
X. Sousa Fernández (2004): "A base de datos do *Atlas Lingüístico Galego*", en R. Álvarez / F. Fernández Rei / A. Santamarina (eds.): *A lingua galega: historia e actualidade.* Santiago de Compostela: Consello da Cultura Galega / Instituto da Lingua Galega, vol. 2, 637-647.

A lingua galega, historia e actualidade. Actas do I congreso internacional:
Consello da Cultura Galega, Instituto da Lingua Galega, 2004: 637-647

A BASE DE DATOS DO *ATLAS LINGÜÍSTICO GALEGO*

Xulio C. Sousa Fernández
Instituto da Lingua Galega. Universidade de Santiago de Compostela

Confeccionar un atlas lingüístico é unha empresa cobizada por calquera filólogo, pero é tamén unha tarefa custosa, de anos e que precisa de moita constancia e paciencia. Todos sabemos de proxectos de atlas sen rematar ou que levan décadas suspendidos, e non pola neglixencia dos investigadores, senón por causas que pouco teñen que ver cos seus azos. No caso do *Atlas Lingüístico Galego* (*ALGa*), e a pesar de tódolos atrancos, xa hoxe se poden ve-los primeiros froitos do proxecto.

No ano 1976 os investigadores encargados de realiza-las enquisas de campo remataban o seu percorrido polo territorio de fala galega con 167 cadernos de campo ateigados de información que a partir daquela comezou a ser empregada para fins moi diversos. Dos cadernos do *ALGa* tiráronse datos moi valiosos para a elaboración das normas ortográficas e morfolóxicas, tamén se aproveitou material das enquisas para a realización de traballos de dialectoloxía e lexicografía. Pero fóra desta porfillaxe menor, foi na década na que andamos cando apareceron os primeiros resultados propios do proxecto[1].

A elaboración dos dous volumes xa publicados foi case artesanal[2]. Para o primeiro volume a información dos cuestionarios pasouse ós mapas manualmente, para o segundo utilizouse un programa de deseño gráfico que alixeirou a corrección dos mapas e perfeccionou a súa aparencia. Durante a redacción do último volume publicado, e despois de ter noticia doutros proxectos semellantes, comezou a pensarse na necesidade de informatiza-lo *ALGa*. Por unha parte, cumpría automatiza-lo proceso de confección dos mapas para facelo máis áxil, mais tamén era tarefa urxente buscarlle un novo soporte á

[1] Para unha descrición do proxecto inicial do *ALGa* pódese ver: Constantino García / Antón Santamarina / Rosario Álvarez Blanco / Francisco Fernández Rei / Manuel González González, "O Atlas Lingüístico Galego", *Verba* 4, 1977, pp. 5-17.

[2] Instituto da Lingua Galega, *Atlas lingüístico galego.* Vol. 1: *Morfoloxía verbal*. A Coruña, Fundación Barrié de la Maza, 1990; e Instituto da Lingua Galega, *Atlas lingüístico galego.* Vol. 2: *Morfoloxía non verbal*. A Coruña, Fundación Barrié de la Maza, 1995.

información das enquisas que asegurase a súa conservación e permitise un mellor aproveitamento.

OS CUESTIONARIOS

O material que se ía tratar estaba contido nos cuestionarios utilizados polos enquisadores. Os cuestionarios son os cadernos nos que se foron anotando as respostas dos entrevistados. De cada un dos 167 puntos que constitúen a rede do *ALGa* existe un caderno coa transcrición das 2712 respostas correspondentes. As preguntas repártense no cuestionario en catro grupos: fonética (1-148), morfoloxía (149-386), sintaxe (387-526) e léxico (527-2712). As respostas dos informantes transcribíronse utilizando o alfabeto fonético que se empregara no *Atlas Lingüístico y Etnográfico de Andalucía*, con algún engadido. Nas páxinas pares complementarias á pregunta: explicacións, debuxos, refráns, referencias a obxectos dos cadernos anotáronse informacións relacionados, contextos nos que se emprega a palabra, etc. Facendo unha cálculo simple, contabamos que a base de datos que contivese a información dos cadernos había de ter preto de medio millón de rexistros.

O primeiro labor no desenvolvemento do proxecto foi procurar un programa de tratamento de datos que cubrise as nosas necesidades. Debía ter capacidade para tratar un número de rexistros non inferior ó medio millón, e permitir que a introducción e consulta de datos resultase doada. Ademais, tiña que ser posible que nos campos do rexistro se puidese varia-lo formato da fonte, de xeito que fose posible transcribir en alfabeto fonético as respostas dos cuestionarios. Esta última condición foi a que dificultou máis a busca do programa. En proxectos semellantes que daquela coñeciamos, adoptarase a solución de crear un alfabeto convencional propio, que nalgúns casos se asemellaba un pouco ó alfabeto fonético internacional[3].

