasificación tradicional bajo el epígrafe de *catalán oriental,* hay que decir que las diferencias observadas en ambos tratamientos son poco remarcables. En todo caso, se puede afirmar que en el dendograma fonológico las agrupaciones parecen en general más compactas, con una distancia lingüística interna menor.

Así, en ambas representaciones las hablas insulares, las variedades de las Baleares *y* Pitiusas, constituyen un clúster bien definido y con una clara estructura interna: por un lado, las variedades de la isla de Mallorca (Felanitx, Palma, Manacor, Pollença y, a una mayor distancia, Sóller) se agrupan entre ellas; por el otro, lo mismo ocurre en los casos de Eivissa y Formentera y de Ciutadella y Maó. Emergen, por lo tanto, las tres principales variedades baleáricas: el mallorquín, el menorquín y el ibicenco.

También hay coincidencia en las variedades del catalán central, que en constituyen un grupo especialmente homogéneo en las dos estructuras arbóreas. Pertenecen a este grupo las variedades de Banyoles, Barcelona, Manresa, Sabadell, Terrassa, Granollers, Igualada, Figueres, Girona, la Bisbal d'Empordà, Vic, Vilanova i la Geltrú, Berga, Valls, Santa Coloma de Farners, Montblanc, Tarragona, Mataró, Sant Feliu de Llobregat, Ripoll, Puigcerdà, Olot, el Vendrell, Reus, Vilafranca del Penedès y Solsona, que es la única que se sitúa a una cierta distancia del resto, posiblemente a causa de su carácter de transición entre los subdialectos central y norte-occidental. A este conglomerado, que es el más compacto de todos, se le agrupan a una distancia considerable las variedades del llamado *catalán septentrional:* Perpinyà, Prada, Sallagosa y Ceret.

Donde sí que se aprecian claramente diferencias sustanciales entre ambos análisis dialectométricos es en las agrupaciones de las variedades que tradicionalmente se han reunido bajo el rótulo de *variedades occidentales.* Se trata de diferencias que, desde nuestro punto de vista, justifican por sí solas la necesidad de realizar un análisis fonológico previo para poder determinar adecuadamente la distancia lingüística entre variedades.

La primera diferencia importante tiene que ver con el hecho que el conjunto de variedades formado por el catalán norte-occidental y el tortosino en un sentido amplio (entendido como el conjunto de variedades que constituían la antigua diócesis de Tortosa) se agrupan en primera instancia con el conjunto de hablas orientales y, sólo a continuación, con el resto de hablas occidentales, es decir con el resto de hablas valencianas.

Veamos a continuación otras diferencias relevantes. En el dendrograma fonológico las variedades tradicionalmente adscritas al catalán occidental, presentan la siguiente estructura de grupos. En primer lugar, emerge un clúster que engloba tanto las variedades nord-occidentales como aquellas hablas de transición al valenciano. Concretamente, parece razonable una distinción en cuatro subgrupos principales: (i) un conjunto de variedades homogéneas que se corresponde, a grandes rasgos, con el denominado leridano: Cervera, Lleida, Mollerussa, les Borges Blanques y Tàrrega. A este grupo se añaden también dos variedades pirenaicas (Andorra y la Seu d'Urgell) y, cada vez a mayor distancia Balaguer, en primer lugar; Falset, Móra d'Ebre y Tremp, a continuación; y, finalmente, el habla de Fraga; (ii) un segundo conjunto de variedades, probablemente más conservadoras,[7] que se agrupan, en primer término, entre ellas, y a continuación, con el clúster anterior: se trata de las hablas del Pont de Suert, Sort, Tamarit de Llitera, Gandesa y Vall-de-roures; (iii) un tercer grupo, compuesto por las variedades de Amposta, Tortosa y Vinaròs, que constituyen el núcleo del dialecto tortosino, y, finalmente, (iv) un último cluster, que aunque aparezca físicamente alejado del anterior (entre ambos aparecen los grupos (i) y (ii)) en realidad está a muy poca distancia lingüística, que incluye las variedades de Albocàsser, l'Alcora, Morella y Castelló de la Plana. Las variedades de este último grupo, que a menudo han sido denominadas *variedades de transición entre el catalán y el valenciano,* se unen claramente al clúster de variedades nord-occidentales, con lo cual creemos poder aclarar considerablemente la filiación de estas hablas, que a menudo ha sido objeto de discusión porque, dependiendo de la isoglosa utilizada (1.ª persona del presente de indicativo *cantelcanto*; 3.ª persona del mismo tiempo *cantelcanta*; 2.ª persona del imperfecto de subjuntivo *cantareslcantesses*...), se habían vinculado más a las variedades nord-occidentales o a las variedades valencianas.

En cambio, en el dendrograma obtenido a partir de los datos fonéticos del COD las variedades nord-occidentales presentan una estructura mucho menos coherente. En principio, las agrupaciones en torno al núcleo central, o leridano, son parecidas a las anteriores, pero a partir de aquí las divergencias son manifiestas. De entrada, al lado de este primer grupo, sólo existe un segundo clúster muy poco compacto donde se agrupan las variedades del tortosino estricto (Amposta, Tortosa, Gandesa y Vinaròs) con las hablas más occidentales (Fraga, Pont de Suert, Sort, Tamarit de Llitera y Vall-de-roures). Por su parte, el grupo de variedades del norte de Castelló (con la excepción de Vinaròs), que en la clasificación anterior se agrupaba con las variedades del tortosino, aquí se sitúa en el centro del clúster del resto de variedades valencianas, a una distancia demasiado importante, para ser coherente, del tortosino estricto.

[7] Sobre la base de las conclusiones de Viaplana (1999: 83-109).

Por lo que respecta al resto de variedades valencianas, en el diagrama fonológico las hablas de las actuales provincias de València y Alacant (y también la variedad de Borriana) conforman, por último, un clúster que presenta, a su vez, una subdivisión en dos grupos: uno mayoritario, que engloba a las variedades del valenciano central y meridional;[8] y otro formado tan sólo por cuatro localidades, representativas del denominado valenciano *apitxat*. Se integran en el primer grupo las hablas de Alacant, Elx, Gandia, Ontinyent, Xàtiva, Alcoi, Borriana, Cocentaina, Dénia, la Vila Joiosa, Sueca y, a mayor distancia, Novelda y Guardamar. El segundo grupo incluye, en cambio, las variedades de Alzira, Sagunt, Llíria y València.

Por el contrario, en el dendrograma obtenido a partir de los datos fonéticos, estas variedades se agrupan en diferentes subgrupos con escasa coherencia. Aparte de la inclusión en el centro del clúster de las variedades del norte de Castelló, que ya hemos comentado, parece poco coherente que al grupo de variedades *apitxadas* nucleares (València, Llíria y Sagunt) se le añada antes Guardamar que Alzira, una variedad que tiene muchos rasgos del valenciano *apitxat*. Tampoco nos parece coherente la distribución de las variedades pertenecientes al valenciano meridional.

6. Conclusiones

Después de analizar las diferencias entre los dendrogramas de los análisis dialectométricos del COD, con y sin análisis fonológico previo, creemos que esta comparación realizada a partir de un corpus de datos considerable corrobora las hipótesis a las que habíamos llegado con los ensayos de laboratorio anteriores. En el sentido que los resultados de un análisis dialectométrico pueden ser considerablemente diferentes si partimos de datos fonéticos o si lo hacemos después de analizar fonológicamente estos mismos datos.

Es cierto que las diferencias no son tan grandes como las que se podían prever a partir del ensayo realizado con los clíticos pronominales, pero de todos modos consideramos que para describir adecuadamente la distancia lingüística entre un grupo de variedades cuanta más información pongamos en contraste más próxima a la realidad será la representación resultante.

Por otra parte, un análisis lingüístico de este tipo permite discriminar claramente el carácter lingüístico de los diferentes fenómenos que entran en juego en la variación. Nos permite introducir distinciones cualitativas en los resultados cuantitativos. Podemos discernir si la distancia obtenida está relacionada con los elementos subyacentes o con los procesos fonológicos, por ejemplo. Como se afirma en Viaplana (1994):

> La distinción entre elementos predecibles y elementos impredecibles en la estructura lingüística permite la discriminación de fenómenos diferenciales cruciales en la variación dialectal que, en ausencia de esta distinción, quedan amalgamados en la simple distinción de las formas.

En más de una ocasión se ha esgrimido como crítica a la descripción cuantitativa de la variación lingüística un cierto grado de menoscabo del análisis lingüístico. Así se ha señalado que una de las deficiencias que presentan algunos tratamientos cuantitativos de la variación dialectal tiene que ver con la falta de un análisis lingüístico

[8] En el sentido de Clua (1999a).

coherente previo a la transposición de los datos fonéticos a las variables de comparación que sirven de base para el análisis cuantitativo. A causa de la gran variación observable en cualquier estudio dialectal, a menudo se ha tendido a una cierta tipificación de los resultados, es decir, a una simplificación de dicha variación; si este proceso no se sustenta en criterios lingüísticos coherentes puede pasar que se utilicen criterios muy dispares (sincrónicos y diacrónicos, interdialectales e intradialectales, etc.) o que las variables de las que parte el proceso clasificatorio no reflejen adecuadamente la variación real. De ahí la importancia que tiene desde nuestro punto de vista el análisis lingüístico de los datos.

Referencias bibliográficas

Bonet, E. & M.-R. Lloret, 2005, «More on alignment as an alternative to domains: The syllabification of Catalan clitics», *Probus* 17, 1, 37-78.

Clua, E., 1999a, *Variació i distància lingüística. Classificació dialectal del valencià a partir de la morfologia flexiva*, tesis doctoral, Universitat de Barcelona.

—, 1999b, «Distància lingüística i classificació de varietats dialectals», *Caplletra* 26, 11-26.

—, 2007, «Distancia lingüística entre los dialectos del catalán a partir de los datos del COD», comunicación presentada en el *XXVième Congrès International de Linguistique et de Philologie Romanes*, Innsbruck. (Aparecerá publicado en P. Danler *et alii* (eds.), *Actes du XXVième Congrès International de Linguistique et de Philologie Romanes (Innsbruck 2007)*, Niemeyer, Tübingen).

— & M.-R. Lloret, 2006, «New tendencies in geographical dialectology: The Catalan Corpus Oral Dialectal (COD)», en J-P. Montreuil (ed.), *New Perspectives on Romance Linguistics. Vol. 2: Phonetics, phonology, and dialectology*, Amsterdam/Philadelphia, John Benjamins, 31-47.

— & —, 2007, «Clasificación de variedades dialectales mediante técnicas de análisis multivariante, a partir de un corpus oral», en P. Cano López *et alii* (eds.), *Actas del VI Congreso de Lingüística General (Santiago de Compostela, 3-7 de mayo de 2004)*, vol. III: *Lingüística y variación de las lenguas*, Arco/Libros, Madrid, 3057-3068.

—, Valls, E. & J. Viaplana, 2008, «Analisi dialettometrica del catalano partendo dai dati del COD. Una prima approssimazione alla gerarchia tra varietà», en G. Blaikner-Hohenwart *et alii* (eds.), *Ladinometria Festschrift für Hans Goebl zum 65. Geburtstag*, vol. 2, Universität Salzburg *et alii*, Salzburg, 27-42.

—; Lloret, M-R & E. Valls, 2009, «Análisis lingüístico y dialectométrico del *Corpus Oral Dialectal* (COD)», en P. Cantos Gómez & A. Sánchez Pérez (eds.) *A survey on corpus-based Research*, Asociación española de lingüística de corpus: 1033-1045.

Durand, J. P., 1889, «Notes de philologie rouergate», *Revue des langues romanes* 33, 47-84.

Lloret, M.-R., 2004, «The phonological role of paradigms: The case of insular Catalan», en J. Auger, C. Clements & B. Vance (eds.), *Contemporari Approaches to Romance Linguistics*, John Benjamins, Amsterdam/Philadelphia, 275-279

— & J. Viaplana, 1998, «Variació morfofonològica. Variants morfològiques», *Caplletra* 25, 43-62.

Sneath, P. H. A. & R. R. Sokal, 1973, *Numerical Taxonomy. The Principles and Practice of Numerical Classification*, W. H. Freeman and Company, San Francisco.

Veny, J., 1982, *Els parlars catalans (síntesi de dialectologia)*, Moll, Palma.

Viaplana, J., 1999, *Entre la dialectologia i la lingüística*, Publicacions de l'Abadia de Montserrat, Barcelona.

—; Lloret, M.-R.; Perea, M.-P.; Clua, E., 2007, *COD. Corpus Oral Dialectal.* Barcelona, PPU. (Publicación en CD-rom).

EL PROCESAMIENTO INFORMÁTICO DE LOS MATERIALES DEL *ATLAS LINGÜÍSTICO DE LA PENÍNSULA IBÉRICA* DE TOMÁS NAVARRO TOMÁS

Pilar García Mouton
ILLA-CCHS (CSIC)

Abstract

Estas páginas dan noticia de la puesta en marcha de un proyecto intramural del CSIC para elaborar y editar definitivamente, en soporte informático, los materiales del Atlas Lingüístico de la Península Ibérica, *que Tomás Navarro Tomás dirigió en la primera mitad del siglo XX.*

These pages announce the launch of the internal CSIC project to produce and definitively edit, in digital format, the materials from the Linguistic Atlas of the Iberian Peninsula, *directed by Tomás Navarro Tomás during the first half of the 20th century.*

Palabras clave: *Procesamiento informático del ALPI, Geolingüística peninsular s. XX.*

Key words: *ALPI Information Technology, Peninsular Geolinguistic XXth century.*

1. Introducción

Redacto estas páginas en calidad de coordinadora del proyecto intramural del Consejo Superior de Investigaciones Científicas (CSIC)[1] para elaborar y editar los materiales del *Atlas Lingüístico de la Península Ibérica* (ALPI).

Los colegas conocen bien las características de este atlas, que fue un proyecto emblemático del Centro de Estudios Históricos de la Junta para la Ampliación de Estudios, dirigido por Tomás Navarro Tomás e impulsado por Ramón Menéndez Pidal, en la línea de los atlas de gran dominio que se plantearon en Europa en la primera mitad del siglo XX. Desde que se tomó la decisión de llevarlo a cabo hasta que empezaron los trabajos de campo pasaron muchos años, pero las encuestas del ALPI ya estaban prácticamente terminadas en las partes correspondientes a las hablas asturleonesas, aragonesas, castellanas y catalanas cuando la Guerra Civil española las interrumpió.

---

[1] Proyecto intramural del CSIC, de referencia 200410E604; Investigadora principal: Pilar García Mouton; Título: *Elaboración y edición de los materiales del* Atlas Lingüístico de la Península Ibérica (ALPI).

Para la zona castellanohablante los encuestadores fueron Aurelio M. Espinosa hijo, Lorenzo Rodríguez Castellano, Aníbal Otero y Manuel Sanchis Guarner; para la zona gallegoportuguesa, Aníbal Otero y Rodrigo de Sa Nogueira, sustituido primero por Armando Nobre de Gusmão y, después, por Luís F. Lindley Cintra, y para la zona de hablas catalanas, Francesc de Borja Moll y Manuel Sanchis Guarner.

Años después, acabadas las encuestas pendientes, el CSIC publicó un tomo de *Fonética* con 75 mapas en el año 1962, pero después los trabajos de edición se abandonaron hasta hoy. En los últimos años David Heap, que había localizado los materiales de encuesta, fue colgando fotocopias de ellos en internet, y de esa manera les volvió a dar vida en cierta forma.

2. Nuestro proyecto de elaborar y editar los materiales del ALPI

En el verano del año 2007 el CSIC se planteó la posibilidad de reiniciar la edición del ALPI y decidió encargar un nuevo proyecto para elaborar y editar los materiales del atlas a un equipo, coordinado por Pilar García Mouton (CSIC), integrado por Inés Fernández Ordóñez (Universidad Autónoma de Madrid), David Heap (Universidad de Western Ontario), Maria Pilar Perea (Universidad de Barcelona), João Saramago (Centro de Lingüística de la Universidad de Lisboa) y Xulio Sousa (Instituto da Lingua Galega de la Universidad de Santiago de Compostela). El CSIC asumió como propio el proyecto en el mes de marzo de 2010, como uno de sus proyectos intramurales, después de haberlo sometido a una estricta evaluación internacional. El resultado previsto no mantendrá la forma del primer volumen del ALPI, en soporte papel, porque se concibe como un atlas interactivo de libre acceso que se podrá consultar desde un geoportal del CSIC.

2.1. Cuestiones metodológicas

Es evidente que desde que se hicieron las encuestas del ALPI hasta ahora han cambiado muchas cosas, pero aquí resulta especialmente apropiado referirse a dos de ellas. En primer lugar, han cambiado radicalmente la cultura y las hablas de los territorios estudiados y, en segundo lugar, también ha cambiado mucho la metodología geolingüística en general. Ahora bien, ninguna de las dos circunstancias merma el interés del ALPI, más bien al contrario: los atlas lingüísticos se refieren siempre a una época determinada que no tiene por qué coincidir con la de quienes los utilizan; lo contrario los convertiría en obras casi efímeras. De ahí que sean precisamente las fechas en las que se recogieron sus materiales las que hacen del ALPI un archivo histórico fiable y segmentado. Como ya señaló el mismo Navarro Tomás (1975: 14) en el prólogo a su libro *Capítulos de geografía lingüística de la Península Ibérica,* "Noticia histórica del ALPI": "Por virtud principal de su información fonética, el ALPI es como una especie de acta documental del carácter y fisonomía del habla popular de la Península en los años inmediatamente anteriores a la guerra civil. La honda conmoción producida por esta guerra en todo el país, y el movimiento de población ocasionado después por motivos económicos y sociales, habrán modificado sin duda alguna las líneas del ALPI, lo cual acentúa su interés como testimonio de valor histórico."

La segunda circunstancia, la relativa a la actualización de su metodología, sólo sería exigible a un atlas que se planteara hoy, no a uno ideado a principios del siglo xx. Con algunas características propias, el ALPI es un atlas equiparable a los similares de su época. Y, finalmente, no conviene olvidar que, a estas alturas, no disponemos de ningún otro atlas de gran dominio peninsular en el que se puedan contextualizar los distintos atlas zonales para establecer comparaciones productivas. En este sentido, buscando algún otro paralelo, se podría aducir que en las últimas décadas se están publicando el *Atlante Linguistico Italiano* (1995-1999), cuyas encuestas se hicieron después de la primera gran guerra; acaba de aparecer el tomo IV del *Atles Lingüístic del Domini Català* de Joan Veny y Lídia Pons i Griera (2009), a partir de unas encuestas relativamente lejanas en el tiempo; se están digitalizando otros atlas históricos, y algunos colegas, como Hans Goebl (2002), revisitan con metodología dialectométrica el *Atlas Linguistique de la France* (ALF) o el clásico atlas italosuizo, el *Sprach- und Sachatlas Italiens und der Südsweiz* (AIS).

Además de estas razones científicas, existen otras de tipo ético que el CSIC ha tenido en cuenta a la hora de apoyar este proyecto, fundamentalmente el compromiso de poner a disposición de la comunidad científica y de los hablantes en general un patrimonio que les pertenece y saldar así una deuda histórica con unos investigadores que pusieron un gran esfuerzo profesional en los trabajos de este atlas (García Mouton 2007).

2.2. Descripción del proyecto

Nuestro proyecto tiene como objetivo elaborar y editar los materiales del ALPI utilizando las posibilidades que la tecnología informática proporciona actualmente a los trabajos geolingüísticos, no sólo para el cartografiado automático, sino también para realizar búsquedas de todo tipo en el corpus completo de datos lingüísticos y etnográficos que conforman un atlas de estas características. Se trata de hacerlas sobre un fondo georreferenciado que cubra las necesidades fundamentales de cualquier proyecto geolingüístico. Una vez concluido, se podrán realizar búsquedas más o menos avanzadas y obtener resultados puntuales, pero también mapas —mapas clásicos (con la respuesta recogida en cada uno de los puntos de encuesta), mapas por áreas o por isoglosas, mapas simbólicos— y cualquier otro tipo de elaboración que esté previsto en la herramienta informática que se está diseñando para tratar los datos y en la estructura del geoportal que los va a acoger. De este modo, cabe esperar que las mismas circunstancias que supusieron un grave retraso en la publicación del ALPI a la larga acaben beneficiando la difusión de estos materiales históricos, que finalmente se podrán consultar en la forma más novedosa.

Como se ha dicho, la edición del atlas está planteada casi exclusivamente en soporte informático a través de la red, a partir de una base de datos georreferenciada alojada en un servidor del CSIC, con las posibilidades que los actuales Sistemas de Información Geográfica (SIG) proporcionan para convertir la futura web del ALPI en una IDE (Infraestructura de Datos Espaciales). Esto significa que sus resultados podrán ser confrontados con otros cualesquiera que hayan sido tratados de forma similar. El soporte papel se reservará, en todo caso, para algunos mapas que puedan servir testimonialmente como muestras representativas institucionales.

2.3. Plan de trabajo

Se trabajará con los cuestionarios correspondientes a los 527 puntos de encuesta del ALPI, que se distribuyen así: 53 para Galicia: 93 para Portugal; 78 para Asturias, León y Extremadura; 90 para las dos Castillas y Albacete; 71 para Andalucía y Murcia; 40 para Navarra y Aragón; y 104 para Andorra, el Rosellón, Cataluña, Valencia y las islas Baleares. Cada cuestionario está dividido en un *Cuaderno I*, con unas preguntas sobre el informante y la localidad y unas notas de orientación fonética, más el cuerpo del cuestionario propiamente dicho, con 411 cuestiones de fonética, morfología y sintaxis, y un *Cuaderno II*, con cuestiones de léxico que alcanzan hasta la pregunta 828 en la versión sintética IIG, mientras que la versión IIE amplía muchas de ellas con varias subpreguntas.

2.3.1. Digitalización y herramienta para introducir datos

Para elaborar los materiales se están abordando actualmente dos cuestiones previas: la digitalización de los cuestionarios —que se ha hecho básicamente en el Instituto da Lingua Galega (ILG) de la Universidad de Santiago de Compostela bajo la responsabilidad de Xulio Sousa— y el diseño de una herramienta informática específica —que estamos elaborando en el Centro de Ciencias Humanas y Sociales del CSIC con el apoyo de la Unidad TIC (Tecnologías de la Información y la Comunicación), especialmente con el de su responsable, Juan Carlos Martínez, y el de Ángel Díaz del Castillo, y la ayuda de la Unidad SIG (Sistemas de Información Geográfica), especialmente la de su responsable, Isabel del Bosque, y la de Carlos Fernández Freire—. La digitalización está prácticamente resuelta, algo que estimamos fundamental desde el primer momento porque va a proporcionar imágenes de los materiales de mucha calidad para trabajar, con la ventaja de tenerlas permanentemente disponibles en pantalla y de poder ampliarlas, unas imágenes que resultan imprescindibles por las características específicas de este proyecto.

Al encarar el proceso de elaboración de los materiales, había que partir del hecho de que los encuestadores del ALPI utilizaron muy a conciencia el alfabeto fonético de la *Revista de Filología Española* (AFE), un alfabeto creado por el propio Tomás Navarro Tomás, publicado en la revista en 1915, y que hoy cuenta con una larga tradición en el mundo hispánico. Ese alfabeto reflejaba el interés fonético de su autor, y de su época, con un lujo de detalles que hoy puede parecer prolijo en exceso y, por ese motivo, puede reducir su consulta por parte de los especialistas que no están familiarizados con él.

A la hora de abordar la edición del ALPI, se planteaba un dilema científico: conservar la estricta transcripción que tanto preocupó a su director y sus encuestadores o intervenir, simplificándola, para hacerla más asequible y volcarla al Alfabeto Fonético Internacional (AFI), lo que aseguraría al ALPI una difusión mayor y posibilitaría la recuperación ágil de los datos en un sistema de búsqueda complejo. Creemos haber encontrado una solución equilibrada que va a permitir conservar la transcripción fonética original de los encuestadores —porque siempre se dará la posibilidad de acceder a la imagen digitalizada de las respuestas originales manuscritas— y conseguir búsquedas ágiles a partir de su conversión al Alfabeto Fonético Internacional, que facilitará la elaboración de los datos.