Hoxe calquera base de datos relacional que traballe co sistema Windows cumpre estes requisitos, pero naquel momento aínda eran moi poucos os programas deste tipo que funcionaban nun sistema gráfico, e menos os que deixaban deseñar un formulario de entrada de datos con distintos formatos de fonte. O programa que finalmente escollemos foi Microsoft Access. Este programa, amais de cumprir tódalas nosas esixencias, permite a organización da información en distintas bases relacionadas, polo que é posible separa-los datos constantes para cada cuestionario ou comúns para moitos rexistros (localización, número e texto das preguntas, caracterización semántica e gramatical das

[3] Exemplos da complexidade deste labor poden verse en Roland Bauer / Silvio Gislimberti / Elisabetta Perini / Tino Szekely / Hans Goebl, "Arbeitsbericht 3 zum ALD I", *Ladinia*, nº 12, 1988, pp. 17-56; W. A. Kretzschmar, "Phonetic Display and Output", *Journal of English Linguistic*, 22.1, 1989, pp. 47-53; e Lawrence M. Davis / Charles L. Houck / Brian B. Kelly, "Easily Manipulatable Databases for Linguistic Atlases: Using Intergraph", *Journal of English Linguistic*, 22.1, 1989, pp. 30-39.

preguntas, etc.) da información variable das respostas. Os dous compoñentes do programa sobre os que baseámo-la organización do proxecto son os que reciben o nome de *táboa* e *formulario*. A *táboa* é unha peza esencial da base de datos, nela determínase o número de campos e as súas características. Na *táboa* almacénase nun único formato toda a información dos rexistros. Calquera operación e aplicación do programa (buscas, ordenacións, substitucións, etc.) ten como orixe este módulo. O *formulario* é unha aplicación que serve para mostra-la información e tamén para realizar operacións ligadas a unha *táboa* (introducción e consulta de datos, apertura doutras aplicacións, etc.). Ó se-lo *formulario* esencialmente un módulo de presentación, as posibilidades de personaliza-lo deseño son moito maiores: podemos ordena-los campos segundo as nosas necesidades e mesmo definir un formato independente para cada campo. Nun mesmo formulario é posible ter á vista rexistros que pertencen a táboas distintas.

O DESEÑO DA BASE DE DATOS DO *ALGA* (BDALG)

O deseño da base de datos actual, resultado xa de moitas modificacións e engadidos, xustifícase polos obxectivos iniciais para os que foi creada. O núcleo principal da súa estructura constitúese como soporte de toda a información dos cuestionarios. A outra parte é en realidade un engadido utilizado para tirar da base o seu primeiro e principal rendemento, a elaboración dos mapas do *ALGa*.

No menú de entrada á base están distinguidas estas dúas partes. A principal está constituída por tres formularios, vinculados ás tres respectivas táboas. A cada un destes formularios déuselle un nome de acordo co seu contido principal: *Puntos*, *Cuestionario* e *Respostas*. *Puntos* e *Cuestionario* foron os que primeiro se crearon, por seren os máis pequenos e máis tamén por funcionaren como formularios de contido básico respecto de *Respostas*. Vexámo-las características de cada un deles.

1. Puntos

No formulario *Puntos* introducíronse os datos da localización xeográfica de cada un dos lugares que forman a rede do *ALGa*: clave do punto (a letra ou letras iniciais da provincia separadas por un guión do número asignado), o nome do concello, o nome da parroquia e mailo nome do lugar en que se realizou a enquisa. Estes son os datos que identifican cada un dos cuestionarios. Unha vez introducida, revisada e corrixida, a información da táboa serve para identificar cada un dos puntos, e ó vinculala á táboa *Respostas* é posible coñece-la localización de cada resposta. A táboa correspondente a este cuestionario conta con dous campos máis, os das coordenadas que serven para situar no mapa os símbolos atribuídos ás respostas. En realidade existen catro pares de coordenadas asociadas a cada punto, un par por cada tipo de resposta diferente para a mesma pregunta (A, B, C e D). A utilidade destes datos aparece explicada máis abaixo. No

formulario vinculado a esta táboa non se mostran os dous campos que sinalan as coordenadas (Figura 1).