La cuestión del respeto a la transcripción original no deja de ser básica. Las siguientes palabras de Navarro Tomás (1975: 19) prueban la importancia que los investigadores del ALPI daban a su minuciosa transcripción: "El valor demostrativo de estos mapas [...] es mérito de la detallada precisión de sus transcripciones. Una investigación realizada con menos rigor analítico y con transcripción menos estrecha, habría resbalado por encima de muchos de estos pormenores tan significativos para el cabal conocimiento de la materia. [...]. Los aparatos mecánicos sólo prestan una ayuda relativa. El oído convenientemente ejercitado sigue siendo el instrumento más perfecto. Muchas veces, al repasar en la oficina una cinta magnetofónica, se echa de menos la presencia de la persona que la inscribió."

2.3.2. La simplificación de las transcripciones originales

Antes que nada hay que tener en cuenta que los mismos editores del tomo I del ALPI, de acuerdo con Navarro Tomás, decidieron proceder desde el principio a una primera simplificación de las transcripciones de sus encuestas. Por el *Epistolario* que acaban de publicar Vicent García Perales y Santi Cortés (2009: cartas 135, 138, 157, etc.) sabemos que existió una "plantilla" elaborada con la ayuda del maestro, plantilla que trajeron de su viaje a Nueva York Manuel Sanchis Guarner y Lorenzo Rodríguez Castellano y que no se ha conservado. La única solución pasaba por tratar de reconstruirla confrontando la transcripción de los cuestionarios y los resultados impresos. Decidimos trabajar después en otra etapa más en la simplificación, para establecer una tabla de equivalencias entre los signos empleados con los del AFI, el alfabeto en el que se van a introducir los datos. Para poder hacerlo con garantías, antes de implementar la tabla de equivalencias fonéticas en la herramienta informática, se pidió ayuda a varios especialistas en la fonética de los distintos dominios lingüísticos —Amélia Andrade, para las hablas portuguesas; Francisco Dubert, para las gallegas; Ralph Penny, para las hablas asturleonesas; Daniel Recasens, para las catalanas y Juana Gil, para las castellanas— que trabajaron un tiempo con los materiales. Los días 7 y 8 de mayo de 2009 el equipo de nuestro proyecto se reunió con ellos, gracias a la hospitalidad de la Universidad Autónoma de Madrid,[2] y allí quedó establecida la correspondencia que se ha decidido adoptar entre los signos fonéticos del Alfabeto de la *Revista de Filología Española* de los originales y los signos del Alfabeto Fonético Internacional, que es mucho más sencillo. Partimos de la base de que, como elaboradores y editores del ALPI, no entramos a juzgar unas transcripciones heredadas, ni intervenimos en la transcripción original más que para establecer las equivalencias con el AFI, en nuestro interés de evitar que se pueda llegar a perder cualquier matiz imprescindible para la caracterización de las hablas estudiadas. Publicaremos con el atlas la tabla de equivalencias y, de todas formas, quienes consulten el ALPI, si lo desean, podrán *ver* literalmente la transcripción original de cada respuesta en cada uno de los cuestionarios.

[2] Gestionada por Inés Fernández Ordóñez.

2.3.3. El diseño de la herramienta para la introducción de los datos

Hecho lo anterior, empezamos el diseño de la herramienta informática —que parte de una base de datos relacional— para introducir los datos. La herramienta, aún sin terminar, está bastante avanzada: incluirá la imagen digitalizada de las respuestas en transcripción AFE, los distintos teclados que permitirán introducir esas respuestas en transcripción AFI —con una ayuda permanente—, la grafía en las distintas lenguas iberorrománicas, las imágenes y los contenidos etnográficos, y tendrá funciones para la premarcación fonética, morfológica, sintáctica y léxica previsibles en cada una de las cuestiones. En otra vertiente, dispondrá de controles de acceso personalizados y jerarquizados para los distintos colaboradores —ya que se trabajará en ella a través de la red desde distintos sitios—, y también de funciones para optimizar el control y la corrección del trabajo.

2.4. Lo etnográfico en el ALPI

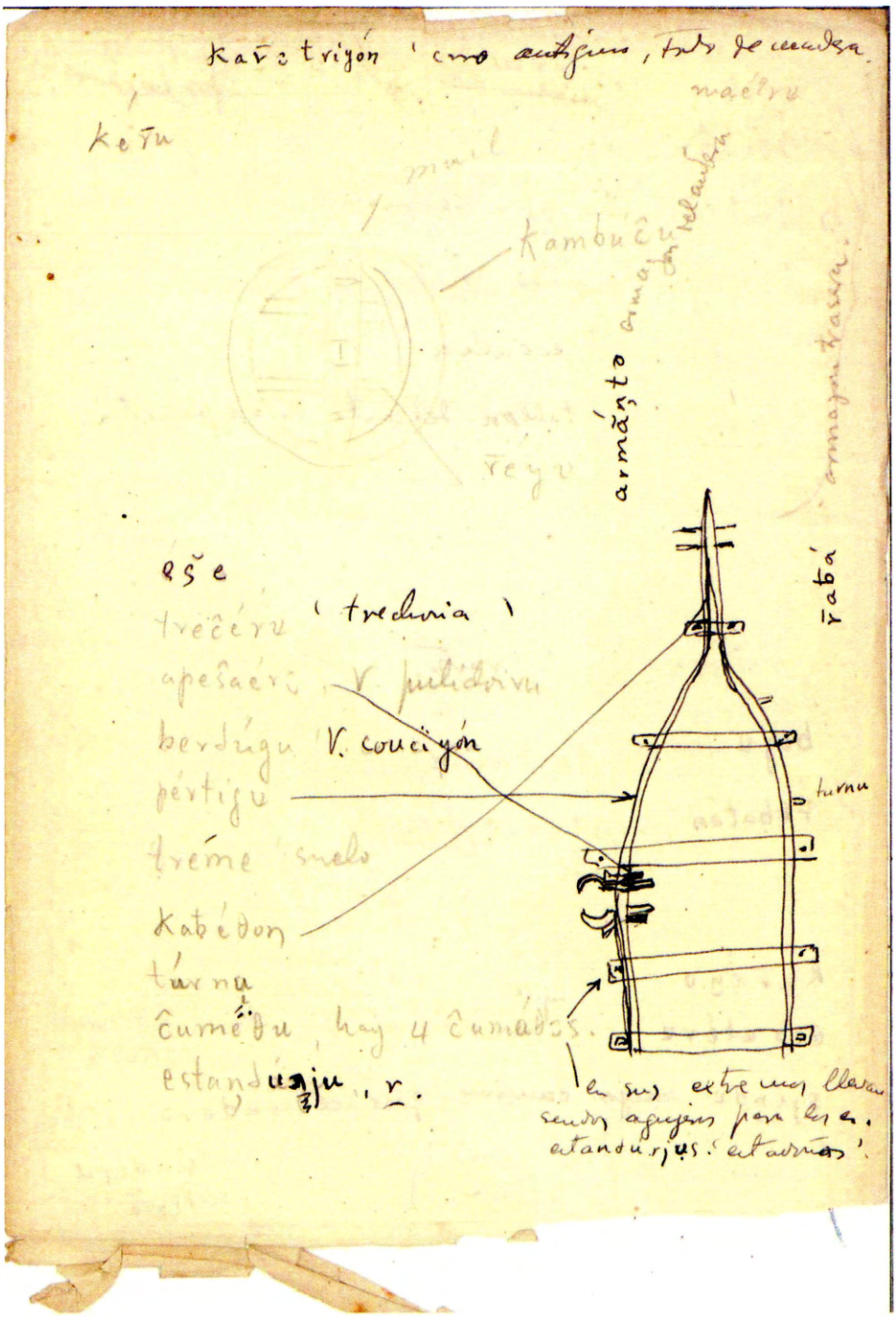
Conviene advertir que el proyecto de edición tiene muy en cuenta el hecho que, junto a las respuestas, en los cuestionarios aparecen con alguna frecuencia informaciones complementarias sobre cultura material y espiritual que los encuestadores anotaron a lo largo de la encuesta, a veces en forma gráfica, con dibujos sobre enseres o utensilios (ver figura en la página siguiente). Está previsto dar un espacio adecuado en nuestro proyecto a estos datos etnográficos, ya que, además de su valor intrínseco, para la mayor parte de la Península, son los últimos recogidos por especialistas antes de la Guerra Civil, circunstancia que los convierte en un testimonio excepcional de una cultura que ya no existe.

De nuevo la voz de Navarro Tomás (1975: 12-13) defiende el peso de estos contenidos en el ALPI, que, en ese sentido, fue acusado de ser un atlas desfasado, meramente fonético. Defiende Navarro la atención que habían prestado a lo etnográfico, aunque sus palabras sólo afirmen que el cuestionario seguía una organización temática similar a la del AIS: "Para la sección de léxico fue de gran ayuda el Atlas italosuizo de Jaberg y Jud, cuyos volúmenes empezaron a aparecer por esa fecha. Adoptamos su organización por temas etnográficos siguiendo el orden de fenómenos atmosféricos, accidentes geográficos, flora, fauna, cuerpo humano, familia, hogar, labores agrícolas, oficios artesanos, herramientas, animales domésticos, etc. Sobre esta base, el ALPI hubiera podido llamarse Atlas lingüístico y etnográfico, como de hecho lo es, aunque no pareciera indispensable indicarlo en el título."

2.5. La estructura del geoportal

Otra parte del proceso de elaboración y edición será la dedicada al diseño de sistemas que faciliten el volcado automático de datos para responder a las búsquedas en línea del geoportal, que permitirán hacer mapas a la carta, búsquedas de todo tipo sobre los datos, enlaces con las referencias previstas a otros mapas sobre los mismos *items* de cualquier otro atlas, acceder a la traducción de las cuestiones a las principales lenguas de cultura, etc.

Estamos aún en las primeras fases. Nos hubiera gustado aportar —como pensamos en nuestros planteamientos más optimistas— el diseño definitivo de la herra-


kaṽ e triyón 'cuno cubierta, todo de madera'
keṽu
kambúču
armáṇto
ṽabá
ṽeyu
ešẽ
trečéru ('trechuia')
apešaéru, V. julidiru
berdúgu V. coucíyon
pértigu
tréme 'suelo'
kabédon
túrna
ĉumédu, hay 4 ĉumados.
estanduúju, r.
turnu
en sus extremos llevan
cuñas agujeros para los
estanduúju: estacones'

mienta para introducir datos y avanzar el planteamiento del resto, pero en estos momentos estamos definiendo cuidadosamente las cuestiones relacionadas con el etiquetado morfológico y sintáctico y las distintas posibilidades de búsqueda. En nuestras conversaciones con los colegas de SIG tenemos que aclarar continuamente que no es el momento de plantear elaboraciones avanzadas a partir de los datos y de su posibilidad de georreferenciación —esas elaboraciones vendrán después—, sino que estamos ante la necesidad de editar y publicar definitivamente unos materiales que podrán ser consultados en línea sin elaboraciones metodológicamente orientadas que los condicionen. Nuestro primer objetivo es poner a disposición de la comunidad científica los materiales inéditos del ALPI.

Referencias bibliográficas

García Mouton, P., 2007, «La JAE y la filología española», en M. Á. Puig-Samper Mulero (ed. científico), *Tiempos de investigación. JAE-CSIC, cien años de ciencia en España*, CSIC, Madrid, 155-159.

Goebl, H., 2002, «Analyse dialectométrique des structures de profondeur de l'ALF», *Revue de linguistique romane* 66, 5-63.

Istituto dell'Atlante Linguistico Italiano, 1995-1999, *Atlante Linguistico Italiano*, I-IV, dir. M. G. Bartoli, G. Vidossi, B. A. Terracini, G. Bonfante, C. Grassi, A. Genre & L. Massobrio, Istituto Poligrafico e Zecca dello Stato, Roma.

La historia interna del Atlas Lingüístico de la Península Ibérica *(ALPI). Correspondencia (1910-1976)*, 2009, introducción, selección y notas de S. Cortés Carreres & V. García Perales, UPV, Valencia.

Navarro Tomás, T., 1975, *Capítulos de Geografía Lingüística de la Península Ibérica*, Instituto Caro y Cuervo, Bogotá.

Veny, J. & L. Pons i Griera, 2001-2009, *Atles Lingüístic del Domini Català*, I-IV, Institut d'Estudis Catalans, Barcelona.

UN RETRATO DEL ARTÍCULO VASCO EN EL AÑO 1895, MEDIANTE EL PROGRAMA *VDM*

Ekaitz Santazilia
UPV/EHU

Abstract

In this paper we offer a concrete application of the computer program VDM, which, given the input of the late XIXth century Bourciez corpus, provides us with clear information about the geographical and functional distribution of the article in Basque at that time. After describing the advantages and defects of the method employed, we discuss a series of linguistic theories about the article, to see how they fit in with the new data provided. Furthermore, by means of this method, which provides data about the geographical distribution of the article, we offer valuable data to explain its progressive introduction into the Basque language. At the same time, we show how the diatopic, as well as the diachronic use of the determiner depends on the type of clause, making it possible to establish a hierarchy and a relative chronology.

Keywords: *Article, Basque, Bourciez collection, VDM program, variation.*

Osaba Jesusendako
Para el tío Jesús
Sortzen denak hiltzea zor

0. Introducción[1]

Son pocas las reflexiones realizadas a propósito de los métodos empleados para la investigación lingüística, al menos en lo que al vasco se refiere. La renovación, crítica y mejora de las teorías son el pan de cada día de la actividad científica, pero es necesario, al mismo tiempo, realizar un análisis de los métodos empleados a tal efecto. Son todavía menos lo simposios que he conocido como ponente (cualidad intrínseca del neófito, no me puedo preocupar): gracias, por tanto, al grupo de investigación EUDIA,

[1] Este trabajo ha podido llevarse a cabo, en parte, gracias a una beca de colaboración del MEPSyD, en el curso 2008-2009. La ponencia original fue realizada en euskara, bajo el título: "Euskal artikuluaren 1895eko argazki bat, *VDM* programaren bitartez".

por potenciar dichas reflexiones tan necesarias para la lingüística y por ofrecer a novatos como yo la opción de comenzar a caminar: y todo eso en un sólo simposio.[2]

Entremos definitivamente en el tema. Hemos realizado una aplicación de un programa informático desarrollado para la dialectometría sincrónica, con el fin de conocer la forma, extensión y estatus del artículo en vasco. No somos nosotros, claro está, los primeros en invertir tiempo pensando en este tema; ni siquiera somos los que más tiempo hemos invertido, pero esta inversión pretende ofrecer un paso más en la descripción de la extensión sincrónica y diacrónica del artículo vasco, empleando para ello los recursos que la informática nos ofrece.

Perdone, pues, el lector, desde el primer momento, los errores que pueda encontrar en el texto y prosigamos.

1. Un poco de luz a la oscuridad del título

Este capítulo pretende ofrecer algunas aclaraciones para la correcta inteligencia del título de este trabajo. En primer lugar, daremos unas generalidades sobre el artículo; explicaremos después el porqué de 1895 y hablaremos también de ciertas cualidades del programa informático *VDM*.

1.1. El artículo

No tenemos tiempo ni espacio aquí para buscar una definición de *artículo*. De todos modos, todo aquél que haya tenido contacto con la lingüística europea ha oído hablar con frecuencia, o al menos esporádicamente, sobre el artículo. Pero de lengua a lengua la presencia y empleo de éste cambia; es decir, en algunas lenguas es necesario sólo el sintagma nominal (SN), donde en otras, para lograr la misma interpretación semántica, es necesario el sintagma determinante (SD).

Longobardi (2001) hace una clasificación tipológica en función del nivel de empleo del artículo. Divide las lenguas de maneras diferentes, dependiendo del nivel de tolerancia de sintagmas nominales desnudos (sin artículo), y de la interpretación semántica que les corresponde:

(1) a) Lenguas sin SN desnudos.
b) Lenguas con contados SN desnudos.
c) Lenguas con SN desnudos más libres.
d) Lenguas con SN singulares desnudos indefinidos.
e) Lenguas con SN desnudos exclusivamente (carecen de artículo y tienen interpretaciones ambiguas).

Longobardi inserta al vasco junto con el francés en el grupo de (1a), asumiendo que mientras que en estas lenguas los SD pueden ser argumentos, los SN no (Lon-

[2] Quiero expresar mi agradecimiento a Ana Gándara, pues con ella comencé en este tema; a Javier Ormazabal, porque dirigió esos primeros pasos; a Xabier Artiagoitia, Urtzi Etxeberria, Iván Igartua y Julen Manterola, porque las conversaciones mantenidas con ellos han sido realmente esclarecedoras y finalmente, al grupo de investigación EUDIA y en especial a Gotzon Aurrekoetxea y Aitor Iglesias, porque han sido compañeros imprescindibles para que este trabajo haya llegado hasta aquí.

gobardi 2001: 581-582). En cualquier caso, entre las lenguas hay diferencias sincrónicas en lo que a presencia del artículo e interpretación semántica se refiere. Nos parece interesante comprobar qué encontramos en vasco. Más adelante (§3) traeremos algunas discusiones entre vascólogos, con el fin de observar de qué manera les afecta nuestra aportación.[3]

1.2. Un corpus del año 1895

Para reunir datos sobre el artículo, es necesario seleccionar un corpus. Nosotros hemos elegido la colección de textos recogida por Edouard Bourciez. Antes de describirla con precisión, hablemos un poco de los trabajos de este tipo.

1.2.1. Las recogidas sistemáticas del siglo XIX

La política lingüística de Napoleón estimó necesario recabar información sobre las lenguas y *patois* que se hablaban en el Imperio Francés; saber cuál era su extensión geográfica y número de hablantes (Oyharçabal 1992a: 351). Los primeros en desarrollar dicha labor fueron Coquebert de Montbret padre e hijo, durante el primer cuarto del siglo XIX aproximadamente. Por encargo del gobierno, se pidió que la Parábola del Hijo Pródigo (y otros textos, cf. Oyharçabal 1992b) fuera traducida a las lenguas y hablas de cada lugar del Imperio.

Las desavenencias políticas evitaron que el trabajo fuera concluido debidamente (Oyharçabal 1994, 1995; Simoni-Aurembou 1989), pero el método perduró en la dialectología francesa, pues posteriormente también será utilizado.

1.2.2. El trabajo de Bourciez

Debemos enmarcar la recopilación realizada por Edouard Bourciez dentro de esa corriente dialectológico-sociolingüística. La realizó a finales del siglo XIX; allá por los años 1984-85. En el trabajo *Récueil des idiommes de la Région Gasconne* recopiló, de nuevo mediante las traducciones de la Parábola del Hijo Pródigo, información sobre las hablas de Gascuña. Logró reunir más de 4.000 traslaciones y fueron los maestros de cada pueblo los que asumieron el encargo de realizar las mencionadas traducciones al habla local (Videgain 2005). En la geografía del vasco se recopilaron 150 textos que fueron publicados recientemente por Aurrekoetxea & Videgain (2004).[4]

Por lo tanto, ¿con qué contamos para extraer información sobre el artículo? con textos de unos 150 pueblos, traducciones de la Parábola del Hijo Pródigo realizadas por gente alfabetizada, a finales del siglo XIX. Esto puede ser una ventaja, teniendo en cuenta que el estatus de los traductores es similar, se tradujeron todos en la misma

---

[3] En vasco el artículo es pospositivo y se inserta al final del sintagma. Haciendo una simplificación, tal vez demasiado osada (pero consciente y necesaria) para el vascoparlante y sobre todo para el vascólogo, podríamos decir que *txakur-Ø* es 'perro', *txakurr-**a*** 'el perro' y *txakurr-**ak*** '(los) perros'.

[4] Por lo tanto, nuestro corpus se circunscribe exclusivamente a la zona de habla vasca al norte de los Pirineos, denominada en euskara *Iparralde*, término que emplearemos aquí. Iparralde se subdivide a su vez, en tres territorios, que son de oeste a este, Lapurdi, Baja Navarra y Zuberoa.

época y porque tuvieron todos el mismo texto en francés como modelo. Eso nos permite hacer búsquedas sistemáticas con un grado de coherencia alto. De todos modos, en contra de este corpus está la brevedad de los textos, que nos impide poder encontrar todos los ejemplos que quisiéramos. El nivel cultural de los traductores nos hace pensar en que la variedad lingüística recogida pueda distar del habla estrictamente local (al haber introducido cultismos o estructuras no autóctonas). En algunas localidades han sido recogidas dos traducciones y las diferencias entre ámbas son notables. En el texto de Izura por ejemplo, hay diferencias en el léxico, grafía y morfología (Aurrekoetxea & Videgain 2004: 145-146):

(2) a) *Ordou icit içan naïn nihaoun buriaz yabé eta oukhan déçaan sos**a**.*
b) *Ordu da izan nadin ene naüsi eta ukhan dezadan dihuru**Ø**.*

Por mentar algunas cosas, aunque tal vez salten a la vista, (2a) emplea <ou> para el sonido /u/, mientras (2b) opta por <u>. Para 'dinero', (2a) tiene *sos* y (2b) *dihuru*. Además, (2a) emplea el artículo, la forma determinada (*sos**a***), pero es la forma desnuda la que aparece en (2b): *dihuru**Ø*** (vs. *dihuru**a***).

También puede resultar un problema el haber traducido demasiado fielmente el texto original, por ejemplo, al haber mantenido el orden de palabras del francés, en lugar del del vasco. Reproducimos a continuación un ejemplo de Anhauze (Aurrekoetxea & Videgain 2004: 175):

(3) a) *Zenbait egunen burian tzarra seme zen yuan herritic eguinez fierrain eta gabe erran adioric nehori.*

Tenemos *tzarra seme* con el adjetivo antepuesto, en lugar del esperable *seme tzarra*; *eguinez fierrain* con el verbo antepuesto al objeto, en lugar de *fierrain eguinez* y *gabe erran* en lugar de *erran gabe*, donde la posposición *gabe* ha sido misteriosamente transformada en preposición.

De cualquier modo y aunque el corpus se limite exclusivamente a Iparralde, el de Bourciez es un tesoro que hay que explotar, puesto que son escasas, en el caso del vasco al menos, estas recopilaciones sistemáticas: nos viene a la mente, aparte del de Bourciez, el trabajo de Sacaze. Si bien la edición moderna de esa recopilación realizada en 1887 verá la luz en breve gracias el grupo de investigación EUDIA, J. Allières ya dio a conocer los datos sobre el vasco de dicha recolecta (Sacaze 1887, Alliéres 1960-1961 y esta página web: http://www.garae.fr/spip.php?article191). Regresando al corpus de Bourciez, tenemos que decir que como la edición de Aurrekoetxea y Videgain nos ofrece los textos en soporte digital y transcritos, la realización de búsquedas y bases de datos ha resultado una labor más amena.

1.3. ¿Qué es VDM?

Visual Dialectometry (*VDM*) es un programa informático desarrollado por el informático E. Haimerl, bajo la supervisión del lingüista austríaco H. Goebl, para cuya utilización el grupo de investigación EUDIA posee una licencia. La labor del programa es transformar corpus en mapas:

(4) CORPUS → Datos estadísticos → MAPA

Tras introducir el corpus en tablas y lematizarlo, el programa ofrece la posibilidad de realizar numerosas operaciones estadísticas, cruzando datos de tantas tablas como queramos, para medir la variación dialectal conforme a diversos parámetros. En el último paso tenemos la opción de mostrar los datos en mapas cartográficos. Hemos transcrito el segundo paso de (4) en minúsculas, porque es optativo: en nuestro caso al menos, hemos omitido los datos estadísticos para observar directamente el mapa.

Dejando a un lado las enormes posibilidades que el programa ofrece, nosotros hemos hecho una simple pero productiva aplicación, trasformando directamente el corpus en mapas; es decir, hemos usado *VDM* para transformar los datos extraídos pueblo a pueblo del corpus de Bourciez en mapas coropletas de Volnoi. De esa manera, podemos saber con un simple vistazo si de pueblo a pueblo hay coherencia geolingüística en el empleo del artículo. Hemos realizado un mapa por cada frase que hemos elegido, siguiendo el procedimiento que a continuación detallamos.

2. Procedimiento y metodología de trabajo

2.1. La explotación del corpus

Es un trabajo manual. En la edición de Bourciez de Aurrekoetxea & Videgain (2004) los textos están transcritos y en formato digital, pero se han transcrito en su grafía original. Esto nos impide poder hacer búsquedas automáticas con un buscador y hay que mirar los textos uno a uno, ya que como se muestra en (2), la grafía puede variar enormemente, incluso dentro del mismo pueblo.

Para comenzar, elegimos textos de localidades distantes entre sí (unos 5 textos al azar) y buscamos oraciones con sintagmas desnudos. Una vez hallados, observamos cómo se había traducido ese sintagma o esa oración en todos los pueblos. Así pues, por cada frase que incluía un sintagma nominal desnudo interesante, creamos una tabla en *Access*.

Tabla 1

Ejemplo de una tabla de la base de datos

Datubase5-gasna			
_KarteNr	_OrtNr	Testu hitza	Lema
217	1193	etchia bethiaüc muthilez çoïnec baïtute ogia eta arnua, arrolziac eta gasna	2
217	1198	bethia da sehis, soïnec baïtute oguia eta arnoa, arroltciak eta gasna	2
217	1201	séhis béthia da, çoïgnek baïtouté ogui éta arno, arroltcé eta gasna // 1202: sehiz bethea da, oghi, arno, arrautze eta gasna baïtouté ogui éta arno, arroltcé eta gasna	1
217	1203	bethea da sehiez doutenac oguia eta arnoa arroltceac eta gasna	2

He aquí el fragmento de la tabla correspondiente a una oración del corpus. En la primera columna aparece el código que le hemos dado a la oración que comprobaremos puelo a pueblo, así como al mapa que extraeremos después (217). En la siguiente vemos el código que corresponde a cada pueblo (el 1193 por ejemplo, pertenece a Lehuntze, etc.). En tercer lugar, tenemos el texto extraído del corpus: el tener el texto en formato electrónico, en CD, facilita enormemente este trabajo, pues no hay más que copiar y pegar. Finalmente, hemos realizado una lematización.

Tal y como se puede inferir de la tabla 1, la lematización la hemos hecho con números. Corresponde el número 1 a las frases con SN desnudos (*ogiØ*, *arnoØ*, *arraultzeØ*) y el número 2 a los SD con artículo (*ogi**a**, arno**a**, arraultze**ak***). El número de lemas lo podemos incrementar a placer, añadiendo más números. Nosotros, por ejemplo, si en un pueblo no hemos encontrado datos sobre el empleo del artículo en la oración que estamos trabajando (ya sea por que la frase tiene otra estructura, por que es una traducción más libre, etc.), le asignamos el lema 0. Cada lema se representará con un color diferente después en el mapa.

Es así como al final tendremos una tabla por cada oración que hayamos decidido investigar, que reúna los datos de todos lo pueblos y los clasifique por lemas.

2.2. Los datos de Access a VDM

Aunque resulte sencillo realizar búsquedas concretas en las tablas de la base de datos, es decir, aunque es fácil consultar los datos de una oración concreta en un pueblo concreto (supongamos que queremos conocer el dato del pueblo de Mugerre en la frase 217: con mirar en la tabla 1, obtendríamos la respuesta), extraer conclusiones generales es más difícil, debido al volumen de datos que contiene cada tabla.

VDM se alimenta de estas tablas de *Access* y las muestra en mapas. Tomemos como ejemplo la tabla 1: obtendríamos el mapa referente a la oración 217 mediante *VDM*; esto es, el mapa llamado 217. El programa une el código correspondiente a cada localidad con el número de lema que le hayamos dado, asignando a cada lema un color. Por lo tanto, el pueblo de Lehuntze, cuyo número en el mapa es 1193, aparecerá en rojo mientras que el pueblo cuyo código es 1203 (Azkaine) aparecerá en azul, ya que éste tiene el lema 2, en tanto que a Lehuntze le corresponde el 1. Esto nos ofrece un mapa con tantos colores como queramos. De manera muy gráfica y con un simple vistazo, podremos saber, como mostraremos más tarde, si el empleo de SN desnudos o de SD presenta alguna coherencia geográfica.

2.3. Limitaciones del método y de VDM

Hemos descrito ya en §1.2.2 los problemas intrínsecos del corpus. Aparte de eso, hemos creído conveniente traer aquí, ya que estamos hablando sobre tecnologías para la variación lingüística, las limitaciones que nos plantea tanto el método como *VDM*, ante la esperanza de que algún día estén en situación de ser superadas.

Este método nos obliga a trabajar en un sistema binario; es decir, a presuponer que en una localidad obtendremos un sola respuesta y a asumir que esta será *a* o *b*, blanco o negro. Teniendo en cuenta las características del corpus, así ha sido en la mayoría de los casos, puesto que tenemos un sólo texto por pueblo y por tanto, un sólo ejemplo de la oración que buscamos. Pero hemos dicho ya al hablar del corpus, que en algunas localidades tenemos dos textos (§1.2.2). Cuando ambos textos dan la misma respuesta, no hay problema: si por ejemplo, en los dos textos, en la oración que estamos estudiando, nos aparece el SN desnudo, sin artículo, no cabe duda de que el lema que asignaremos a esa localidad será 1; por el contrario, si en ambos textos tenemos sintagmas determinados (SD) en esa frase, le asignaremos el lema 2 a esa localidad. Pero la casuística es más amplia. Tomemos un ejemplo:

(5) a) 1368 (Donapaule): *Oguen**Ø** dut.* // 1269 (Donapaule): *Oguen**Ø** dut.*
b) 1315 (Baigorri): *Hoben handi**a** izan nîn (nièn).* // 1316 (Baigorri): *Hoben handi**a** izan nien.*
c) 1311 (Iholdi): *Oguén handi**a** oukhan nicin.* // 1312 (Iholdi): *Oben haündi**Ø** ukhan nuen.*

En los tres pueblos contamos con dos textos; dos respuestas para un solo pueblo. Evidentemente, (5a) y (5b) no son problemáticos, puesto que los resultados de los dos textos concuerdan; en (5a) ámbos son indeterminados, esto es, les corresponde el lema 1, y en (5b) ámbos son determinados, lematizados con 2. El conflicto viene con (5c), ya que los textos difieren: mientras en uno tenemos un SN desnudo, el otro tiene un SD. *VDM* acepta un sólo resultado por localidad, es decir, sólo le puede asignar un color a una localidad. La solución es la proliferación de más lemas: podemos crear tantos lemas como queramos, pero tan sólo podemos asignar uno por localidad. Si 1 es para los indeterminados (SN), 2 para los determinados (SD) y 0 para la falta de datos, 3 será para los datos divergentes: lo emplearemos cuando en localidades con dos textos, los datos de uno y de otro difieran. Esto solventa el problema parcialmente. Visualmente el problema está resuelto, puesto que el que vea el mapa observa claramente esa distribución mediante colores diferentes, pero *VDM*, a la hora de obtener estadísticas, tratará los lemas 1, 2 y 3 como respuestas diferentes, y eso no es estrictamente así: el lema 3, asignado a los datos divergentes, no es una respuesta diferente, sino aquella que toma en cuenta las anteriores dos (con el lema 1 y 2); es decir, el lema 3 no es *c*, sino *a+b*. Pero, como decíamos, para *VDM* es estadísticamente igual la ausencia de datos (lema 0), el tener una sola respuesta (lemas 1 y 2), tener ambas respuestas (lema 3) o cualquier otro tipo de respuesta (lemas 4, 5, 6, etc.). Es una característica del método cuantitativo: lo que cualitativamente es diferente, cualitativamente se mide igual.

Otro problema del sistema cualitativo binario es la imposibilidad de juntar resultados cruzados. Expliquémonos. Para investigar el nivel de utilización del artículo (núcleo D del sintagma), hemos cogido como referencia oraciones de un corpus. Hemos hecho un mapa de cada oración. ¿Cómo comparar entre sí todos esos mapas? No es difícil crear en la base de datos una tabla que una las respuestas de todos los mapas:

Tabla 2

Los resultados de todos los mapas por frases

_OrtNr	KARTIEL	DIRU	GASNA	OGEN	AHATE	NAGUSI	ERO
1193	2	2	2	2	7	1	2
1194	1	2	2	2	2	1	2
1196	0	0	1	2	8	1	0
1198	1	1	2	6	1	1	1
1201	2	1	1	2	3	1	2
1203	0	2	2	6	2	1	0
1204	2	2	2	0	2	1	0
1205	2	1	1	2	7	1	2
1206	1	1	1	2	7	4	0
1207	2	2	2	1	0	1	0
1208	0	0	2	0	0	0	0
1209	2	2	2	2	2	1	2
1210	2	2	2	2	8	1	2

En primer lugar tenemos el código de cada pueblo y después los datos de siete tablas diferentes (*Kartiel, Diru, Gasna, Ahate, Nagusi, Ero* y *Ogen*). Cada una de las siete corresponde a una oración y su mapa. En esta tabla hemos juntado todas y podemos ver qué ha respondido cada localidad sobre cada frase. El pueblo 1194, Mugerre, nos ha dado los siguientes resultados:

(6) a) KARTIEL: *Behar da (...) ikus deçadan herri**Ø**.*
b) DIRU: *Dembora da (...) içan deçadan diru**a**.*
c) GASNA: *bethea da sehiz çoinec baituté oguia eta arnoa, arroltzeac eta gasna.*
d) OGEN: *Hutz handia eguin nuen.*
e) AHATE: *Hartcen ahalco ditutçué éré oilarrac, ahateac eta ekartcen aratché on bat hiltçeko.*
f) NAGUSI: *Dembora da içan nadin éné nausi**Ø**.*
g) ERO: *Soroa çare.*

Tal y como muestran los datos de la tabla 2 y de (6), mientas unas oraciones aparecen con el determinante (*Diru, Gasna, Ahate, Ero, Ogen*), otras presentan un SN desnudo (*Kartiel, Nagusi*). En otras localidades, por supuesto, los resultados serán otros.

¿Cómo hacer un mapa que tenga en cuenta los siete mapas? En la tabla 2 tenemos 7 lemas por localidad y recordemos que *VDM* sólo acepta un lema por pueblo. Las combinaciones posibles entre esos 7 lemas son demasiadas (no hay más que observar la tabla 2, para percatarse de que en ningún pueblo se da la misma combinación de lemas): asignar un lema (un color en el mapa) a cada combinación posible, haría el mapa ilegible; incoherente y lleno de colores.

La solución a este problema parte de fijarse, más que en la cantidad, en la cualidad. Todos los SN desnudos no son iguales; tampoco los SD. La presencia del artículo cambia la semántica y los SN desnudos son sólo posibles en contextos sintácticos concretos. La labor del lingüista es clasificar los SN y los SD recopilados en el corpus, en función de su interpretación semántica y sintáctica. Tal y como mostraremos luego, realizar comparaciones cruzadas entre mapas será más sencillo tras hacer la clasificación.

Para darse cuenta de la importancia de la función sintáctica y semántica, veamos de manera breve lo dicho por algunos lingüistas sobre el empleo del artículo en vasco. Aprovecharemos la síntesis de sus teorías para presentar los mapas correspondientes a cada oración, mostrando si avalan o contradicen esas teorías.

3. Lo dicho por los lingüistas y nuestros mapas

Ya hemos añadido (§1.1) que se han escrito unas cuantas líneas sobre el artículo en vasco. Los sincronistas han hecho un gran esfuerzo para fijar el contexto de empleo y aparición del artículo y los diacronistas han perseguido el objetivo de saber desde cuándo se encuentra en vasco y de qué manera. Esta cita de Michelena servirá para introducirnos en el meollo del asunto:

> Es un lugar común de la lingüística histórica vasca la afirmación de que el artículo determinado es entre nosotros de introducción relativamente reciente, como lo es en las lenguas románicas o germánicas. Pero aquí, lo mismo que mucho antes en griego, su aparición es un hecho documentado, mientras que en vasco se trata de una presunción (...) referente a la prehistoria de la lengua. Lo que sí es un hecho plenamente histórico es su proliferación: durante los últimos siglos ha ido y va ganando posiciones en las cuales su empleo tiende a hacerse obligatorio. (Michelena 1978: 208)

Por lo tanto, es imaginable que en la evolución, partiendo de la ausencia absoluta del artículo al paradigma actual, el vasco ha conocido diferentes fases cronológicas como geográficas; esto es, no hallaremos el mismo empleo del artículo en un lugar y época, o en otra. Asimismo, aparecerá en unas estructuras sintácticas antes que en otras. Mientras los diacronistas han tratado de concretar esa cronología, los sincronistas persiguen fijar la distribución sintáctica y geográfica.

Los que han hecho frente al problema, sea desde el punto de vista sincrónico como desde el diacrónico, han dado por buena la separación de los sintagmas en dos grupos, dependiendo de su función. Por un lado estarían los predicados, que cumplen una función atributiba, con un verbo copulativo. Por el otro, tendríamos sintagmas que satisfacen un argumento, puesto que un verbo les asigna un rol temático.

3.1. Sobre los sintagmas con función atributiva

En el trabajo Azkarate & Altuna (2001) se nos dice que los atributos que tengan un verbo auxiliar de raíz *izan* 'ser', serán indeterminados si son adjetivos simples; si tienen una estructura de *nombre + adjetivo*, aparecerán definidos con el artículo (Azkarate & Altuna 2001: 74-75). De todos modos, en las hablas del oeste (Bizkaia, Gipuzkoa, etc.), también los adjetivos simples aparecerán principalmente determinados. En las hablas del este, por tanto, la presencia de un nombre obligaría a la presencia del artículo. Sirvan como ejemplo estas frases de nuestra cosecha:

(7) a) *Mikel aberatsØ da.* 'Mikel es rico' (Adjetivo simple, que en el oeste aparecería también con artículo)
b) *Mikel gizon aberatsa da.* 'Mikel es un hombre rico' (estructura *nombre + adjetivo*)

La unión entre la presencia del nombre y el artículo fue expresada ya por Michelena (1978: 213): "...el adjetivo, como tal, queda indeterminado, mientras que la determinación es normal cuando el predicado nominal está formado por un sintagma sustantivo + adjetivo" y un poco más adelante: "...permitiría (...) atribuir la determinación a los sustantivos". Zabala, por el contario, da a los nombres o adjetivos simples la opción de ir indeterminados (desnudos), y añade que sólo las estructuras *nombre + adjetivo* tienen obligación de portar artículo (Zabala 2001: 330).

La semántica del predicado también tiene su importancia, como ya han destacado otros trabajos. Zabala mismo nos muestra que lo que Txillardegi o Lafitte definieron como *transitivo / permanente* y que ella define como *individual / stage level*, tiene mucho que decir sobre la presencia o ausencia del artículo (Zabala 2001: 329). La cualidad de los predicados *individual level* es perpetua o intransitoria y en el caso de los *stage level* esa cualidad tiene una limitación temporal; no es para siempre:

(8) a) *Mikel irakaslea da.* 'Mikel es profesor' *Individual level*
b) *Mikel irakasleØ da(go).* 'Mikel está de profesor' *Stage level*

Como se puede observar, los predicados de tipo *individual level* (8a) se forman con el auxiliar *izan* 'ser'; los de tipo *stage level*, por el contrario, pueden formarse en vasco con los auxiliares *izan* 'ser' o *egon* 'estar': la elección de uno u otro depende del dialecto. Los dialectos que emplean *izan* para ambos casos (8 a y b), hacen la diferenciación semántica empleando la ausencia o presencia del artículo. Eguren (s. d.) también utilizará esta diferenciación para demostrar que el morfema *-a* no es el artículo, sino el núcleo de un sintagma predicativo. En todo caso, no es éste el lugar para profundizar en la hipótesis de Eguren.

Resumiendo; la necesidad de un SD o de un SN desnudo se ha asociado a la interpretación semántica, pero en algunas hablas, en palabras de Zabala en las del este (Zabala 2001: 330), en esas mismas que permiten que los sintagmas de interpretación *individual level* puedan ser SN desnudos, la presencia del determinante artículo iría asociada a la presencia del nombre.

Vista la distribución sincrónica, fijémonos en la descripción realizada desde el punto de vista diacrónico. Manterola nos dice claramente que en su opinión, en la fase antigua del vasco no había artículo (cf. la cita de Michelena §3) y que por tanto,

a día de hoy los dialectos del este son los que mayor vínculo conservan con dicha fase antigua (Manterola 2008: 1).

En los trabajos más recientes, como el de Manterola, se han hecho matizaciones a la distribución anterior. Zabala nos decía que en las hablas del este había sintagmas desnudos de carácter *individual level*; Manterola nos confirma que, además de con nombres, con adjetivos simples a veces tampoco son posibles los SN desnudos. Los ejemplos del dialecto bajonavarro son suyos (Manterola 2008: 7).

(9) a) *Jestu hori / jestu horren*[5] *egitea ez da pollitØ.* 'Hacer ese gesto no es bonito'
b) **Ume hori ez da politØ vs. Ume hori ez da pollita.* 'Ese niño no es bonito'

Ambos sintagmas (9a, b) son adjetivos de tipo *individual level*, pero el segundo necesita del artículo en bajonavarro (la presencia del nombre no condiciona nada en este caso, ya que no lo hay). Para justificar esto, Manterola nos recuerda el trabajo de Milsark, que, tras aplicarlo al vasco, le lleva a pensar en una posible relación entre el sujeto y el predicado de la oración: en los ejemplos de (9), alguna propiedad del sujeto sería la que condicione que mientras en (9a) el SN desnudo es posible, no lo sea en (9b). ¿Cuál es, pues, el motivo de que en bajonavarro haya entrado el artículo antes en oraciones del tipo (9b)? que el sujeto deíctico de (9b) *ume hori* 'ese niño' tendría una "carga referencial" mayor: sería más tangible que el de (9a). El niño tendría más propiedades de sujeto, de agente, que el gesto (Manterola 2008: 8-9).[6]

En nuestro corpus tenemos algún ejemplo de predicado atributivo de raíz *izan* 'ser'.

En el mapa de la página siguiente tenemos los resultados de la traducción de la oración francesa *Etes-vous fou?*[7] He aquí sendos ejemplos de las posibles traducciones que hemos recogido en el mapa:

(10) a) *EroØ zara?*[8] – determinado
b) *Ero**a** zara?* + determinado
c) *Ero**tua** zara?* + determinado: participio
d) *Ero**turik** zara?* + determinado: partitivo

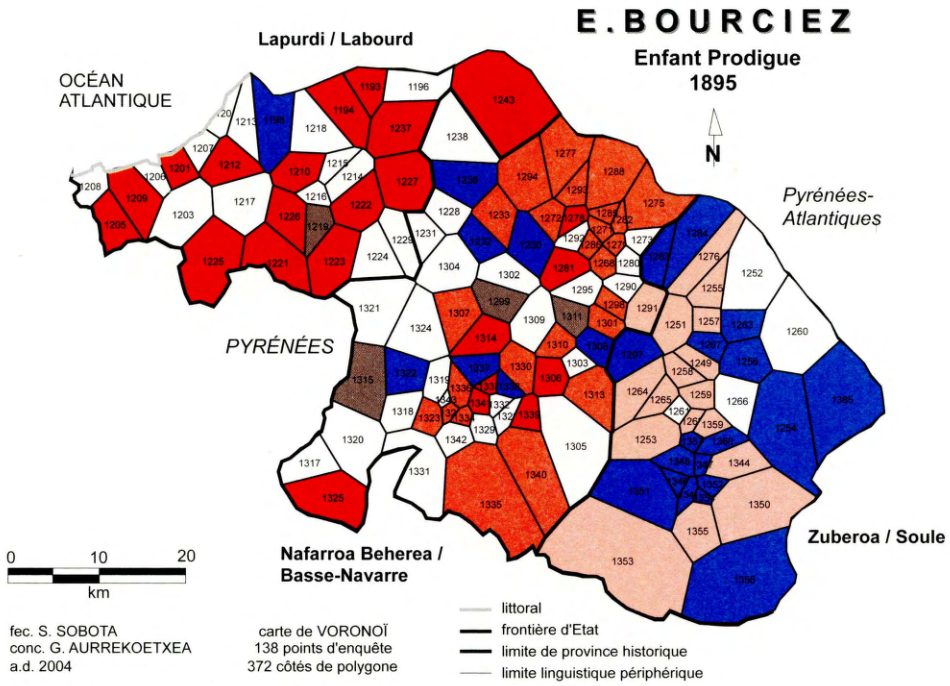
---

[5] En las hablas del este el objeto suele aparecer con caso genitivo (*gestu horREN*) en lugar de absolutivo (*gestu horI*), cuando el verbo está nominalizado (*egitea* 'hacer'). Es por eso que Manterola da las dos formas, pero en principio no tiene que ver con el tema que tratamos.

[6] El propio Manterola reconoce que la argumentación para clasificar un sintagma como fuerte o débil conforme a su carga referencial, no está todavía demasiado clara. Alguien podría pensar que, en lugar de tener en cuenta la carga referencial, es la estructura sintáctica la que condiciona la presencia del artículo, ya que mientras en (9a) el sujeto es una oración (tiene verbo), en (9b) nos encontramos con un sintagma determinante cumpliendo dicha función. En todo caso, y reiterando que ésta es una hipótesis no contrastada, Manterola sospecha que una oración como **Jestu hori ez da pollit* tampoco sería posible a pesar de contar con un SD como sujeto, puesto que éste sería un sintagma débil; de todos modos, este ejemplo que damos en la nota al pie, a diferencia de los de (9), no ha podido ser escuchado por Manterola a hablantes en conversación natural, por lo tanto, es un asunto que queda por dilucidar.

[7] En Aurrekoetxea & Videgain (2004: 421) viene *Etes-vous devenu fou?* en la transcripción, pero es una errata. Cf. La versión corregida del texto en Santazilia (2007).

[8] Los ejemplos correspondientes a los mapas los hemos adaptado y estandarizado dejando de lado variantes dialectales, para mostrar de manera más clara el asunto que nos concierne: el de la utilización del artículo. Si alguien quisiera ver la oración original, no tiene más que acudir a Aurrekoetxea & Videgain (2004).



Mapa 1

Predicado atributivo de raíz *izan* 'ser'

El retrato de finales del siglo XIX que nos ofrece el mapa es esclarecedor. En Lapurdi son mayoría las formas determinadas por el artículo (10b) y en Zuberoa las indeterminadas (10a); es más, en esta región no encontramos ninguna forma indeterminada. Además de eso, en este mapa tenemos dos estructuras más; las de las oraciones de (10c) y (10d). La estructura con participio (*-tu*) lleva artículo; por otro lado, si tomamos el partitivo como determinante, ambas estructuras son determinadas. En el siglo XIX las formas sin artículo (indeterminadas) aparecen con más profusión en Zuberoa. Las estructuras con partitivo y participio también muestran una distribución clara: las primeras las hallamos sólo en Zuberoa y las segundas corresponden sólo a la Baja Navarra. Estas últimas, llevarían el artículo, según Michelena, puesto que si no lo llevaran se podrían confundir con una estructura de *verbo principal + auxiliar*: *Erotu zara* 'has enloquecido' (Michelena 1978: 215).

Siguiendo la pauta de Michelena, podríamos pensar que el sintagma de (10a) es un adjetivo y el de (10b) un nombre; es decir, que mientras el primero considera *erho* 'loco' como adjetivo, el segundo lo interpreta como nombre (sustantivo). Si no, mirando al original en francés, podríamos decir que mientras en algunas traducciones le han dado una interpretación como *individual level*, en otras localidades lo han interpretado como *stage level*.[9] Pero, si miramos el mapa, otro tipo de ideas nos vienen a la cabeza, por la coherencia geográfica que muestra. Si fuera un mero caso de interpretación semántica, no podríamos diferenciar isoglosas claras. Es más económico pensar que para entonces en Lapurdi se habría extendido la tendencia que era ya habitual en las hablas al sur del Pirineo: que el artículo aparezca con todo tipo de atributos. De facto, Michelena mismo reconoce que "se diría que [el artículo] es empleado más profusamente en el centro y oeste del país que en el este, más en general al sur que al norte" (Michelena 1978: 208). Nos parece lo más apropiado pensar que la oración ha sido en todos los lugares interpretada como *stage level*, puesto que las formas con participio (*erotua*) de Baja Navarra y las formas con partitivo (*eroturik*) no admiten más que una interpretación transitoria (*stage*). Siguiendo la opinión de Manterola, podríamos justificar las pocas formas determinadas por el artículo que encontramos en Baja Navarra, aduciendo a una posible interpretación *individual level* por parte de los traductores, en el que se han visto obligados a emplear el artículo, al tener un sujeto con carga referencial grande (*strong*).