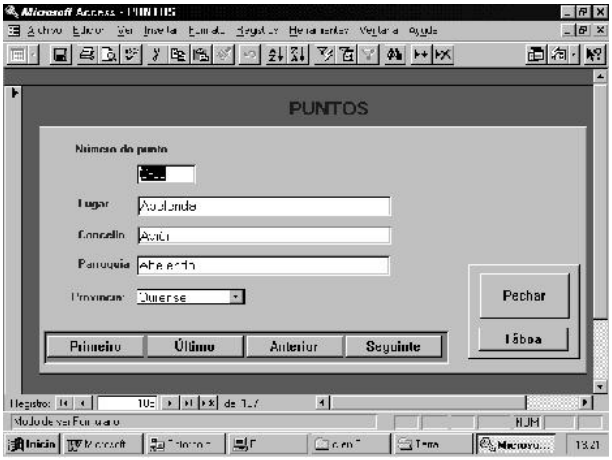

Figura 1, Formulario *Puntos*

2. Cuestionario

En *Cuestionario* fixouse o texto base dos cadernos: o número e o enunciado da pregunta. Inseríronse dous novos campos que non aparecen no cuestionario orixinal: un coa información sobre a categoría léxica da palabra e outro con información semántica. O número final de rexistros desta táboa non coincide co número total de preguntas dos cadernos de campo; en moitos casos foi preciso dispoñer novos rexistros para datos complementarios. Por exemplo, a partir de datos da pregunta 2508 *Entroido*, creáronse catro novos espacios para cada un dos días en que se celebran as festas do entroido (2508,1 domingo de entroido; 2508,2 luns de entroido, 2508,3 martes de entroido e 2508,4 mércores de entroido).

O campo *categoría* non ten correspondente nos cadernos; a selección da categoría léxica de cada pregunta foi traballo dos manipuladores. O espacio reservado para a información semántica tampouco tiña un correlato exacto no caderno, aínda que si un antecedente. Como sinalei antes, as preguntas distribúense no cuestionario en catro seccións, e na sección do léxico distínguense dezasete grupos semánticos[4]. Ó deseñármo-la base de datos coidamos conveniente conservar unicamente a división semántica, as outras son prescindibles, pois a información sempre sería recuperable a través da numeración das preguntas. As 2186 preguntas sobre léxico máis as 526 das primeiras seccións foron clasificadas semanticamente empregando un sistema máis sinxelo, o

[4] As divisións desta sección teñen as seguintes cabeceiras: 1. O tempo; 2. Accidentes topográficos: o espacio e as súas relacións; 3. Agricultura; 4. Viño, aceite, fariña, panificación; 5. Plantas; 6. Insectos, aves, animais salvaxes. 7. Pesca e peixes, caza; 8. Vida pastoril; 9. Animais domésticos; 10. A casa, ocupacións domésticas; 11. O corpo humano; movementos e accións; 12. O vestido; 13. A familia, a vida humana; 14. O mundo espiritual; 15. Xogos e diversións; 16. Oficios; 17. Pesas e medidas.

mesmo que se estaba a seguir no proxecto do Ficheiro do léxico oral[5]. Os campos nocionais que se distinguiron son:

A-1 O ceo e a atmosfera
A-2 A terra
A-3 As plantas
A-4 Os animais
B-1 O home ser físico
B-2 A alma e o intelecto
B-3 O home ser social
B-4 A organización social
C-1 O apriori (calidades, existencia, etc.)

Para a introducción dos datos nesta táboa deseñouse un formulario con controis de selección que axudaron a simplifica-lo labor de clasificación dos datos. (Figura 2).

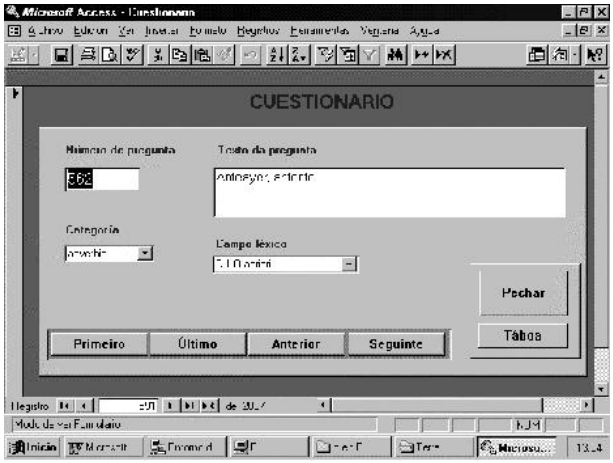

Figura 2, Formulario *Cuestionario*

3. Respostas

A táboa *Respostas* é a principal da BDALG. O deseño desta táboa estableceuse considerando que o formulario dependente dela ía se-lo elemento da BDALG máis empregado polos manipuladores (Figura 4). Este formulario está vinculado ós anteriores: o campo *número de punto* está relacionado coa táboa *Puntos* e os campos *número da pregunta* e *texto da pregunta* están ligados á táboa *Cuestionario*. O resto das compoñentes son propias da táboa *Respostas*: *resposta fonética*, *resposta ortográfica*, *notas*, *tipo*, *xénero* e *número* (Figura 3).