De todos modos, Michelena admite que el empleo de ciertos predicados indeterminados es común a la totalidad del vasco (Michelena 1978: 210). Tenemos un ejemplo (en la página siguiente, Mapa 2) de esto en nuestro corpus.

Son las siguientes, las oraciones representadas aquí:

(11) *Il est temps que je sois mon maître.*

a) *Denbora da izan nadin nere buruaren nagusi**Ø**.* – determinado
b) *Denbora da izan nadin ene nagusi**a**.* + determinado
c) *Ordu da ene buru**az** nagusi**Ø** izan nadin.* instrumental

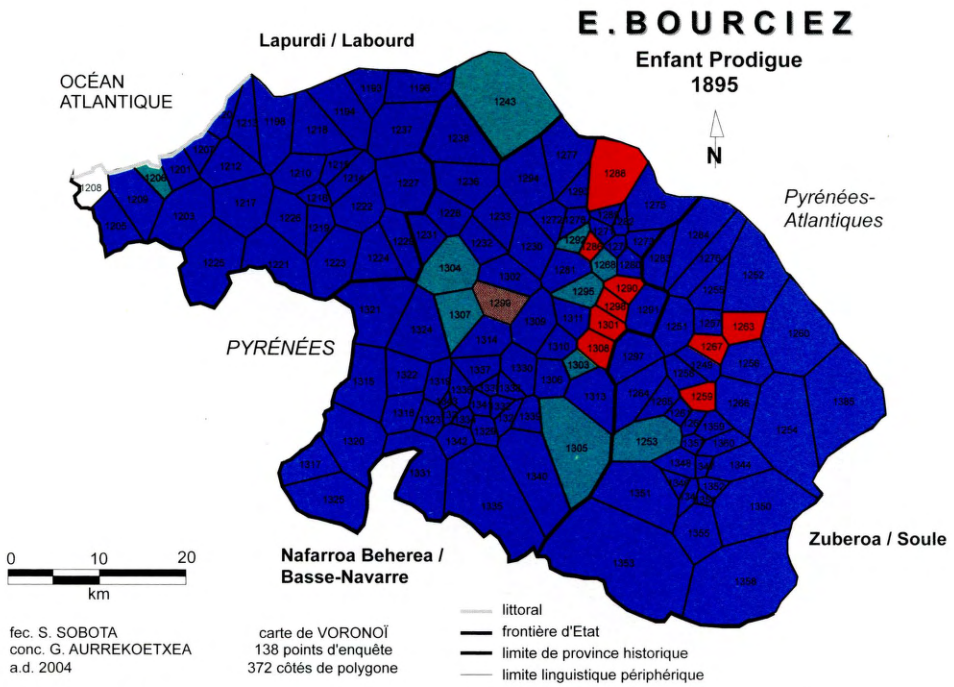
¿Por qué es en este caso general el empleo de la forma desnuda, también en las hablas del sur del Pirineo? A nuestro parecer en este tipo de oración sólo es posible una interpretación *individual level*, por tanto, sería lo habitual que apareciera el artículo. Es más, si asumimos que la presencia de un nombre condiciona la aparición del artículo, podría pensarse que *nagusi* 'dueño' es un adjetivo, pero al aparecer con un genitivo (*ene buruaREN* 'de mí mismo'), queda asegurado su carácter de nombre, puesto que los genitivos sólo pueden aparecer con éstos.

Podríamos tomarlo como una estructura lexicalizada, si tenemos en cuenta que en castellano también es posible esa estructura sin artículo:

(12) a) *Soy **Ø** dueño de mis actos.* a') *Soy **el** dueño de mis actos.*

Podría proponerse que tanto en castellano como en vasco, la ausencia de artículo sugiriera un cambio semántico progresivo hacia un *stage level*, que transformaría la cualidad de propiedad en algo transitivo, a pesar de formarse con el verbo *izan* 'ser'

[9] Recordemos que en francés ambas interpretaciones se realizan con el verbo *être* 'ser/estar'.



Mapa 2

Ejemplo de profusión de indeterminados

(y no con *egon* 'estar'). No seremos nosotros quienes aportemos la solución definitiva al asunto, pero nos parece que puede abrirse una vía de investigación en torno a la estructura semántica de *nagusi* 'dueño'. Diríamos que es un nombre con estructura argumental y por tanto, tanto los sintagmas con genitivo como los de instrumental son necesarios para que la frase satisfaga dicha estructura argumental. Haremos una anotación a modo de anexo, para marcar la importancia de esos complementos en genitivo e instrumental: los pueblos que han utilizado formas determinadas (9 pueblos) no han utilizado como complemento una forma reflexiva (*nere burua* 'mí mismo'), sino un pronombre simple (*nere* 'mi'); pero sí hay quienes aunque no hayan usado la forma reflexiva, han empleado la forma indeterminada, sin artículo:

(13) a) *Denbora da izan nadin* ***ene/neure*** *nagusi***Ø**.
a') *Denbora da izan nadin* ***ene/neure*** *nagusia.*

b) *Denbora da izan nadin* ***ene buruaren/neure buruaren*** *nagusi***Ø**.
b') ***Denbora da izan nadin* ***ene buruaren/neure buruaren*** *nagusia.*

Sin duda, este asunto merecería un análisis más profundo.

3.2. Sobre los sintagmas de función complemento

Las demás oraciones extraídas del corpus y sus mapas corresponden a sintagmas de función complemento, es decir, son argumentos de un verbo y cumplen la función de complemento directo. Hemos elegido ejemplos de esa función sintáctica, puesto que no hemos encontrado en el corpus sintagmas argumentales con otra función, construidos mediante SN desnudos. En el vasco estándar de hoy es imposible encontrar sintagmas argumentales sin artículo (Artiagoitia 2004: 27-28, Zabala 2001: 328), pero en el vasco histórico mismo y en otras lenguas tenemos ejemplos: fijémonos en la oración inglesa *Dogs are dangerous*, donde el sujeto *Dogs* no lleva artículo; o veamos estos ejemplos extraídos por Michelena del texto de Axular, del siglo XVII (Michelena 1978: 211-212):

(14) a) *Aita saindu***Ø**-*ri*. Complemento indirecto, con marca de dativo (-*ri*) y sin artículo (Ø).
b) *Ainguiru***Ø**-*c*. Sujeto, con marca de ergativo (-*c*) y sin artículo (Ø).

Como decíamos, en el texto de Bourciez no hay ni rastro de SN desnudos en función de sujeto ni de objeto indirecto, pero sí que hay, aunque no sean posibles en el estándar de hoy, SN desnudos con función de objeto directo: es en éstos en los que nos fijaremos.

Asumiendo, pues, que los dialectos orientales son los más arcaizantes en este aspecto y que el artículo, además de ser tardío, se ha extendido de oeste a este (Manterola 2006: 11), tendríamos la hipotética esperanza de encontrar el mayor número de objetos desnudos en el este. Desde el punto de vista diacrónico, siguiendo a Manterola, los nombres incontables son tipológicamente los últimos en recibir el artículo; primero lo adquieren los sintagmas de interpretación plural indeterminada (Manterola 2006: 7):

(15) a) *Nik ardo***a** *erosi dut.* a') *Nik ardo***Ø** *erosi dut.* 'Yo he comprado vino'
b) *Nik arraultze***ak** *erosi ditut.* b') *Nik arraultze***Ø** *erosi dut.*[10] 'Yo he comprado huevos'

Los objetos de las primeras oraciones (15a y 15a') son conceptos de masa; incontables. Los otros dos son contables, pero tienen interpretación indeterminada, genérica: no sabemos a qué huevos hacemos referencia, ni necesitamos precisarlo. En castellano también son indeterminados estos sintagmas: *He comprado* **Ø** *huevos*. En los dialectos conservadores del este, a día de hoy todavía esperaríamos ejemplos como los de (15a') y (15b'). En el trabajo de Urtzi Etxeberria y Rikardo Etxepare, limitan la existencia de SN desnudos al dialecto suletino (de Zuberoa)

[10] Aunque no entraremos en este tema, es interesante en el caso de (15b) y (15b'), cómo la presencia o ausencia del artículo condiciona la concordancia verbal de objeto, siendo plural en (15b) (d**it**ut) y singular en (15b') (d**Ø**ut). Hay algo sobre este asunto en el trabajo Etxeberria & Etxepare (2008).

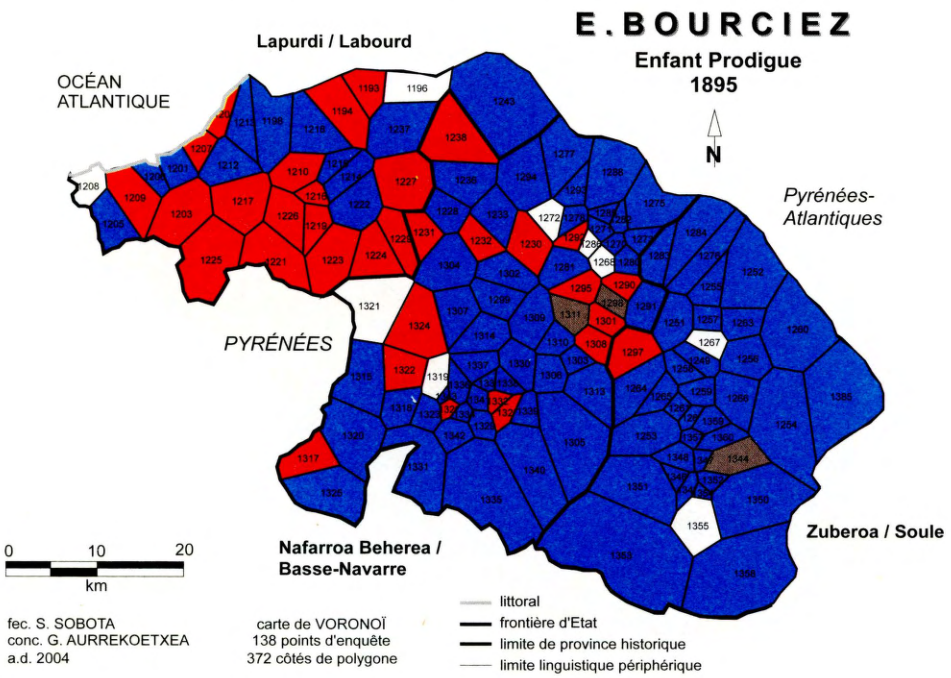
(Etxeberria & Etxepare 2008).[11] Cuando mostremos los mapas, trataremos de fijar las fronteras de ese "este" en el siglo XIX, atendiendo a la oposición recién descrita de *contable / incontable*. A su vez, dentro de los incontables separaremos los sintagmas con interpretación existencial (genérica) y específica.

3.2.1. Los sintagmas incontables

En el mapa de debajo veremos la distribución de las siguientes oraciones:

(16) *Il est temps* (...) *que j'aie de l'argent.*

a) *Denbora da* (...) *izan dezadan diru***Ø**. – determinado
b) *Denbora da* (...) *izan dezadan diru***a**. + determinado



0: sin datos [9]
1: – determinado [93]
2: + determinado [33]
3: formas divergentes [3]

Mapa 3

Concepto de masa

[11] De todos modos, el objetivo de Etxeberia & Etxepare en ese trabajo no es establecer isoglosas claras, por lo tanto, no se tome *dialecto suletino*, demasiado *sensu stricto*.

El dinero es un concepto de masa; no se puede contar (**tres dineros*). Como vemos en el mapa, el empleo de SN desnudos con conceptos de masa está bastante extendido en el corpus de Bourciez. Las formas determinadas aparecen principalmente en el territorio del labortano, aunque también tenemos ejemplos en bajonavarro. Teniendo en cuenta que Manterola y Etxepare & Etxeberria circunscriben ese fenómeno a día de hoy a los dialectos orientales exclusivamente, parece ser que el artículo habría entrado en estos contextos en época muy reciente, debido a la rápida evolución que ha tenido desde finales del siglo XIX hasta hoy, si comparamos el mapa con la distribución actual.

3.2.2. Los sintagmas contables

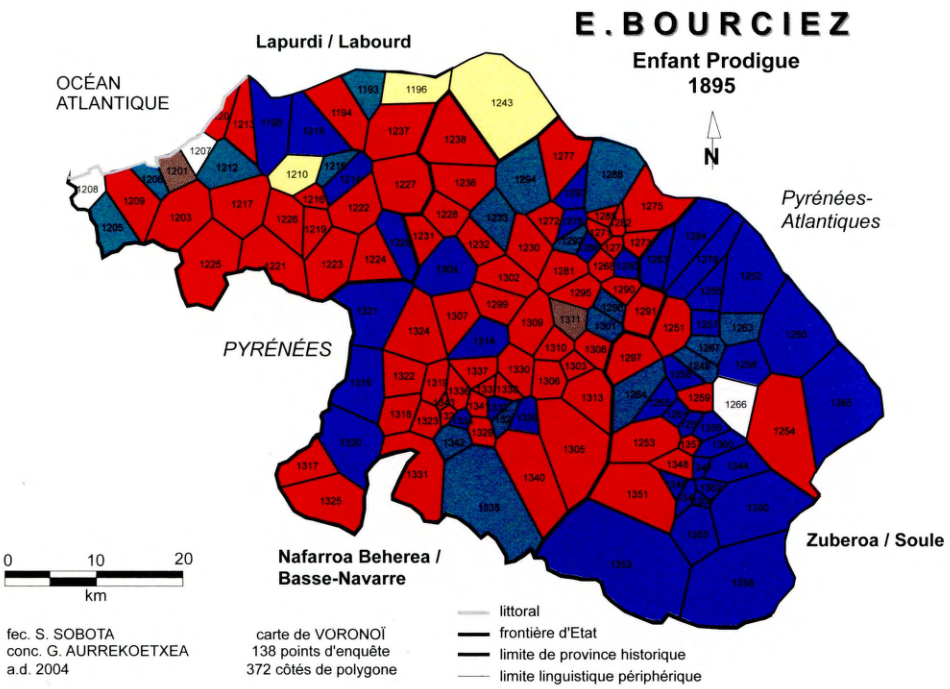
Tenemos oraciones como éstas:

(17) *Vous pourrez aussi prendre des coqs, des canards, et amener un veau (...).*

a) *Hartzen ahal duzue oilar**Ø**, ahate**Ø**, eta ekartzen txahal bat* (...).
– determinado
b) *Hartzen ahal dituzue oilarr**ak**, ahate**ak**, eta ekartzen txahal bat* (...).
+ determinado
c) *Hartzen ahal dituzue oilar eta ahate **zenbait**, eta ekartzen txahal bat* (...).
det. indefinido
d) *Hartzen ahal dituzue oilar eta ahate**etarik**, eta ekartzen txahal bat* (...).
posposición *-etarik*

Las oraciones de (17), por lo tanto, tienen sintagmas con conceptos contables, pero mediante la falta de artículo expresan el carácter genérico del sintagma; nos da igual qué pato o gallo coger. Las oraciones (17c) y (17d) son claros ejemplos de esta interpretación genérica, ya que el determinante indefinido (*zenbait* 'algunos') y la posposición *-etarik* no permiten otra interpretación.

Según muestra el mapa 4, mientras que las formas indefinidas son más abundantes en Zuberoa, encontramos las determinadas en Lapurdi y la Baja Navarra principalmente. Lo más interesante es comparar este mapa con el mapa 3: a finales del siglo XIX es bastante más común la presencia de SN desnudos con conceptos incontables que con contables, ya que con estos últimos no encontramos sintagmas indeterminados de manera general, más que en Zuberoa. Que la jerarquía de Manterola mencionada arriba (§3.2) se cumple en estos mapas, es evidente: han tomado antes el artículo los nombres contables (por eso está más extendido el uso del artículo con éstos), a pesar de tener una interpretación genérica, que los conceptos de masa.



0: sin datos [3]
1: – determinado [41]
2: + determinado [71]
3: formas divergentes [2]
4: det. indefinido [18]
5: posposición-*etarik* [3]

Mapa 4

Sintagmas contables genéricos

3.2.3. Los sintagmas contables e incontables

Nos ha parecido interesante hacer el mapa de una oración que presenta a la vez nombres contables e incontables.

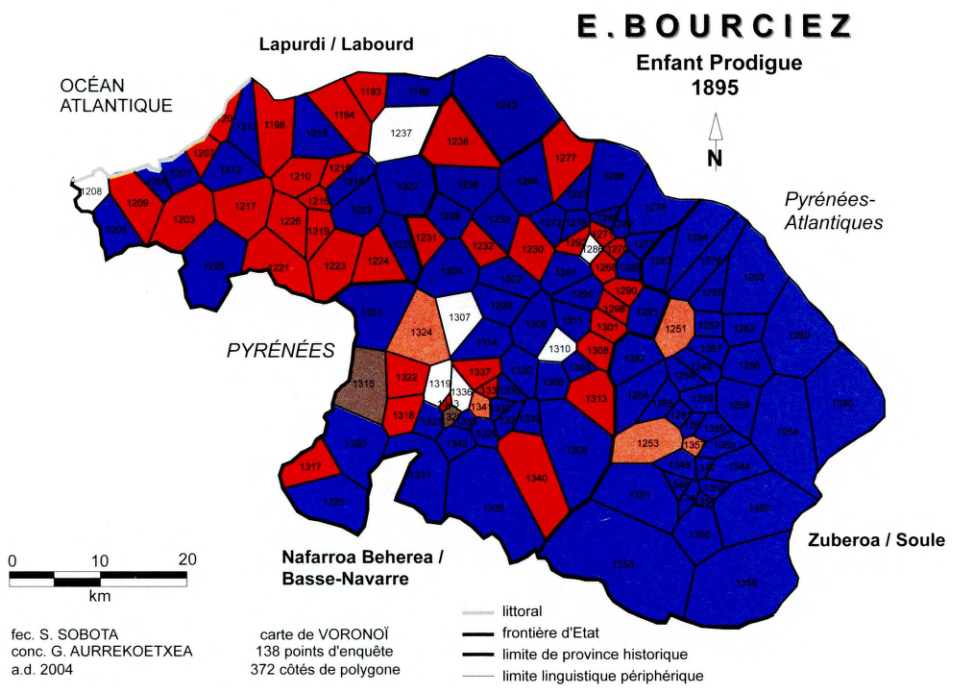
Como veremos a continuación en los ejemplos, tres son las combinaciones posibles:

(18) *[La maison de mon père] est pleine de domestiques qui ont du pain et du vin, des oeufs et du fromage.*

a) *Betea da mutilez, zeinek baitute ogi**a** eta arno**a**, arraultze**ak** eta gasna.*
+ determinado

b) *Betea da mutilez, zeinek baitute ogi**Ø** eta arno**Ø**, arraultze**Ø** eta gasna.* – determinado
c) *Betea da mutilez, zeinek baitute ogi**Ø** eta arno**Ø**, arraultze**ak** eta gasna.* contable + determinado

Y he aquí el mapa que hemos obtenido:



Mapa 5

Conceptos contables e incontables

En estos ejemplos de (18), los dos primeros sintagmas (*ogia* 'pan', *arnoa* 'vino') son incontables y el tercero (*arraultzeak* 'huevos') contable.[12] Las opciones son tres: todos

[12] Hemos dejado de lado el sintagma *gasna* 'queso', porque la *-a* final forma parte de la raíz y no cambia, sea determinado o no: *gasna* + *-a* = *gasna*; *gasna* + *-Ø* = *gasna*. Por lo tanto, no hay manera de saber si el ejemplo de la oración lleva artículo o no.

los sintagmas con artículo (18a); todos indeterminados (18b) o todos indeterminados salvo los contables (18c). En teoría podríamos esperar una cuarta opción: que aparecieran los incontables con artículo y los contables sin él, pero no hemos hallado en la práctica ningún ejemplo de eso, por lo que la oración siguiente debemos darla con asterisco:

(19) a) **Betea da mutilez, zeinek baitute ogi**a** eta arno**a**, arraultze**Ø** eta gasna.*

Pero no es casualidad que estos casos teóricamente posibles no aparezcan, y un hallazgo tal, no hace sino fortalecer la hipótesis defendida junto a Manterola en este trabajo: si los nombres incontables han adquirido el artículo diacrónicamente después de los contables, podemos diferenciar tres épocas o situaciones de la lengua:

(20) a) Los contables y los incontables son indeterminados.
b) Mientras los contables son determinados, los incontables permanecen indeterminados.
c) Tanto los contables como los incontables aparecen con el artículo.

Si quisiéramos representar esto en una tabla tetracórica (Croft 1993: 48), tendríamos algo así:

Tabla 3

Diagrama tetracórico de los sintagmas contables e incontables

	Contables e incontables	Sólo incontables
+ determinado	✓	✗
– determinado	✓	✓

Por lo tanto, esta tabla nos descubriría una relación de implicación: si una lengua tiene nombres incontables determinados, entonces tendrá también nombres contables determinados.

El mapa nos atestigua las tres situaciones descritas en (20): no puede existir, en nuestra opinión, ni un lugar ni una época en la que los sintagmas contables fueran indeterminados y los incontables determinados, tal y como confirma el mapa.

De facto, se ve perfectamente en el mapa 5 que el este es más conservador. Tanto los nombres contables como los incontables con artículo los encontraremos de manera más frecuente en el territorio del labortano y también en el del bajonavarro tenemos unos cuantos ejemplos: sin embargo, no hay ni uno en Zuberoa, donde todos son indeterminados. Conforme vayamos al oeste, nos irá apareciendo el artículo de manera progresiva.

Son pocos los ejemplos de (20b), una situación de transición entre sistemas: no son representativos geográficamente, pero sí tipológicamente, puesto que podemos estar seguros de que si en esos pueblos que están entre el sistema indeterminado completo y el determinado completo, ha sucedido algún cambio, ha sido a hacer todas las formas con el artículo, y no a potenciar las formas indeterminadas; esto es, el camino será de (20b) a (20c), nunca a (20a).

Es interesante observar el asunto desde el prisma de lo marcado/no marcado. Ese concepto de *markedness* cuya entrada en vigor puede ser atribuida a los neogramáticos, ha conocido a posteriori una infinidad de empleos imprecisos y ambiguos que han llevado a devaluar el concepto como término científico. No obstante, Andersen recupera ese valor de *markedness* como un concepto universal, para crear el *Principle of Markedness Agreement*, según el cual los elementos marcados intervendrán antes en entornos marcados, mientras que los no marcados lo harán en los no marcados (Andersen 2001a). Se considerará marcado aquello que el hablante no tenga asimilado en su gramática base, y cuya utilización conozca por reglas de empleo superficiales.

Si partimos de la hipótesis de que el artículo es de incorporación reciente en vasco, su ausencia sería, al menos históricamente, el suceso no marcado. Debemos pensar que en determinado momento se introduce el artículo en vasco, un morfema marcado y posiblemente apre(he)ndido,[13] inicialmente como cuantificador en sintagmas contables, produciéndose una alternancia sincrónica. El contar con este morfema para cuantificar sintagmas contables haría finalmente obligatoria su presencia en sintagmas contables (más fácilmente cuantificables que los incontables), considerando marcados los contextos en los que tenemos sintagmas contables sin cuantificar (es decir, sin artículo), ya que necesitarán de ese morfema marcado, cumpliendo así el *Principle of Markedness Agreement*: el morfema marcado aparecerá primero en un contexto que también lo es, para luego extenderse a los contextos no marcados (Andersen 2001: 31). El mapa mostraría esas tres fases: en Zuberoa el morfema marcado apenas habría comenzado a aparecer, salvo en unos pocos ejemplos donde lo encontramos en los contextos marcados. Conforme nos aproximamos a Lapurdi, el morfema marcado (el artículo) se habría extendido ya incluso a los contextos no marcados, que serían los de los sintagmas incontables.

En resumidas cuentas y asumiendo la cronología sobre la direccionalidad del cambio evolutivo propuesta por Andersen (Andersen 2001b), los hablantes del vasco labortano habrían comenzado a emplear el artículo para cuantificar sintagmas contables, al principio esporádicamente y en alternancia con la forma tradicional indeterminada. A medida que el uso de ese morfema marcado se va extendiendo, se considera el SN contable desnudo como contexto marcado, y se generaliza el empleo del artículo en él. A posteriori, serán los contextos no marcados los que tomarán el artículo. Y este proceso irá sucediendo de igual manera de oeste a este. Así las cosas, se producirá al final un *markedness shift* (Andersen 2001b: 238) y lo que en principio era marcado y sucedía en la gramática superficial del hablante, provocará un reanálisis y una actualización de su gramática base, pasando a ser considerado no marcado, en detrimento de la forma antigua (sin artículo) que será considerada marcada (por arcaica). Este *markedness shift* sólo se producirá cuando la nueva forma haya sido asimilada completamente por la gramática base del hablante.

El labortano presentaría así una situación más evolucionada donde el artículo sería ya un morfema no marcado con el *markedness shift* ya concluido y extendido a todos los contextos, mientras que de camino a Zuberoa encontraríamos fases más arcaicas donde el empleo del artículo no estaría aún completamente gramaticali-

[13] No entraremos aquí a discutir si se trata de un préstamo del romance o no.

zado. Esto demuestra, obviamente, que el artículo en vasco se ha ido extendiendo de oeste a este.

3.2.4. Un caso particular: el sintagma de nombre + adjetivo

El corpus siempre es escaso para el investigador; nunca le dará tantos ejemplos como sean necesarios para obtener conclusiones completas. De todos modos, en el texto de Bourciez hemos encontrado unas traducciones que pueden ser un buen tema sobre el cual hablar:

(21) *J'eus grand tort.*

a) *Ogen handi**a** ukan nuen.* + determinado
b) *Ogen handi**Ø** ukan nuen.* – determinado
c) ***Anitz** ogen ukan nuen.* con adverbio

La particularidad de este sintagma reside en la presencia del adjetivo por un lado y en las características del argumento por otro. Manterola muestra, contra la creencia de Michelena, que hay estructuras *nombre + adjetivo* predicativas indeterminadas en los textos antiguos, como el de Etxepare en el siglo XVI (Manterola 2008: 18), pero, ¿y a modo de complemento en el corpus de Bourciez?

Veamos el mapa 6 de la página siguiente.

Si tomamos el sintagma *ogen handi(a)* 'gran culpa' como complemento del verbo *ukan* 'tener', lo deberíamos tratar como concepto incontable. Aunque el adjetivo *handi* 'grande' puede aparecer con nombres contables pero no con incontables (cf. **el gas grande*), no parece apropiado algo como **tres culpas grandes*. Si es incontable, deberíamos esperar encontrarnos una utilización indeterminada amplia (§3.2.1): pero no es así. La forma determinada es hegemónica sin duda.

Hay que subrayar el carácter especial de un nombre como *ogen* 'culpa'. Por un lado, como hemos dicho ya, puede tomar el adjetivo *handi* 'grande', propio de los nombres contables, pero luego no se puede contar (**Hiru ogen handi* 'tres grandes culpas'). Por otro lado, aunque sea un argumento del verbo *ukan* 'tener', tiene una semántica predicativa grande; es fácilmente parafraseable de la siguiente manera: *ogendun handia izan nintzen* 'fui un gran culpable'. No sería pues, demasiado inverosímil tomarlo como atributo. Haciendo referencia a la limitación que hemos mencionado ya varias veces, las estructuras de *nombre + adjetivo* indeterminadas serían imposibles o muy esporádicas en los predicados ya para el siglo XIX, es decir, para la época de nuestro corpus.

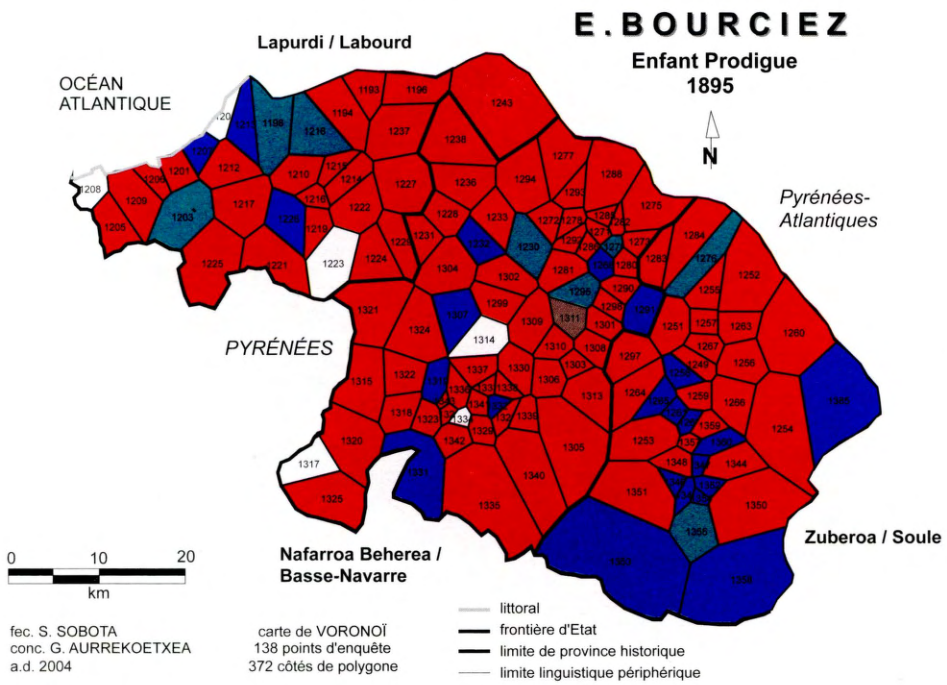
Las pocas apariciones indeterminadas que hay se podrían explicar también mirando el texto francés original. Dice *J'eus grand tort*, forma indeterminada, sin el artículo *du* ni otros. Ésta sería, también en francés, una forma fosilizada.

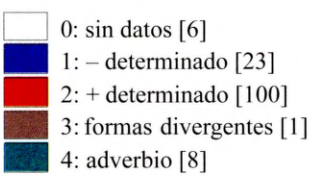
3.2.5. Un sintagma contable digno de mencionar

En las oraciones de debajo tenemos un sintagma que tomaríamos por contable:

(22) *Il faut que (...) je vois du pays.*

a) *Behar dut (...) ikusi herri**Ø**.* – determinado
b) *Behar ditut ikusi herri**ak**.* + determinado



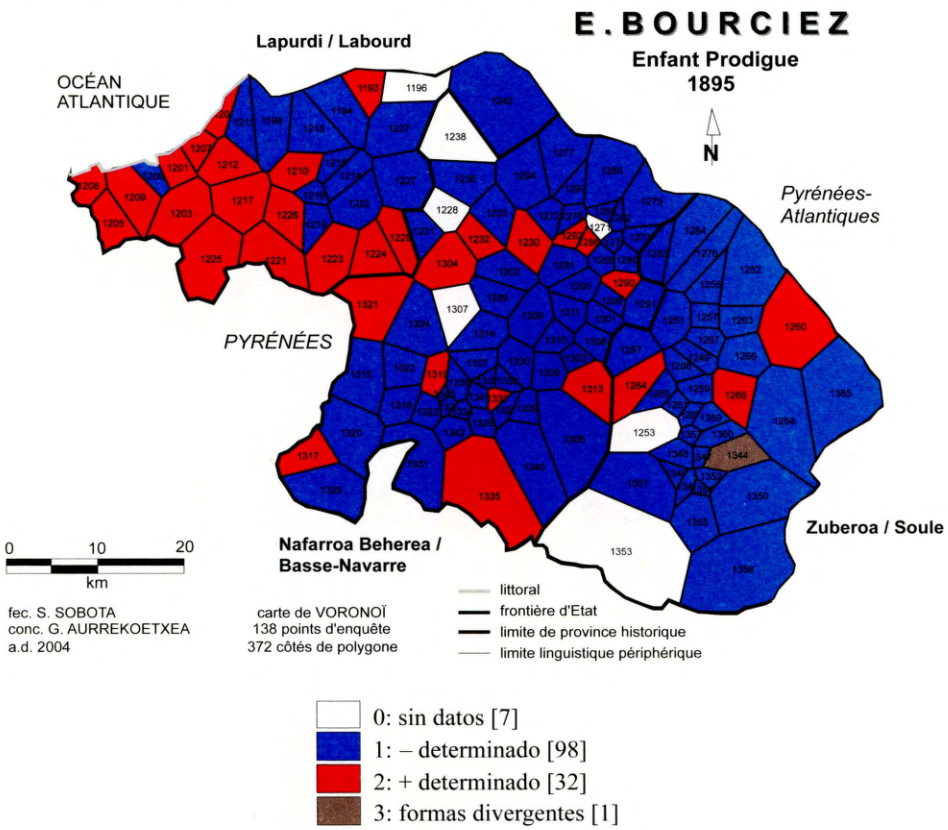


Mapa 6

Estructuras de *nombre + adjetivo*

Los pueblos o países son contables: podríamos decir sin inconveniente *cinco pueblos*. En estos casos de (22a), el empleo de la forma indeterminada se le debe achacar a la voluntad de no precisar los pueblos o países en concreto. De momento no hay ninguna razón para decir que las oraciones de (22) sean algo diferente a las de (17). Así las cosas, esperaríamos que el mapa que a continuación traemos fuera bastante similar al Mapa 4, puesto que ámbos hacen referencia a términos contables. Pero veamos el Mapa 7 de la página siguiente.

Para empezar, en este mapa no tenemos ni un solo sintagma que contenga algún determinante indefinido, como lo tenemos en (17c): No nos topamos con nada como *Behar dut ikusi kartiel* ***zenbait***: todos son SN desnudos o SD con artículo. A su vez, las formas indeterminadas son muchas, sobre todo en Zuberoa, pero en Baja Bavarra y en el territorio de Lapurdi de habla bajonavarra también son numerosas, cosa que no sucede en el mapa 4. En el territorio del labortano, por supuesto, hay más formas determinadas.



Mapa 7

Concepto contable peculiar

¿Por qué son tan diferentes los dos mapas? o, es más, ¿por qué son tan similares este mapa de nombres contables y el que hemos mostrado para los nombres incontables? (cf. Mapa 3). En nuestra opinión la respuesta se halla en el texto original en francés: aquél que Bourciez distribuyó para que fuera traducido. En ése, la oración francesa dice así: *Il faut* (...) *que je voie* ***du*** *pays*.

Hemos marcado el determinante en negrita, para que el lector constate que es el mismo que se emplea con nombres incontables. Es decir, en el texto en francés *pays* no es contable (si no llevaría el determinante *des*), sino incontable. A nuestro parecer, tener una forma incontable en el modelo a traducir habría ayudado a que en las traducciones en vasco la forma indeterminada tenga tal expansión. Las hablas que a la sazón tenían bastante debilitada la opción de emplear la forma indeterminada como recurso para expresar el valor no específico, habrían visto en el modelo francés una excusa para recuperar esa utilización ideterminada, aunque en §3.2.2 no lo hayan hecho.

Para terminar, es digno de remarcar hasta qué punto están extendidas las formas con artículo en el sur de Lapurdi, que tienen necesidad de emplear el artículo incluso superando al modelo francés, que no lo hace.

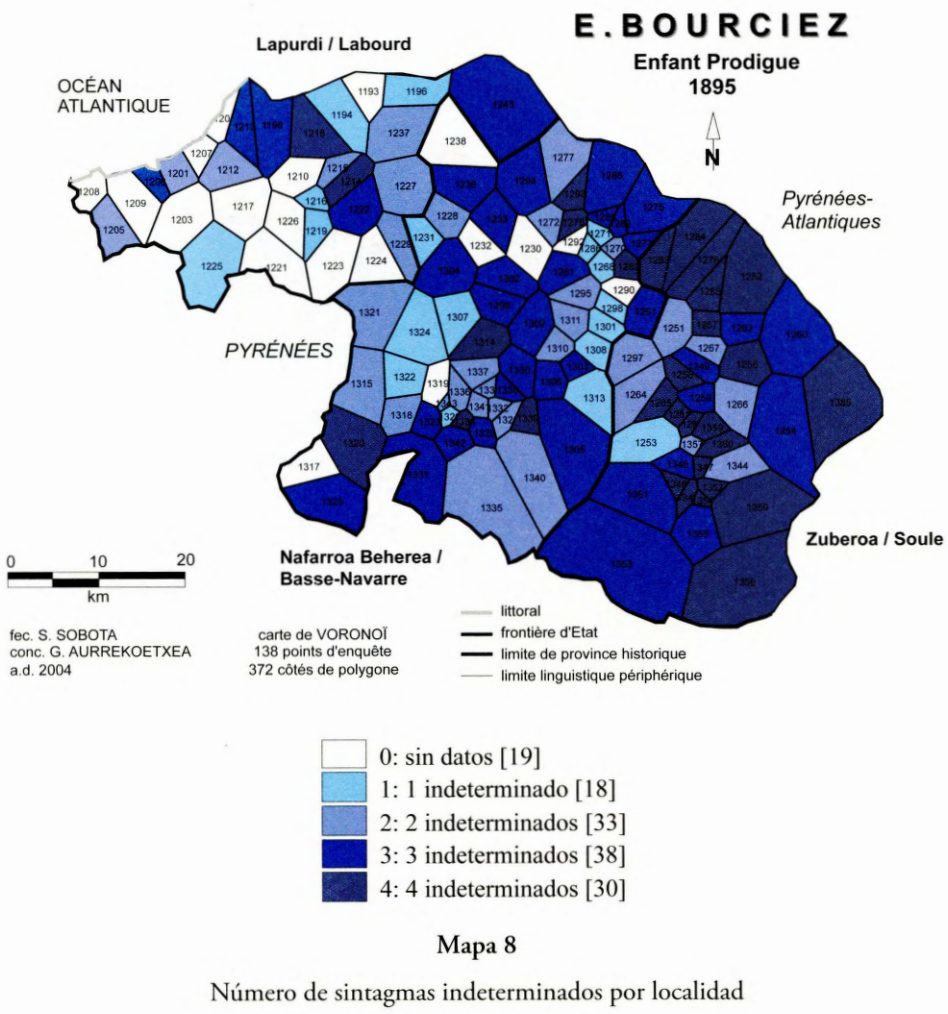
3.3. Mapas con respecto al carácter semántico y sintáctico, o "mapas de los mapas"

Como hemos explicado, el programa *VDM* no permite más que un sólo resultado por localidad (§2.3). ¿Cómo hacer, por ejemplo, la síntesis de todos los mapas con función de complemento directo (§3.2) en un sólo mapa, para poder extraer conclusiones generales?

Lo explicaremos al presentar cada mapa. Se podrían hacer numerosos mapas así y nosotros hemos traído aquí algunos.

3.3.1. Mapas sobre los sintagmas argumentales

En este apartado queremos ofrecer la síntesis de los mapas empleados en §3.2, es decir, los mapas que a continuación se muestran están basados en los mapas de sintagmas argumentales de función objeto directo. Veamos el primero:

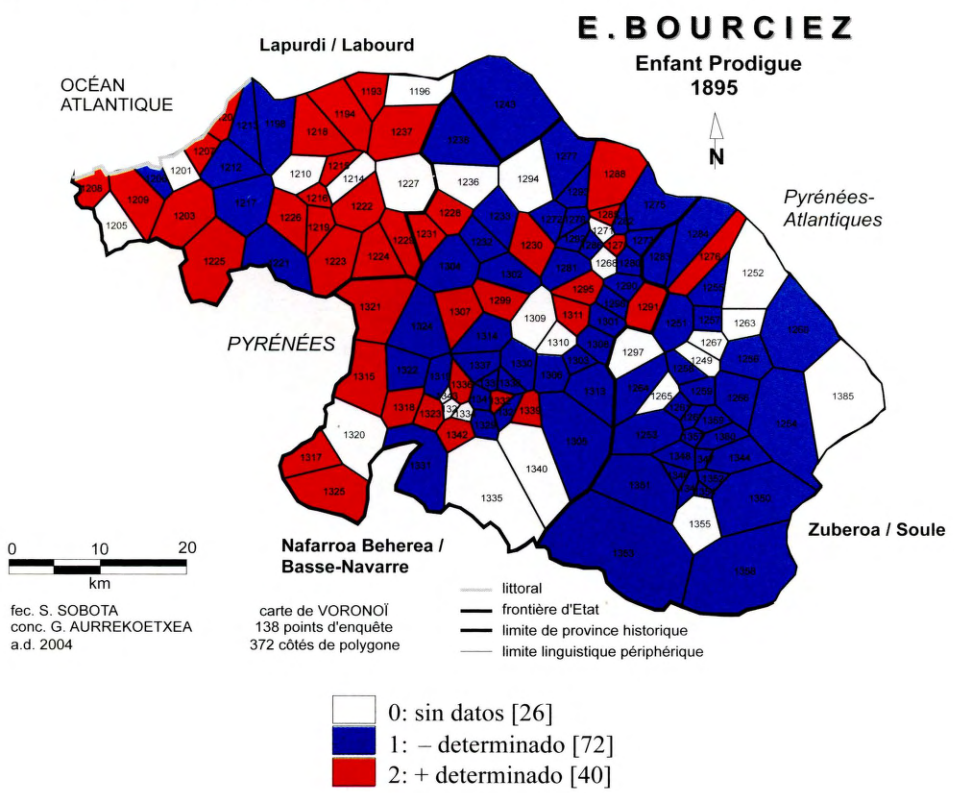


Mapa 8

Número de sintagmas indeterminados por localidad

En este mapa hemos tomado cuatro oraciones cuyo sintagma sometido a estudio tiene función de objeto directo; los del apartado §3.2, y hemos contado cuántos de ellos aparecen sin artículo.[14] Así las cosas, a las localidades que han hecho las cuatro oraciones con el sintagma indeterminado les hemos asignado el mismo lema (y color); lo mismo a las que han hecho con el indefinido tres oraciones de cuatro, etc. En este mapa, por lo tanto, hemos tratado por igual todas las oraciones; no hemos hecho divisiones por semántica o forma (contable vs. incontable, etc.). De todas formas, en este mapa se puede ver de manera clara cómo va penetrando el artículo de oeste a este, siendo una vez más el territorio del labortano el que menos formas indeterminadas presenta, y el del suletino el que más SN desnudos tiene.

En el siguiente mapa hemos trabajado de nuevo con sintagmas argumentales, pero hemos contrastado las formas determinadas y las indeterminadas. A las localidades que han hecho la mayoría de las cuatro oraciones con el artículo les corresponde el color rojo: por el contrario, a las localidades que tienen más oraciones indeterminadas, el azul. Éste es el resultado:



Mapa 9

Sintagmas argumentales determinados vs. indeterminados

[14] No hemos tenido en cuenta el mapa de §3.2.4, porque para cuando lo preparamos ya teníamos hecho éste.

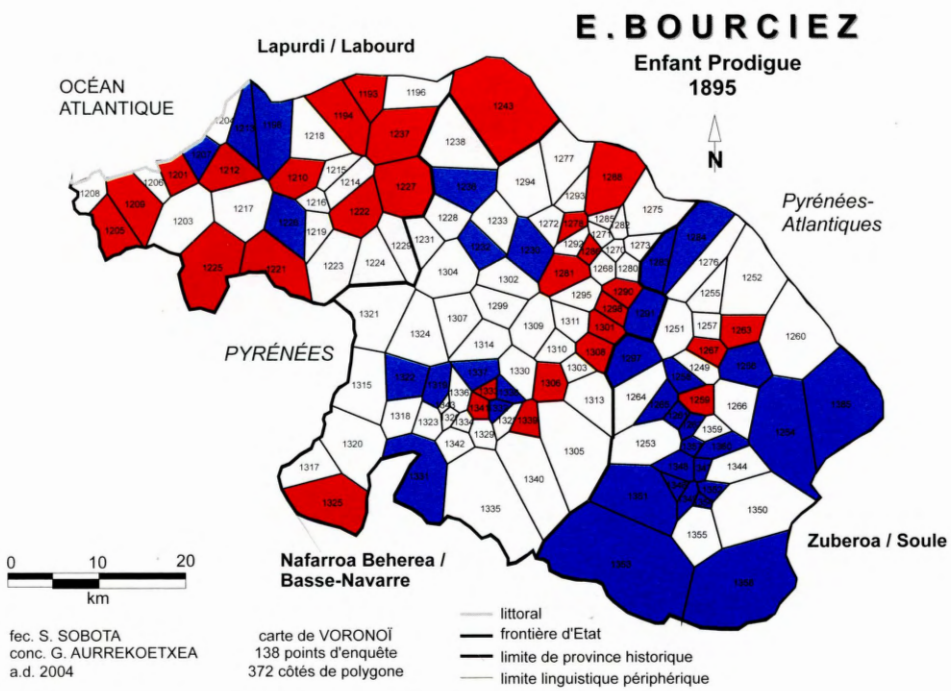
En éste también, los resultados son clarificantes. Mientras que en Zuberoa la mayoría de oraciones se han realizado sin artículo, en Lapurdi es mayoría la presencia de SD con artículo. Baja Navarra es una zona de transición evidente, como se puede ver en el mapa.

3.3.2. Mapas de los sintagmas de función atributiva

Son menos los ejemplos de sintagmas que no son argumentales (§3.1) y es necesario tener eso en cuenta a la hora de valorar los datos.

Así las cosas, esta vez no hemos realizado un mapa como el 8, pues al tener sólo dos ejemplos y ser uno de ellos casi por completo con formas indeterminadas (cf. Mapa 2), el resultado no sería representativo.

Sí hemos hecho, al igual que con los sintagmas de complemento directo, un mapa de mayorías. Corresponde el mismo lema y color a las localidades que han realizado los dos atributos sin artículo y lo mismo para los que emplean atributos determinados. El resto de combinaciones posibles han sido eliminadas, y se representan en blanco en el mapa.



0: sin datos [74]
1: – determinado [35]
2: + determinado [29]

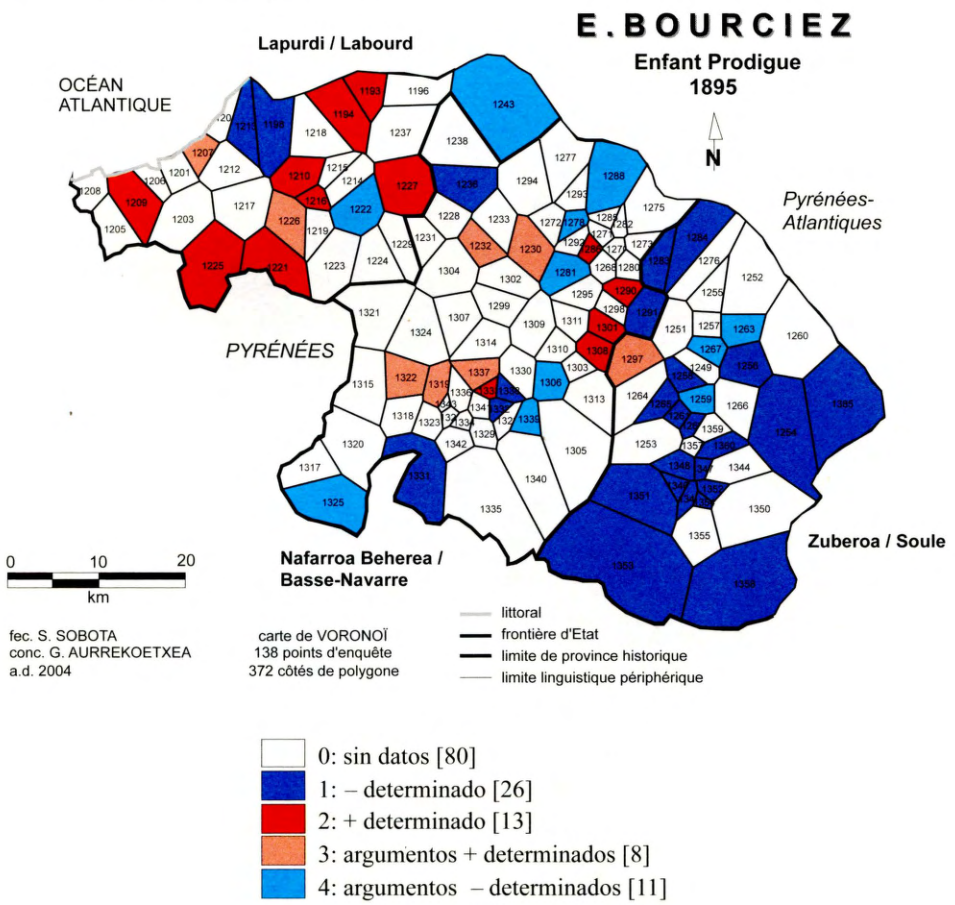
Mapa 10

Sintagmas atributivos determinados vs. indeterminados

Son muchos los espacios en blanco, porque en muchas ocasiones, a pesar de dar sin artículo las oraciones representadas en el mapa 2, las del mapa 1 son determinadas, lo que nos deja en una situación de empate entre los determinados y los indeterminados. Pero lo llamativo es lo siguiente: solamente (y casi exclusivamente) en los lados del mapa tenemos formas extremas, es decir, sobre todo en Lapurdi se han empleado siempre formas determinadas, y en Zuberoa las indeterminadas. Los grandes espacios en blanco de la Baja Navarra dejan entrever la prominencia de combinaciones intermedias, en las cuales, mientras una forma es determinada, la otra no.

3.3.3. Comparando los argumentos con los predicados

Ahora que estamos presentando mapas generales, no podemos dejar de comparar los sintagmas de función atributiva con los de complemento directo. A ello se dispone este mapa, el último, por cierto.



Mapa 11

Sintagmas proposicionales *vs.* argumentales

Como salta a la vista, son cuatro las opciones posibles en este mapa, aparte de la falta de datos, claro está. Han tomado tono azul las localidades que han utilizado principalmente formas indeterminadas, tanto con sintagmas de función complemento directo, como en sintagmas de función atributo. Les hemos dedicado el rojo a las que han preferido los sintagmas determinados, sean atributos o complementos. El rosa es para las localidades en las que, aunque en las oraciones de atributo prevalezcan las formas indeterminadas, presenten más formas determinadas entre las oraciones de complemento directo. Finalmente, el azul claro ha sido asignado a las localidades donde, siendo mayoría los sintagmas indeterminados para los atributos, realizan los complementos directos sobre todo con el artículo.

Como viene siendo habitual, en ambos costados, al este y al oeste, tenemos los colores más prominentes y en medio el resto de combinaciones posibles. De nuevo toparemos con la coherencia lingüística de Zuberoa, que, al igual que en éste, es visible también en el resto de mapas.

4. Conclusiones

Es menester ya resumir todo lo dicho hasta ahora. Partiendo de la base de que podemos clasificar las lenguas tipológicamente dependiendo de la utilización del artículo que hagan, y sabiendo que, ya sea de lengua a lengua o de dialecto a dialecto, dicha utilización cambia, prolifera o se reduce, hemos querido traer a estas páginas un ejemplo del vasco.

Escogido el corpus de Bourciez como muestra de la variación diatópica, hemos mencionado los trabajos de recogida de la época y los hemos aprovechado para presentar nuestro corpus, al mismo tiempo que lo hemos situado en su contexto geográfico y temporal correspondiente, mencionando a su vez las ventajas e inconvenientes que nos brinda.

Hemos dado una visión general del programa *VDM*, describiendo el uso que hemos realizado nosotros. Siguiendo esa línea, hemos descrito paso a paso la metodología y procedimiento de trabajo, a la vez que hemos expuesto los problemas a los que debemos enfrentarnos.

Para que la interpretación de los resultados del corpus sea correcta, hemos descrito las oraciones que hemos elegido para nuestra investigación, basándonos en los trabajos de diversos lingüistas. De esa manera, hemos tenido la oportunidad de comprobar hasta qué punto nuestros datos convergen con lo dicho hasta el momento, añadiendo a su vez, nuevo material para futuras investigaciones.

Hemos comentado los mapas uno a uno, describiendo lo que muestran y explicando de qué manera concuerdan o difieren con las teorías de los lingüistas que acabamos de mencionar. Aparte de eso, hemos expuesto las dificultades que presentan algunas oraciones del corpus y hemos intentado dar una explicación plausible a incoherencias geográficas que nos han podido surgir en los mapas.

También hemos confeccionado unos mapas generales basados en los mapas extraídos directamente del corpus. Gracias a ellos hemos podido sacar unas conclusiones más globales, resumiendo en unas pocas fotografías el empleo del artículo a finales del siglo XIX.

Desgraciadamente, la brevedad del corpus no nos ha permitido reflejar toda la casuística sobre el empleo del artículo, pero hemos podido extraer algunas conclusiones claras:

— Las formas con artículo son más numerosas en el oeste que en el este.
— Aunque antaño los SN desnudos eran posibles en cualquier posición argumental, han pervivido durante más tiempo en función de complemento directo y en vasco estándar, como en muchos dialectos, ya no son posibles ni siquiera en dicha posición.
— En lo que respecta a la aparición del artículo, los sintagmas de función complemento y los atributos presentan una geografía diferente.
— Dentro de los sintagmas de complemento directo, presentan una geografía diferente los nombres contables y los incontables.
— Desde el punto de vista diacrónico, las diferencias de isoglosas mencionadas en los puntos anteriores abogarían por una expansión tardía del artículo del oeste al este, teniendo incluso la posibilidad de distinguir diferentes fases de gramaticalización.
— Mirando a la geografía, Zuberoa muestra una coherencia lingüística muy grande en favor de los indeterminados y el territorio del labortano (el sur de Lapurdi) puede diferenciarse muy bien mediante el empleo del artículo, aunque no es tan claro como el de Zuberoa. La Baja Navarra actúa a menudo como zona de transición.
— Los mapas parecen mostrar un estadio de la lengua en el cual el artículo presenta diferentes grados de gramaticalización. El *markedness shift* habría sucedido ya en algunos puntos de Lapurdi, mientras que en Zuberoa prácticamente no habría comenzado ni siquiera la alternancia entre formas desnudas y determinadas.
— Desde el lado metodológico es conveniente mentar la fiabilidad de los mapas. No se puede realizar una interpretación pueblo a pueblo y hay que tomar el mapa en su totalidad: debido a las características del sistema de recogida de datos (un texto por localidad, traducido por gente alfabetizada, etc.), en algunas localidades aparecen incoherencias dialectológicas. Es labor de los lingüistas hacer interpretaciones más generales, por encima de esas carencias.

No quisiera terminar, sin hacer una proclama en contra de la tradicional (al menos desde tiempos de Saussure) división entre diacronía y sincronía. Entre los objetivos de este trabajo, aunque sea humildemente, estaba mostrar la debilidad de esa frontera. Que la metodología sincrónica utilizada haya proporcionado también a los diacronistas algunas conclusiones a tener en cuenta sería un sueldo más que digno a cambio del trabajo realizado.

5. Referencias bibliográficas

Allières, J., 1960-1961, «Petit atlas linguistique basque-français <Sacaze> I-II», *Via Domitia VII-VIII*, Faculté des Lettres et Sciences humaines de Toulouse, Toulouse, 82-126 & 205-224.

Andersen, H., 2001a, «Markedness and the Theory of Linguistic Change», in H. Andersen (ed.), *Actualization. Linguistic Change in Progress*, John Benjamins Publishing Company, Amsterdam & Philadelphia, 21-57.

—, 2001b, «Actualization and the (Uni)directionality of Change», in H. Andersen (ed.), *Actualization. Linguistic Change in Progress*, John Benjamins Publishing Company, Amsterdam & Philadelphia, 226-248.

Artiagoitia, X., 2004, «Izen Sintagmaren birziklatzea: IS-tik inguruko funtzio buruetara», in P. Albizu & B. Fernández (eds.), *Euskal gramatika XXI. mendearen atarian: arazo zaharrak, azterbide berriak*, Diputación Foral de Álava, Vitoria-Gasteiz, 11-38.

Aurrekoetxea, G. & X. Videgain, 2004, *Haur prodigoaren parabola Ipar Euskal Herriko 150 bertsiotan*, *ASJU*ren gehigarriak, XLIX, UPV/EHU, Bilbao.

Azkarate, M., & P. Altuna, 2001, «3.3.3. Predikatu sintagmak: atributu sintagmak», in *Euskal Morfologiaren Historia*, Elkarlanean, Donostia, 74-76.

Croft, W., 1993[2] [1990], *Typology and universals*, Cambridge U. P., Cambridge.

Eguren, L., s.d., «Non-canonical uses of the article in Basque». Manuscrito de la UAM.

Etxeberria, U. & R. Etxepare, 2008, «Zenbatzaileak komunztatzen ez direnean: Hiru sistema», in *Aldaketak, Aldaerak, Bariazioak Euskaran eta Euskal Testugintzan*, Abenduaren 12-13a, IKER-CNRS, Baiona. Handout del congreso.

Gándara, A. & E. Santazilia, 2007, «Zehaztapen batzuk artikuluaren erabileraz Bourciez korpusean». Manuscrito de la UPV/EHU.

Longobardi, G., 2001, «The structure of DPs: some Principles, Parameters and Problems», in C. Collins & M. Baltin (eds.). *The Handbook of Contemporary Syntactic Theory*, Blackwell, USA/UK.

Manterola, J., 2006, «*-a* euskal artikulu definituaren gainean zenbait ohar», on line en la base de datos ARTXIKER: http://artxiker.ccsd.cnrs.fr/artxibo-00142080/eu/ [consulta: 2009-09-03]. Publicado también en *ASJU* 40: 1-2, 651-676.

—, 2008, «*-a* morfemaren erabilera (eza) ekialdeko euskaretan», on line en la base de datos ARTXIKER: http://artxiker.ccsd.cnrs.fr/artxibo-00352804/eu/ [consulta: 2009-09-03].

Mitxelena, L., 1978, «II. Vasc. ON DA/GAUZA ONA DA», in «Miscelánea filológica vasca», *FLV* X: 29, 208-218.

Oyharçabal, B., 1992a, «Euskararen mugez egin lehen mapak (1806-1807)», in *Luis Villasanteri omenaldia*, Iker-6, Euskaltzaindia, Bilbao, 349-366.

—, 1992b, «Lehenbiziko inkesta geo-linguistikoak Euskal Herrian frantses lehen Inperioaren denboran: ipar aldean bildu dokumentuak», in *Nazioarteko Dialektologia biltzarra. Agiriak*, Iker-7, Euskaltzaindia, Bilbao, 285-298.

—, 1994, «Les documents recueillis lors des enquêtes linguistiques en Pays Basque durant la période révolutionnaire et le Premier Empire», in J. B. Orpustan (ed.), *La révolution française dans l'histoire eta la litterature basques du XIXème siècle*, Izpegi, Baigorri, 62-119.

—, 1995, «Euskararen mugak hego aldean 1807.ean: Coquebert de Montbret-ek bildu dokumentuak», in R. Gómez & J. A. Lakarra (eds.), *Euskal Dialektologiako kongresua (Donostia, 1991ko Irailak 2-6), Suplementos de ASJU* XXVIII, Diputación Foral de Gipuzkoa, Donostia.

Sacaze, J., 1887, *Recueil de linguistique et de toponymie des Pyrenées*.

Santazilia, E., 2007, «Edouard Bourciezen *Haur prodigoaren parabola*: edizio paleografiko konparatua», *ASJU* 41-1, 397-402.

Simoni-Aurembou, M.-R., 1989, «La couverture géolinguistique de l'Empire français: l'enquête de l'enfant prodigue», in *Espaces Romans: Études de dialectologie et de géolinguistique offertes à Gaston Tuaillon*, Université Sthendal-Grenoble, Grenoble, II, 114-135.

Videgain, X., 2005, «Présentation du recueil Bourciez», *Lapurdum* X, 315-324.

Zabala, I., 2001, «4.2 Nominal Predication: copulative sentences and secondary predication», in J. I. Hualde & J. Ortiz de Urbina (eds.), *A Grammar of Basque*, Mouton de Gruyter, Berlin, 426-447.

TECHNOLOGY FOR PROSODIC VARIATION

Gotzon Aurrekoetxea & Aitor Iglesias
UPV/EHU

Abstract[1]

This contribution is included in the "Corpora of Spoken Dialects of the Basque Language-EDAK" project, which aims to create the prosodic multimedia atlas of the Basque language. It presents the first outcome taking into account data from two localities situated in the Western part of the Basque territory: Ondarroa and Larrabetzu. Although both belong to the Biscayan dialect, they have a different accentual system.

Apart from the computer tools being used in the project, this paper deals with the sociolinguistic differences that exist in both localities between two generations (adults and young people) and the geolinguistic differences between these two localities.

Key words: *prosodic variation, technology, Basque language, sociolinguistic variation, geolinguistic variation.*

1. Introduction

The research into Basque prosody has increased considerably in recent years, above all on account of different researchers like G. Elordieta, I. Gaminde and J. I. Hualde.

Even if we assume that this research constitutes great progress in the field, the fact that none of it takes the space of the Basque language as a whole into consideration, and that this work has been carried out using different methodologies has prompted the EUDIA research team to embark on the EDAK project (Corpora of Spoken Dialects of the Basque language).

This research project has two bases: to take an in-depth look at the knowledge of Basque prosodic variation, and to use methodological aspects of the AMPER[2] project.

This paper sets out to present the technology we are using and to submit the first partial analysis of our data.

---

[1] This research has been funded by the Ministry of Science and Innovation of the Spanish Government (HUM2007-65094).

[2] "Atlas multimedia de l'espace romanique" (Contini 1992, Romano 2001). For bibliography about the AMPER project see M. Contini, A. Romano, L. de Castro Moutinho & E. Fernández Rei "L'avancement des recherches en géoprosodie et le projet *AMPER*", *EFE,* ISSN 1575-5533, XVIII, 2009, pp. 109-122.

The first part deals with the itinerary the data has undergone between the recording of them and their publication. We have paid special attention to the technology used to record, to mark and label the data, and subsequently to analyse them.

The second part presents the data in which two generations in the same locality are compared in order to show sociolinguistic variation, and data from the two localities to show geolinguistic variation.

2. The use of technology

Our research team was already familiar with the features of automated technology. So we opted for free software. Nowadays, there is more available and there are more means for accessing the new free technology; there is abundant software, which is becoming increasingly efficient and enables the proposed aims to be achieved without any reduction in quality.

We are convinced that in the near future we will have all the technologies we need free of charge, even though some are not yet available.

2.1. Data gathering

For the data gathering and sound recording we used only laptops, which were equipped with "Audacity" software[3] and USB microphones (PC Headset 960 USB).[4] Once the sound had been recorded, various copies were made in a range of mediums: CD, hard disc, etc. and they were kept on different premises.

The use of laptops allowed the sound material to be managed more effectively and more rapidly. It also facilitated the management of the technical support or means we had, so it could be used either in the data gathering or in the analysis (Aurrekoetxea, Sánchez & Odriozola 2009).

97192: Etorri da? [Has he come?]	V14
97193: Erosi du? [Has he bought it?]	V15
97194: Laguna sartu da? [Has the friend come in?]	V16
97195: Laguna, etorri da? [The friend, has he come in?]	V17
97197: Erosi du ogia? [Has he bought the bread?]	V18
97198: Zer ikusi du? [What has he seen?]	V19
97199: Non ikusi du? [Where has he seen it?]	V20
97200: Zer ikusi du alabak? [What has the daughter seen?]	V21
97201: Alabak, zer ikusi du? [The daughter, what has she seen?]	V22

Questions used to gather the data for this contribution.

2.2. Sound annotation

The annotation task was carried out using the SFSWin program.[5] Even if this tool offers the possibility of conducting automatic annotation, it was decided that it should be done manually because of the way the gathering was structured.

---

[3] *http://audacity.sourceforge.net/*

[4] *www.logitech.com.*

[5] *http://www.phon.ucl.ac.uk/resource/sfs/download.htm.*

Figure 1 shows an annotated sound file: the beginning of the annotation of each question is marked by a "V" followed by a numerical code referring to the number of the question; and the end of the question is marked by the sign "/". We only annotated the "good" answer; in other words, the answer that corresponded to the question and that produced the word sequence we were seeking; we did not take into account the sound produced by the research, nor the translations, which we did not consider to be valid.

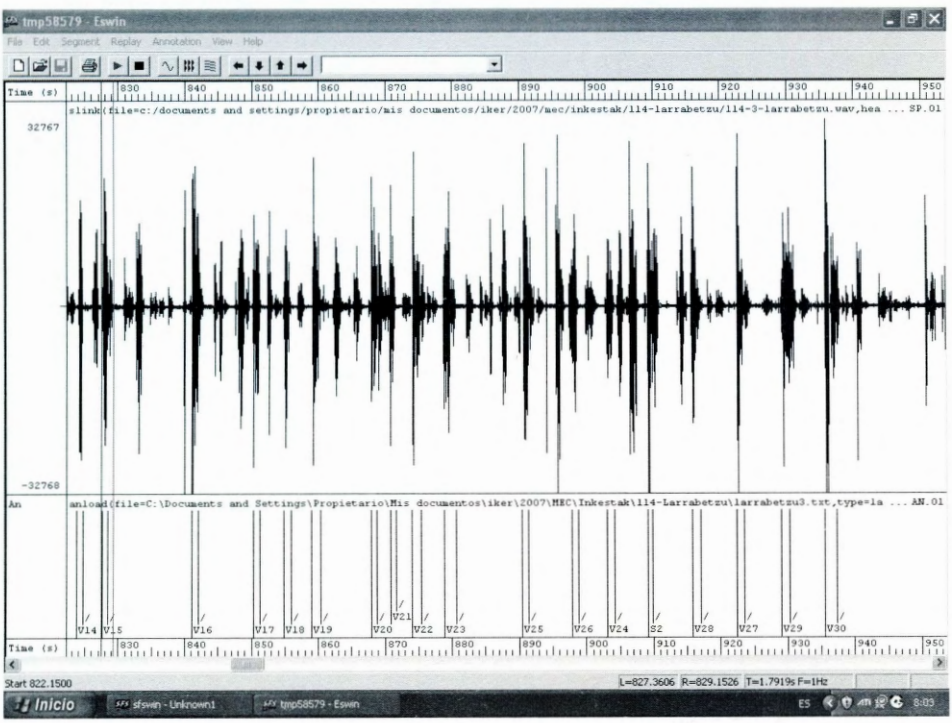


Fig. 1

Image of the annotation of EDAK data using the SFSWin program

2.3. Sound labelling

The output of the SFSWin program is a .txt file which links the annotations of each question (V14, V15...) with the sound location in the recording. Figure 2 shows the annotations of the questions analysed in this paper only.

Once the sound had been annotated, the labelling task was carried out by using the "txertatu_etiketak" Active perl script.[6] This script has been adapted by Aholab,

[6] *http://www.activestate.com/.*

823.60378	V14
824.67929	/
827.55588	V15
828.40101	/
840.99302	V16
842.06804	/
850.21459	V17
851.27490	/
854.78713	V18
855.86241	/
859.00270	V19
860.30397	/
867.96080	V20
868.69909	/
870.85901	V21
871.57844	/
874.02069	V22
875.34611	/

Fig. 2

The output of the SFSWin program

the Signal Processing laboratory *(http://aholab.ehu.es)* of the University of the Basque Country (UPV-EHU), with which we are working.

97192-114-3-1-13:43.603-13:44.679
97193-114-3-1-13:47.555-13:48.401
97194-114-3-1-14:00.993-14:02.068
97195-114-3-1-14:10.214-14:11.274
97196-114-3-1-14:14.787-14:15.862
97197-114-3-1-14:19.002-14:20.303
97198-114-3-1-14:27.960-14:28.699
97199-114-3-1-14:30.859-14:31.578
97200-114-3-1-14:34.020-14:35.346
97201-114-3-1-14:38.763-14:40.635

Fig.3

Image of the labelling of EDAK data

The fig. 3 has different columns separated by hyphens:

— the first column refers to the question
— the second to the locality
— the third to the informant
— the fourth to the answer

— the fifth to the position of the beginning of the sound of the answer obtained by the SFSWin program and measured in minutes, seconds and hundredths of a second
— the last column refers to the end of the sound of the answer

Once these tasks had been carried out, the data were ready to be used in the acoustic analysis or in the audible atlas.

2.4. Data storage

The recorded data and their transcriptions were stored in two formats: MySQL and TEI (Text Encoding Initiative).

The first one was used to enter the data into the computer system and to exploit the data geolinguistically. The second one was used to facilitate transfer to the different systems. The updating between these two systems was automated and the data were entered only once.

2.5. Acoustic analysis

The following step is the acoustic analysis of the data. The Praat program[7] was used for this task. This program is one of the most widespread free programs for acoustic analysis, in addition to the "Segment Data" Script.[8]

2.6. Graphics

For the graphics, Microsoft Office Excel was used for the purposes of this paper, but we are starting to use the OpenOffice Cal program.

3. Background of the field

As in other latitudes, prosodic variation has not been considered in Basque dialectology when comparing dialects and making dialect maps. Only a few pieces of research have studied geo-prosodic variation. It was K. Mitxelena (1972) who first distinguished different accent types and divided the geography of the Basque language into four main types, namely type I (Western and Middle part), type II (Eastern part), type III (Southern part of Navarre) and type IV (accent of Bidasoa). Txillardegi distinguished only two types: the Western accent and the Eastern accent. Subsequently, J. I. Hualde (1990), I. Gaminde (1995) and others have studied different kinds of accents. In all of them, however, the basic materials used for distinguishing the different kinds of accents were not gathered by means of the same methodology and an identical questionnaire.

The studies of intonation are more recent in the Basque language. There are very few people who have studied the features of intonation in Basque. The most impor-

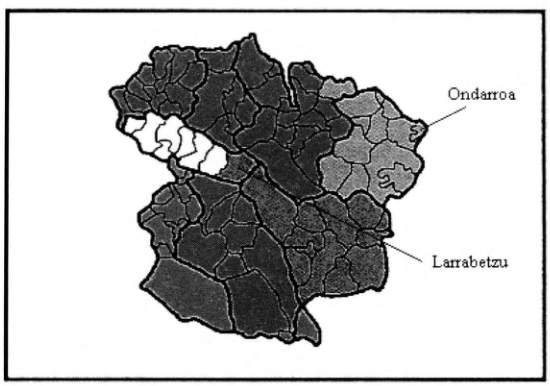
[7] *http://www.fon.hum.uva.nl/praat/.*
[8] *http://www.helsinki.fi/~lennes/praat-scripts/.*

tant researchers in this field are I. Gaminde (1995, 2002, 2004, 2006, etc.), G. Elordieta (1997, 2000, 2007, etc.) and J. I. Hualde (1997, 2002, 2003a, 2003b, etc.).

The EDAK corpus provides us with an opportunity to compare, for the first time, prosodic variation in Basque by using the same questionnaire and gathering information by means of an identical methodology, and not only geo-prosodic variation, but also socio-prosodic variation.

If we compare the utterances of adults with those of young people from the two localities, we can obtain two figures for the differences between the two localities, differences between adults and differences between young people.

The utterances we have selected are questions. We have chosen two kinds of questions: five Yes/No questions and four Wh- questions.



Map 1

Location of Ondarroa and Larrabetzu in the Biscayan accent distribution (Gaminde 2007: 66)

4. The data

For this paper, we have selected data from two localities: Larrabetzu and Ondarroa, each of which is located in a different accent area (as we can see in Map 1). They allow us to research two types of variation: geolinguistic (comparing data from two generations in the same locality) and sociolinguistic (comparing data from the two localities).

We wanted to select two localities with different accent systems in order to analyse the role played by the accent in the intonation system.

The EDAK project questionnaire has 22 items for analysing intonation: 12 are affirmative sentences and 10 are interrogative sentences. We have selected 9 interrogative sentences. All of them have a question structure, although the first five are yes/no questions and the last four are Wh-questions.

These interrogative sentences were asked three times: at the beginning of the data gathering, halfway through it, and at the end. Some of the answers have been

excluded for different reasons; in these cases there will be two, instead of three (see Fig. 5, for example).

5. Intonation patterns in Larrabetzu and Ondarroa

The data gathered enables us to study socio-prosodic and geo-prosodic variation. First of all, data taken from two generations in both localities is compared. By comparing these data it is possible to confirm whether there is socio-prosodic variation or not in both localities.

After that, data from the two localities is compared, but data gathered from the two generations is analysed separately. Furthermore, we will be showing whether the geo-prosodic variation between the two localities is the same in the two generations or not.

5.1. Intonation in Larrabetzu

First of all, the intonation pattern in the two localities, Larrabetzu and Ondarroa, is analysed. This task is carried out separately: firstly, the intonation pattern of adults and secondly, the pattern of young people.

5.1.1. Intonation pattern of adults

We have analysed two kinds of patterns: patterns in y/n questions and patterns in wh-questions. For the first type we have 5 questions with the verb at the end and with the verb at the beginning. For the wh- ones we have 4 questions: three of them begin with a question mark and the last one has an introductory noun.

Y/N questions

Each question is analysed separately because they have a different number of syllables and different patterns.

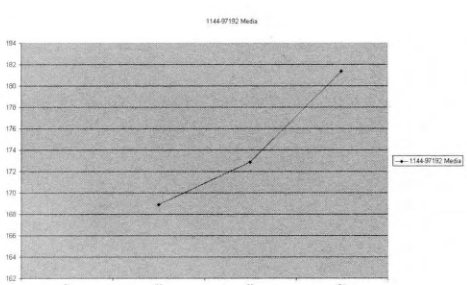


Fig. 4

Etorri da? [Has he come?]

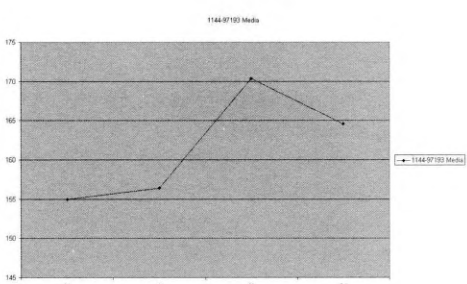


Fig. 5

Erosi du? [Has he bought it?]

The intonation of y/n questions among adults in Larrabetzu has more than one pattern:

— First pattern: the intonation curve is upward as far as the last accentuated syllable and downward on the last syllable (Figs. 5 and 6).
— Second pattern: the intonation curve is only upward (Fig. 4).
— Third pattern: the curve is upward from the beginning as far as the last accentuated syllable; then, there is a downward syllable and on the last syllable there is an upward curve (Fig. 7).

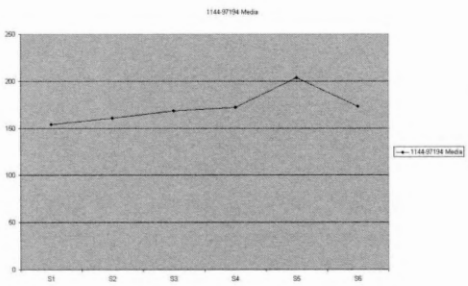


Fig. 6

Laguna sartu da?
[Has the friend come in?]

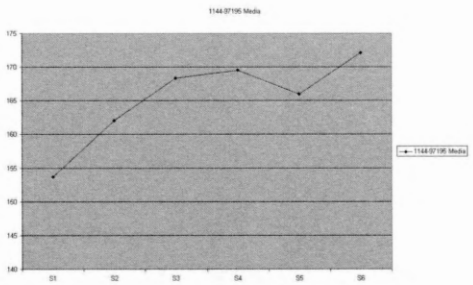


Fig. 7

Laguna, etorri da?
[The friend, has he come in?]

Different patterns emerge in the three sentences used to show the intonation of Wh-questions:

— First pattern: the intonation curve is falling from the first syllable as far as the last one (Figs. 8 and 10).
— Second pattern: the curve is falling as far as the last accentuated syllable and after that it is rising (Fig. 9).

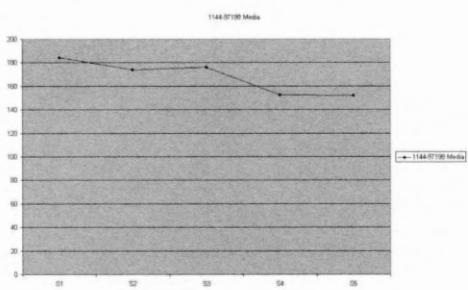


Fig. 8

Zer ikusi du? [What has he seen?]

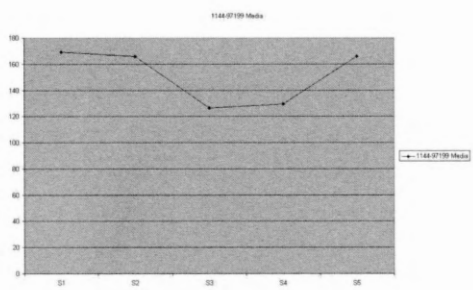


Fig. 9

Non ikusi du? [Where has he seen it?]

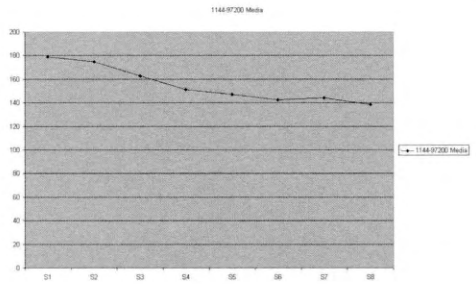


Fig. 10

Zer ikusi du alabak? [What has the daughter seen?]

5.1.2. Intonation pattern of young people

Y/ N questions

We can say that there are two patterns:

— First pattern: the curve is upward (Figs. 11 and 12).

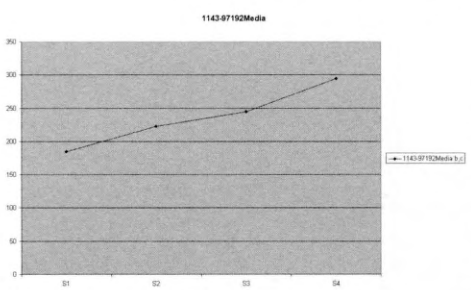


Fig. 11

Etorri da? [Has he come?]

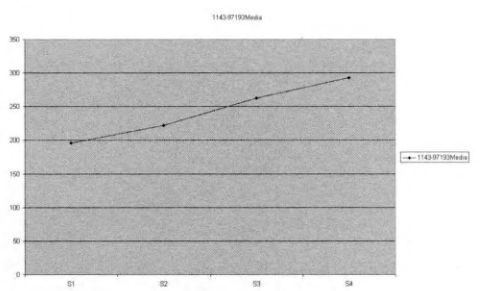


Fig. 12

Erosi du? [Has he bought it?]

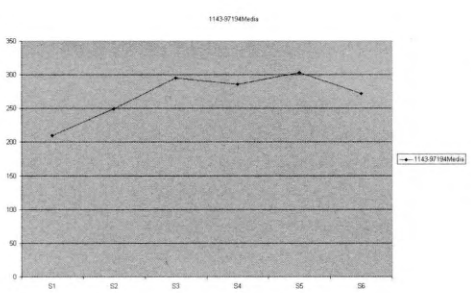


Fig. 13

Laguna sartu da?
[Has the friend come in?]

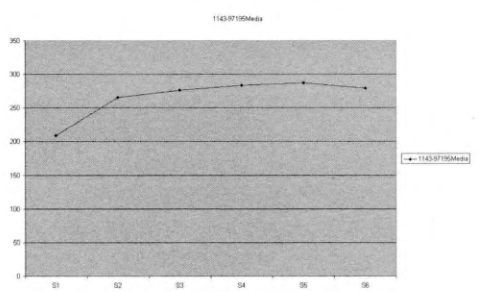


Fig. 14

Laguna, etorri da?
[The friend, has he come in?]

— Second pattern: the curve is upward, but downward on the last syllable (Figs. 13 and 14).

Wh-questions

We have the same pattern in the three utterances: the utterances begin with %H mark, falls as far as the accentuated syllable and rises on the last syllable (Figs. 15 and 16). Fig. 17 shows a similar pattern with little difference, owing to the fact that it ends on a noun after the verb.

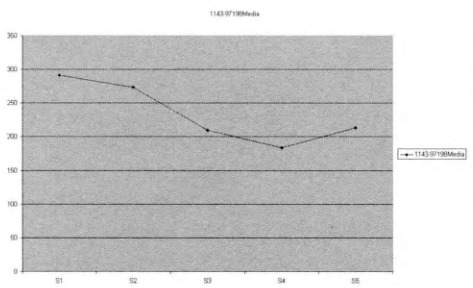


Fig. 15

Zer ikusi du? [What has he seen?]

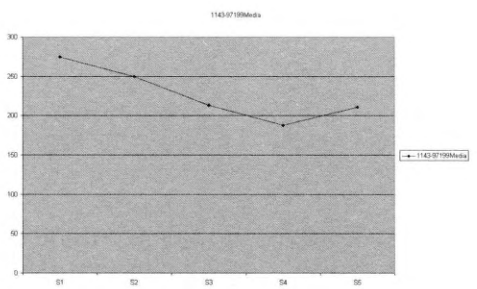


Fig. 16

Non ikusi du? [Where has he seen it?]

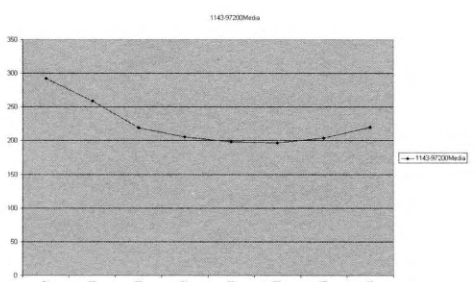


Fig. 17

Zer ikusi du alabak? [What has the daughter seen?]

5.2. Intonation in Ondarroa

5.2.1. Intonation pattern of adults

Y/ N questions

The same pattern appears in all of the utterances: the utterances begin at the lowest position and the F0 gradually increases as far as the end of the utterance (Figs. 18, 20 and 21). But Fig. 19 shows a particular pattern that displays a very marked curve in the first part corresponding to the noun.

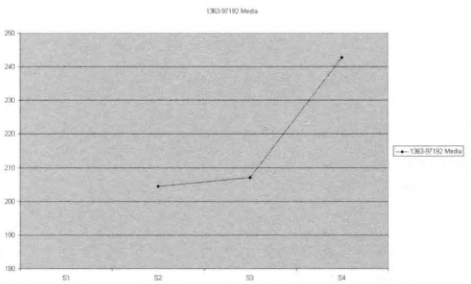


Fig. 18

Etorri da? [Has he come?]

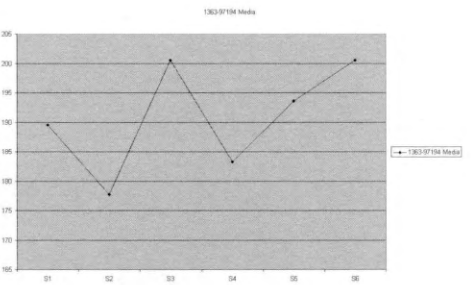


Fig. 19

Laguna sartu da? [Has the friend come in?]

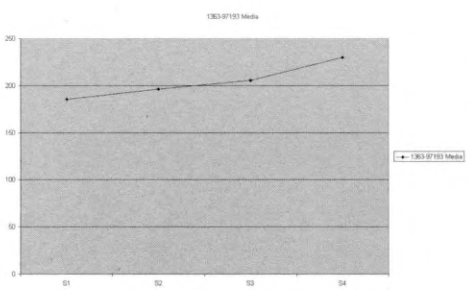


Fig. 20

Erosi du?
[Has he bought it?]

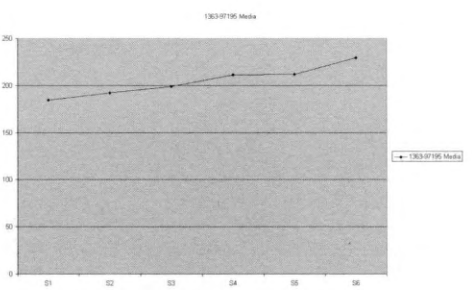


Fig. 21

Laguna, etorri da?
[The friend, has he come in?]

Wh-questions

The pattern of the three Wh-questions is similar (Figs. 22, 23 and 24). If we consider Fig. 23, it displays a slight difference compared with the others owing to the including of a noun at the end of the sentence: the pattern begins with %H and rises on the last syllable.

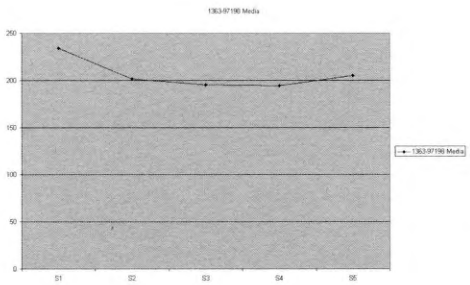


Fig. 22

Zer ikusi du?
[What has he seen?]

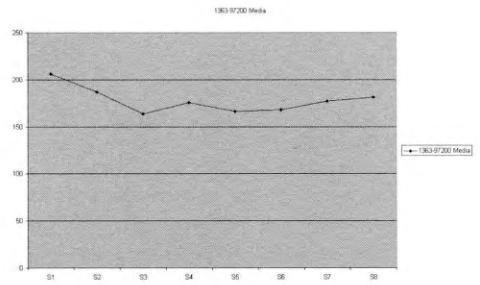


Fig. 23

Zer ikusi du alabak?
[What has the daughter seen?]

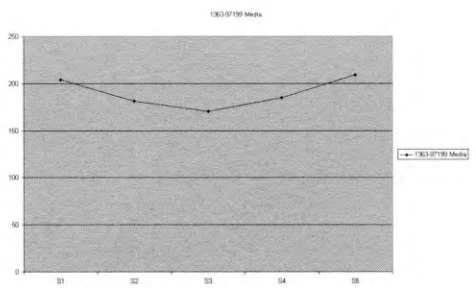


Fig. 24

Non ikusi du? [Where has he seen it?]

5.2.2. Intonation pattern of young people

Y/ N questions

The four utterances show the same intonation pattern: the sentence begins with %L mark and rises right up to the end (Figs. 25-28).

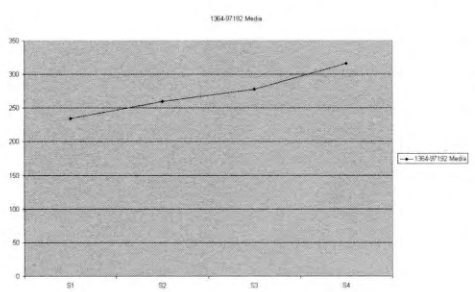


Fig. 25

Etorri da? [Has he come?]

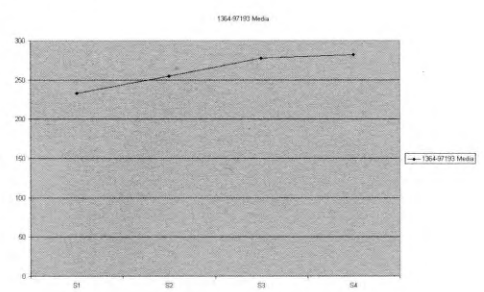


Fig. 26

Erosi du? [Has he bought it?]

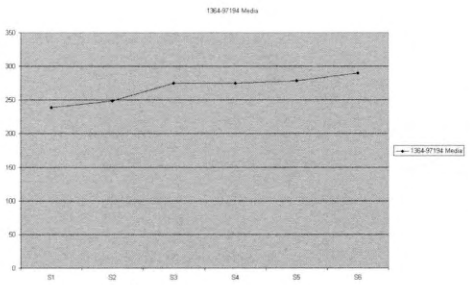


Fig. 27

Laguna sartu da??
[Has the friend come in?]

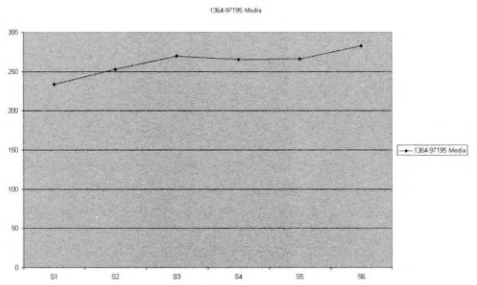


Fig. 28

Laguna, etorri da?
[The friend, has he come in?]

Wh-questions

But in Wh-questions there are two patterns:

— First pattern: the utterance is %H at the beginning and falls as far as the end (Figs. 29 and 30).

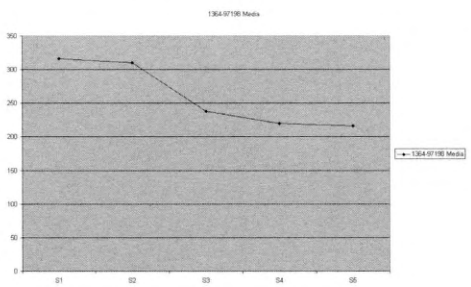


Fig. 29

Zer ikusi du? [What has he seen?]

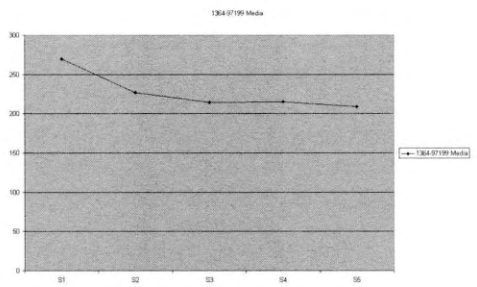


Fig. 30

Non ikusi du [Where has he seen it?]

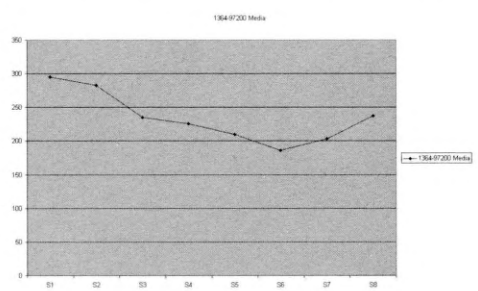


Fig. 31

Zer ikusi du alabak? [What has the daughter seen?]

— Second pattern: the inclusion of the noun after the verb changes the pattern of the intonation, which is upward from the beginning of the noun as far as the end (Fig. 31).

6. Socio-prosodic variation

6.1. Socio-prosodic variation in Larrabetzu

To measure the socio-prosodic variation, data taken from adults and young people will be compared (see Figs. 4 and 11 for the first utterance, Figs. 5 and 12 for the second ; Figs. 6 and 13 for the third ; Figs. 7 and 14 for the fourth ; Figs. 8 and 15 for the fifth ; Figs. 10 and 16 for the sixth ; Figs. 9 and 17 for the seventh).

In y/n questions there are different systems:

— The first utterance shows the same pattern (Fig. 32): the utterance begins with %L mark and rises as far as the end. Young people produce this sentence with one syllable less, because they do not pronounce the first syllable of "etorri".

— In the second utterance (Fig. 33), while the adults' curve rises as far as the end of the sentence, the young people's curve rises and falls slightly on the last syllable, thus indicating variation.

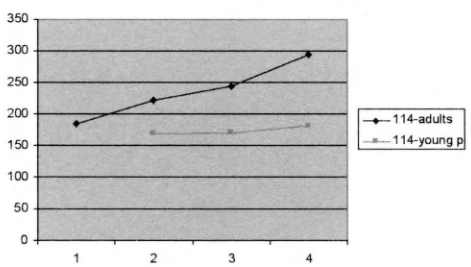


Fig. 32

Etorri da? [Has he come?]

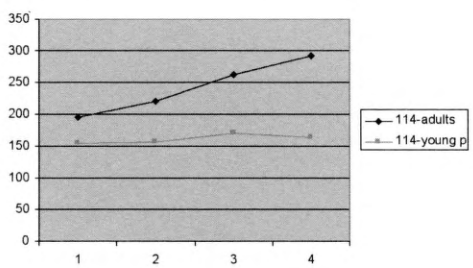


Fig. 33

Erosi du? (Has he bought it?]

— In the third utterance (Fig. 34), the adults' curve rises from the first syllable onwards and maintains its level more or less at the same height and falls on the last syllable, whereas the young people's curve rises at first, and then falls on the last syllable, indicating an absence of variation.

— In the fourth utterance (Fig. 35), there is no prosodic variation except on the last syllable.

— In the fifth utterance (Fig. 36), the adults' curve rises on the last syllable, whereas the young people's curve falls slightly, which indicates variation.

— In the sixth utterance (Fig. 37), there is no variation between the patterns produced by adults and young people.

— In the seventh utterance (Figs. 38a and 38b), the two graphics differ considerably: the pattern produced by adults rises slightly on the last syllable, but in young people the pattern shows a slight fall; thus, there is linguistic variation.

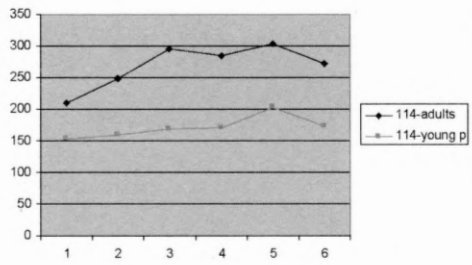


Fig. 34

Laguna sartu da
[Has the friend come in?]

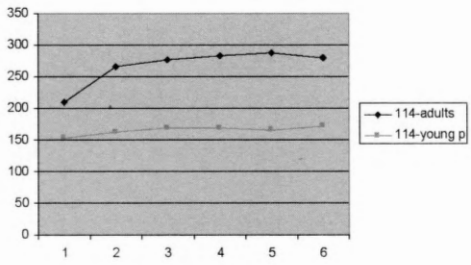


Fig. 35

Laguna, etorri da?
[The friend, has he come in?]

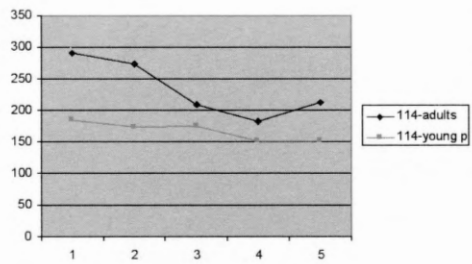


Fig. 36

Zer ikusi du? [What has he seen?]

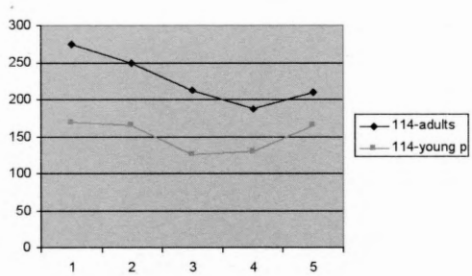


Fig. 37

Non ikusi du? [Where has he seen it?]

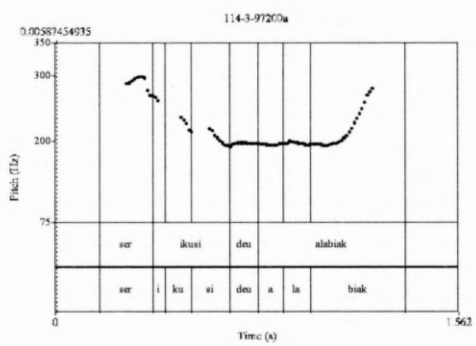


Fig. 38a

Zer ikusi du alabak?
[What has the daughter seen?]

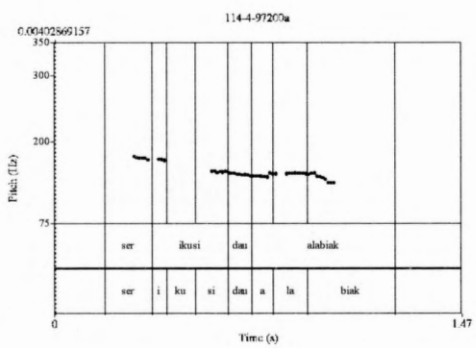


Fig. 38b

Zer ikusi du alabak?
[What has the daughter seen?]

Therefore, three out of the seven utterances display different curves indicating different prosodic structure between data collected from adults and young people (43%).

Consequently, we can affirm from these data in Larrabetzu that socio-prosodic variation between adults and young people can be found.

6.2. Socio-prosodic variation in Ondarroa

To measure the socio-prosodic variation, data taken from adults and young people are compared (see Fig. 39).

Y/N questions. The results are as follows:

— in the first utterance (Fig. 39), the intonation pattern is similar in adults and young people; so there is no variation;

— in the second utterance (Fig. 40), there is a rising curve in adults and young people; so there is no variation.

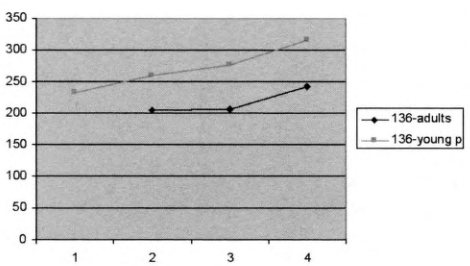


Fig. 39

Etorri da? [Has he come?]

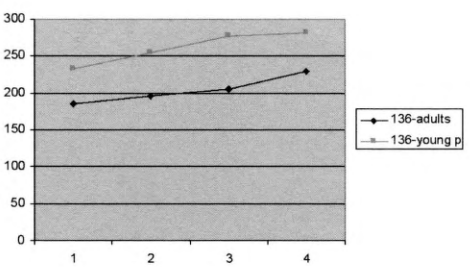


Fig. 40

Erosi du? [Has he bought it?]

— in the third utterance (Fig. 41), even if the curve is not the same in the utterances of the two generations, the main features are identical, so there is no variation.

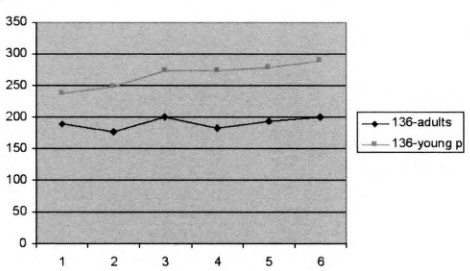


Fig. 41

Laguna sartu da?
[Has the friend come in?]

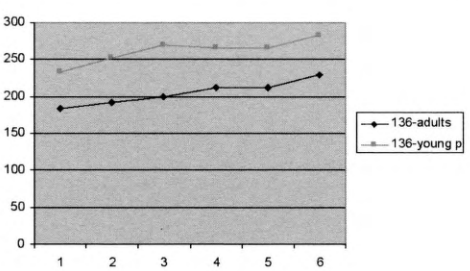


Fig. 42

Laguna, etorri da?
[The friend, has he come in?]

— in the ninth utterance (Fig. 42), the curves in the patterns produced by adults and young people are very similar; so we do not consider socio-prosodic variation.

Wh-question sentences.

— in the first utterance (Fig. 43), the patterns differ at the end of the curve: in adults it rises, whereas in young people it falls; so there is variation.

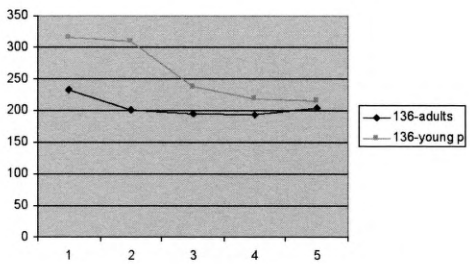


Fig. 43

Zer ikusi du? [What has he seen?]

— in the second utterance (Figs. 44a and 44b), there is socio-prosodic variation because the pattern produced by adults shows a falling curve right from the beginning of the utterance, whereas the curve of young people ends on an upward curve. So, there is variation.

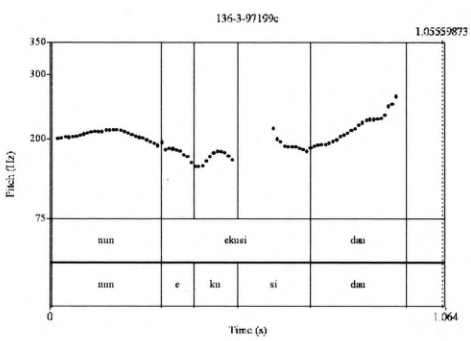


Fig. 44a

Non ikusi du? [Where has he seen it?]

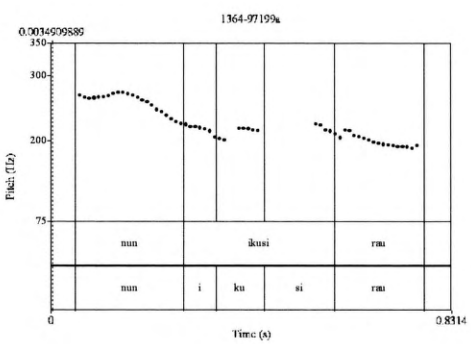


Fig. 44b

Non ikusi du? [Where has he seen it?]

— In the third utterance (Fig. 45), the main features of the intonation curves are similar, some differences notwithstanding. So there is no variation.

In five utterances in the seven cases analysed, the curve of the utterance is similar between adults and young people, while in two of them it differs. Consequently, we have to say that in Ondarroa there is evidence of incipient socio-prosodic variation (28%).

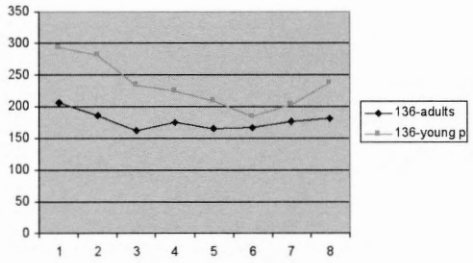


Fig. 45

Zer ikusi du alabak? [What has the daughter seen?]

7. Geo-prosodic variation

Once the phonological pattern occurring in the intonation in both localities and in both generations had been determined, it was then possible to compare the intonational patterns between them. The comparison between different patterns must be made taking into account phonological aspects alone. For that we have got the average of the data.

This kind of comparison is one of the most widespread ways of comparing two intonational patterns, as in studies related to the AMPER project (M. A. Pradilla & P. Prieto 2002; Fernández et al. 2004; Carrera et al. 2004: for Catalan: López et al. 2005; for Asturian, etc.).

As we have gathered data from two generations, we can compare these localities twice: the patterns of adults and the patterns of young people.

7.1. Geo-prosodic variation in adults

The intonation of adults from Larrabetzu and Ondarroa can now be compared. We have two groups of utterances:

a) utterances with a similar or equal intonational pattern (Fig. 4 —Larrabetzu— and Fig. 18 —Ondarroa—), and
b) utterances with a different intonational pattern: in Larrabetzu the last syllable is downward and in Ondarroa upward (Fig. 46).

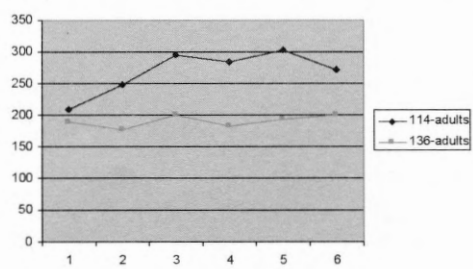


Fig. 46

Laguna sartu da? [Has the friend come in?]

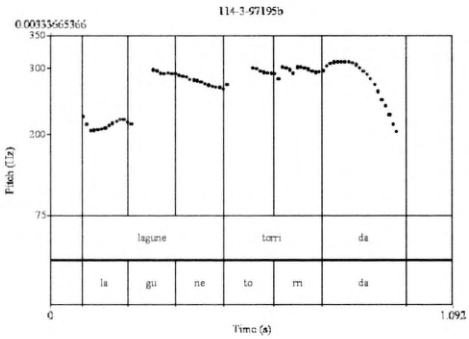


Fig. 47a

Laguna, etorri da?
[The friend, has he come?]

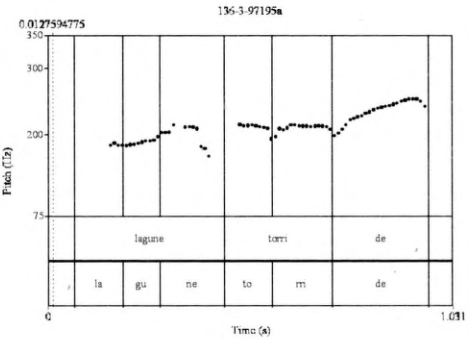


Fig. 47b

Laguna, etorri da?
[The friend, has he come?]

We have used 7 utterances: in five of them there is no geo-prosodic variation between Larrabetzu and Ondarroa, and in two of them variation was found. So, it has to be pointed out that there is 28% geo-prosodic variation between adults from Larrabetzu and Ondarroa, according to these data.

7.2. Geo-prosodic variation of young people

If we take data from young people and compare them, the following emerges:

a) In some questions there is a similar intonational pattern: utterances 97192, 97195 and 97198.

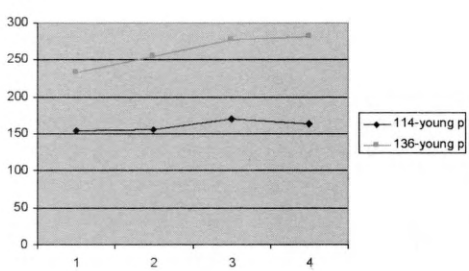


Fig. 48

Erosi du?
[Has he brought?]

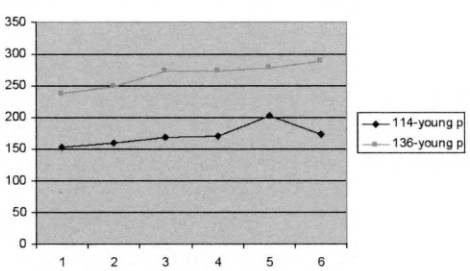


Fig. 49

Laguna sartu da? [
Has the friend come in?]

b) In others it is different: fig. 48, fig. 49, fig. 50 and fig. 51.

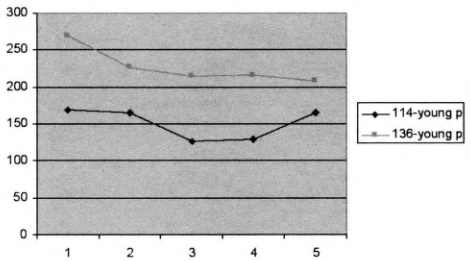


Fig. 50

Non ikusi du?
[Where has he seen it)?]

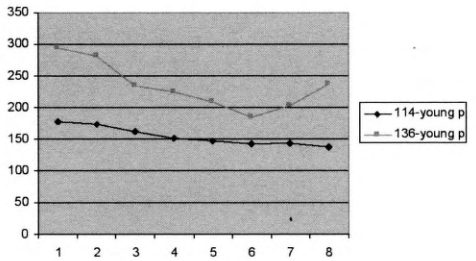


Fig. 51

Zer ikusi du alabak?
[What has the daughter seen?]

Out of the 7 sentences used, three of them exhibited no geo-prosodic variation between Larrabetzu and Ondarroa, and in the other four variation was found. So, it is possible to say that there is 57% geo-prosodic variation between young people from Larrabetzu and Ondarroa, according to these data.

Consequently, it can be said that, according to the data used in this research (bearing in mind the reduced number of sentences), when comparing Larrabetzu and Ondarroa, there is more geo-prosodic variation between young people (57%) than between adults (28%).

8. Conclusions

The technological means used in the EDAK research project in data gathering, sound annotation and labelling, and data analysing have been described.

The use of means of this type is crucial when the objective of the research is the prosodic analysis of the data.

Using this technology we have been able to compare and display the socio-prosodic variation that exists in one locality, and to compare and determine the geo-prosodic variation between two localities.

We have found socio-prosodic variation in both localities, Larrabetzu and Ondarroa. The difference between adults and young people is 43% in Larrabetzu whereas in Ondarroa it is 28%.

With regard to geo-prosodic variation between these two localities, in adults the difference is 28%, whereas in young people it is 57%; in other words, the number of differences among young people is higher than among adults.

This is the first analysis of EDAK data, the first study to approach socio- and geo-prosodic variation in the Basque language. It will require considerable additional work and that is why it cannot be regarded as the definitive study.

9. References

Aurrekoetxea, G., 2010, «Sociolinguistic and Geolinguistic Variation in the Basque language», *Slavia Centralis* III/1, 88-100.

—, Sánchez, J. & Odriozola, I., 2009, «EDAK: A Corpus to Analyse Linguistic Variation», *actas del Congreso Internacional de Lingüística del Corpus-CILC09 (Murcia 2009-05-07/09).*

Carrera, J. et al., 2004, «Les interrogatives al tortosí i al lleidatà. Un element diferenciador de subdialectes», *EFE* XIII, 157-179.

Contini, M., 1992, «Vers une géoprododie», in G. Aurrekoetxea & X. Videgain (eds.), *Nazioarteko Dialektologia Biltzarra. Agiriak,* Euskaltzaindia, Bilbao, 83-109.

—, 2005, «2e Séminaire international du projet AMPER», *Projet AMPER, Géolinguistique-Hors Série* n. 3, Centre de Dialectologie, Université Stendhal Grenoble 3, Grenoble, I-XI.

—, Romano, A. & Roullet, I. S. (forthcoming), «Vers un Atlas prosodique parlante des variétés romanes», *Mélanges en honneur de X. Ravier.*

—, Lai, J.-P., Romano, A., Roullet, S., Moutinho, L. de C., Coimbra, R. L., Bendiha, U. P. & Ruivo, S. S., 2002, «Un Projet d'Atlas Multimédia Prosodique de l'Espace Roman» a B. Bel & I. Marlien (eds.): *Proceedings of the Speech Prosody 2002 Conference, 11-13 Abril,* Aix-en-Provence : Laboratoire Parole et Langage, 227-230.

Dorta, J. & Hernández, B., 2005, «Intonation et accent dans le cadre de AMPER: déclaratives vs interrogatrives sans expansion en Tenerife et Gran Canaria», *Le projet AMPER, Geolinguistique.* Hors série, Grenoble: Centre de Dialectologie, U. Stendhal-Grenoble-3, 187-215.

Elordieta, G., 2003, «Intonation», in J. I. Hualde and J. Ortiz de Urbina (eds.), *A Grammar of Basque.* Berlin: Mouton de Gruyter.

—, 1997, «Accent, Tone, and Intonation in Lekeitio Basque», in Martínez-Gil, F. & Morales-Front, A. (eds.), *Issues in the Phonology and Morphology of the Major Iberian Languages.* Washington, D.C.: Georgetown U. P., pp. 3-78. [Revised version: Elordieta, G., 1998, «Intonation in a Pitch Accent Variety of Basque», *ASJU International Journal of Basque Linguistics and Philology* 32: 511-569.]

—, 1999, «Primer estudio comparativo de tres variedades dialectales vascas», *Actas del I. Congreso de Fonética Experimental,* Barcelona: Universitat Rovirai Virgili y Universitat de Barcelona, 209-215.

—, 2000, «Mendebaldeko intonazioaren inguruan», in, among many others, *Mendebaldeko berbetearen formalizazinoa,* Mendebalde Kultur Alkartea, Bilbo, 111-136.

—, 2002, «From pitch-accent to stress-accent in Basque» (J. I. Hualde, G. Elordieta, I. Gaminde and R. Smiljanic). In C. Gussenhoven and N. Warner, publ., *Laboratory Phonology* 7, 547-584. Berlin and New York: Mouton de Gruyter.

—, 2007a, «A constraint-based analysis of the intonational realization of focus in Northern Bizkaian Basque». In T. Riad and C. Gussenhoven, eds., *Tones and Tunes:* Volume I, *Typological Studies in Word and Sentence Prosody* (Phonology and Phonetics. Series editor: A. Lahiri), Berlin: Mouton de Gruyter, 201-234.

—, 2007b, «Constraints on Intonational Prominence of Focalized Constituents», in Ch. Lee, M. Gordon and D. Büring (eds.), *Topic and Focus: Cross-Linguistic Perspectives on Meaning and Intonation,* Amsterdam: Kluwer Academic Publishers.

—, Gaminde. I., Henáez, I., Salaberria, J. & Martín de Vidales, I., 1999, «Another step in the modelling of Basque intonation: Bermeo», in M. Matoušek, P. Mautnet, J. Ocelíková & P. Sojka (eds..), *Text, Speech and Dialogue,* Berlin: Springer-Verlag, 361-364.

Fernández Planas, A. M., Martínez, E., Carrera, J., Oosterzee, C. van, Salcioli, C., Castellvi, J. & Szmidt, S., 2004, «Interrogatives absolutes al barceloní i al tarragoní (estudi contrastiu)», *EFE* XIII, 2004, 129-155.

—, Martínez, E., Carrera, J., Oosterzee, C. van, Salcioli, C., Castellvi, J. & Szmidt, S., 2007, «Proyecto AMPER: estudio contrastivo de frases interrogativas sin expansión del barceloní y del tarragoní», *Actas del VI Congreso de Lingüística General,* University of Santiago de Compostela, 1931-1944.

Fernández Rei, E., González, M., Xuncal, L. & Camaño, M., 2005, «Achega á entonacion dunha fala do centro de Galicia. Contribució para o Altas Multimèdia Prosodique de L'Espace Roman», *Le projet AMPER, Géolinguistique, Hors série,* Grenoble: Centre de Dialectologie, U. Stendhal Grenoble 3, 87-102.

Gaminde, I., 1995, *Bizkaieraren azentu-moldeez,* Labayru Ikastegia, Bilbo.

—, 2001, «Azentua eta intonazioa. Egoera eta ikerketa baliabideak», in K. Zuazo (ed.), *Dialektologia gaiak,* UPV/EHU-Arabako Foru Aldundia, 263-286.

—, 2004, «Tonuak eta etenak Gatikako intonazioan», *FLV* 97, 519-536.

—, 2006, «Intonazio kurben etenez», in J. Lakarra and J. I. Hualde (eds.), *Studies in Basque and Historical Linguistics in Memory of R. L. Trask. R. L. Trasken oroitzapenetan ikerketak euskalaritza eta hizkuntzalaritza historikoaz, ASJU XL:* 1-2, 351-376.

—, 2007, *Bizkaian zehar euskararen ikuspegi orokorra,* Izaro bilduma VI, Mendebalde Kultura Alkartea, Bilbo.

—, 2010, *Bizkaiko Gazteen Prosodiaz: Euskaraz eta Gaztelaniaz,* Mendebalde Kultura Alkartea, Bilbo.

Hualde, J. I., 1989, «Acentos vizcaínos», *ASJU* 23, 275-325.

—, 1990, «Euskal azentuaren inguruan», *ASJU* 24, 699-720.

—, 1994, «Euskal azentu ereduen sailkapenerako», *Euskaltzaindiaren XIII Biltzarra, Euskera* 39-3, 1569-1578.

—, 1997, *Euskararen azentuerak,* ASJUren Gehigarriak, Donostia-San Sebastian.

—, 2002, «From pitch-accent to stress-accent in Basque». In *Laboratory Phonology VII,* ed. by C. Gussenhoven & N. Warner, Berlin: Mouton de Gruyter, 547-584.

—, 2003a, «El modelo métrico y autosegmental», in Prieto, P. (coord.), *Teorías de la entonación,* Ariel Lingüística, Barcelona.

—, 2003b,«Peak alignment and intonational change in Basque». In *Proceedings of the 15th International Congress on the Phonetic Sciences Barcelona 3-9 August 2003,* ed. by M. J. Solé, D. Recasens & J. Romero. Co-authored: K. Ito, G. Elordieta & J. I. Hualde. 2929-2932.

— & Bilbao, X., 1992, «A Phonological Study of the Basque Dialect of Getxo», *ASJU* XXVI-1, 1-118.

Jun, S. A. & Elordieta, G., 1997, «Intonational Structure of Lekeitio Basque». In Botinis, A., G. Kouroupetroglou & G. Carayiannis (eds.), *Intonation: Theory, Models and Applications, Proceedings of an ESCA Workshop,* Athens, 193-196.

López Bobo, M. J., González, R., Cuevas, M., Díaz, L. & Muñiz, C., 2005, «Rasgos prosódicos del centro de Asturias: comparación Oviedo-Mieres», *EFE* 14, 167-199.

Martínez Celdrán, E., Fernández Planas, A. M., Salcioli Guidi, V., Carrera Sabaté, J. & Espuny Monserrat, J., 2005, «Approche de la prosodie du dialecte de Barcelona». *Géolinguistique. Hors série.* 153-175.

Mitxelena, L., 1958, «À propos de l'accent basque», *BSL* 53, 204-220 [now also in *SHLV* I, 220-239].

—, 1972, «A note on Old Labourdin Accentuation», *ASJU* 6, 110-120 [now also in *PT,* 235-344].

—, 1976, «Acentuación alto-navarra», *FLV* 23, 147-162.

Pradilla, M. I. & Prieto, P., 2002, «Variación entonativa catalana: catalán central versus tortosino». *Actas del II Congreso de Fonética Experimental,* Sevilla: Laboratorio de Fonética, Facultad de Filología, Universidad de Sevilla, 291-295.

Prieto, P., 1998, «L'entonació dialectal del català: el cas de les frases interrogatives absolutes». In A. Bover & M.-R. Lloret i M. Vidal-Tibbits (eds.), *Actes del Novè Col.loqui d'Estudis Catalanas a Nord-Amèrica,*. Barcelona, PAM, 347-377.

—, 2002, *Entonació. Models, teoria, mètodes.* Editorial Ariel: Barcelona.

Romano, A., 2001, «Un projet d'Atlas multimédia de l'espace roman (AMPER)», *Actas del XXIII Congreso Internacional de Lingüística y Filología Románica,* Salamanca, University of Salamanca, 279-294.

Txillardegi, 1984, *Euskal Azentuaz,* Elkar, Donostia-San Sebastián.

Van Oosterzee, C., Fernández Planas, A. M., Romera Barrios, L., Carrera Sabaté, J., Espuny Monserrat, J. & Martínez Celdrán, E., 2007, «Proyecto AMPER: estudio contrastivo de frases interrogativas sin expansión en *tortosí* y en *lleidatà», Actas del VI Congreso de Lingüística General,* University of Santiago de Compostela, 1977-1990.

XXXI. KARLOS OTEGI, *Lizardi: lectura semiótica de "Biotz-begietan"*, 1993. 18 €.

XXXII. AURELIA ARKOTXA, *Imaginaire et poésie dans "Maldan behera" de Gabriel Aresti (1933-1975)*, 1993. 18 €.

XXXIII. MANUEL AGUD - ANTONIO TOVAR, *Diccionario etimológico vasco, VI. Ilpiztu-Korotz*, 1993. 8 €.

XXXIV. JOSÉ I. HUALDE - GORKA ELORDIETA - ARANTZAZU ELORDIETA, *The Basque dialect of Lekeitio*, 1994. 18 €.

XXXV. GEORGES REBUSCHI, *Essais de linguistique basque*, 1997. 18 €.

XXXVI. XABIER ARTIAGOITIA, *Verbal projections in Basque and minimal structure*, 1994. 12 €.

XXXVII. MANUEL AGUD - ANTONIO TOVAR, *Diccionario etimológico vasco, VII. Korpa-Orloi*, 1994. 8 €.

XXXVIII. PATXI GOENAGA (ed.), *De grammatica generativa*, 1995. 18 €.

XXXIX. ANTONIO CID, *Romancero y balada oral vasca. (Literatura, historia, significado)*. En preparación.

XL. AMAIA MENDIKOETXEA - MYRIAM URIBE-ETXEBARRIA (eds.), *Theoretical issues at the morphology-syntax interface*, 1997. 21 €.

XLI. BERNARD HURCH - MARÍA JOSÉ KEREJETA, *Hugo Schuchardt - Julio de Urquijo: Correspondencia (1906-1927)*, 1997. 21 €.

XLII. JOSÉ I. HUALDE, *Euskararen azentuerak*, 1997. 15 €.

XLIII. RUDOLF P. G. de RIJK, *De lingua Vasconum: Selected Writings*, 1998. 15 €.

XLIV. XABIER ARTIAGOITIA - PATXI GOENAGA - JOSEBA A. LAKARRA (arg./eds.), *Erramu Boneta: Festschrift for Rudolf P. G. de Rijk*, 2002. 30 €.

XLV. JOSEBA A. LAKARRA, *Ikerketak euskararen historiaz eta euskal filologiaz*. Argitaratzeko.

XLVI. BEÑAT OYHARÇABAL, *Inquiries into the lexicon-syntax relations in Basque*, 2003. 18 €.

XLVII. BLANCA URGELL, *Larramendiren "Hiztegi Hirukoitza"-ren Eranskina: saio bat hiztegigintzaren testukritikaz*. Argitaratzeko.

XLVIII. IÑIGO RUIZ ARZALLUZ, *"Aitorkizunen" historia eta testua: Orixeren eskuizkributik Lekuonaren ediziora*, 2003. 21 €.

XLIX. GOTZON AURREKOETXEA - XARLES VIDEGAIN (arg.), *Haur prodigoaren parabola Ipar Euskal Herriko 150 bertsiotan*, 2004. 21 €.

L. JOSEBA A. LAKARRA, *Raíz y reconstrucción del protovasco*. En prensa.

LI. XABIER ARTIAGOITIA - JOSEBA A. LAKARRA (arg.), *Gramatika Jaietan. Patxi Goenagaren omenez*. 36 €.

LII. BEATRIZ FERNÁNDEZ - PABLO ALBIZU - RICARDO ETXEPARE (arg.), *Euskara eta euskarak: aldakortasun sintaktikoa aztergai*. 18 €.

LIII. GOTZON AURREKOETXEA - JOSE LUIS ORMAETXEA (eds.), *Tools for Linguistic Variation*. 18 €.