[5] Para unha descrición polo miúdo véxase Manuel González González, "Os traballos lexicográficos no Instituto da Lingua Galega*", Actas do XIX Congreso Internacional de Lingüística e Filoloxía Románicas*, vol. VIII, A Coruña, Fundación Pedro Barrié de la Maza, 1996, pp. 771-786.

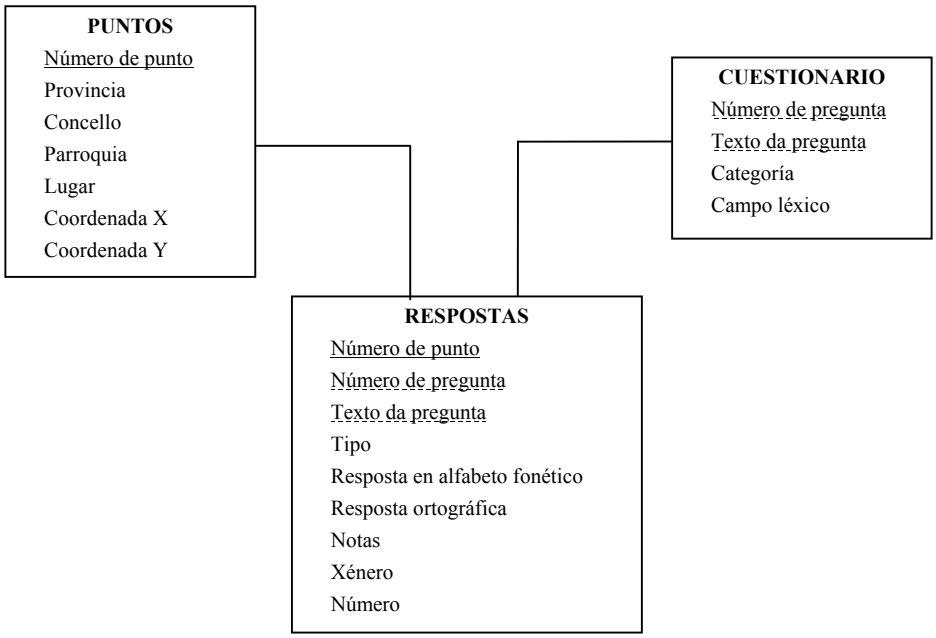

Figura 3, Vínculos establecidos entre as táboas que compoñen a BDALG

O formulario *Respostas* emprégase para introduci-la información dos cadernos na base de datos e mais para facer correccións (Figura 4). Os elementos que constitúen o formulario *Respostas* son os que seguen:

- *Número de punto*: este campo está vinculado á táboa *Puntos* e nel indícase de xeito abreviado o punto ó que corresponde a información do rexistro.
- *Número da pregunta, Texto da pregunta* e *Tipo*: estes campos están relacionados coa táboa *Cuestionario*, de xeito que ó teclea-lo número da resposta que se está copiando aparece automaticamente no recadro o texto da pregunta. O campo *Tipo* emprégase cando existen varias respostas para unha mesma pregunta: a primeira resposta é *Tipo A*, a segunda *Tipo B*, e así sucesivamente ata un máximo de catro respostas distintas. Cada resposta distinta é sempre un novo rexistro.
- *Resposta fonética* e *Resposta ortográfica*: nestes espacios recóllese o texto da resposta do cuestionario; en *fonética* cos caracteres da fonte IPA e en *ortográfica* con grafía convencional. A fonte tipográfica selecciónase automaticamente ó entrar no campo. Para a transcrición fonética os manipuladores dispoñen dun cadro de equivalencias entre o alfabeto utilizado nos cuestionarios e a fonte IPA.
- *Notas*: neste espacio fanse anotacións complementarias á pregunta principal que nos cadernos acostuman aparecer nas marxes ou nas páxinas pares.

- *Información gramatical*: este último espacio está composto por catro campos de selección (masculino, feminino, singular e plural) que serven para clasificar gramaticalmente as categorías léxicas que na resposta presenten esas flexións.

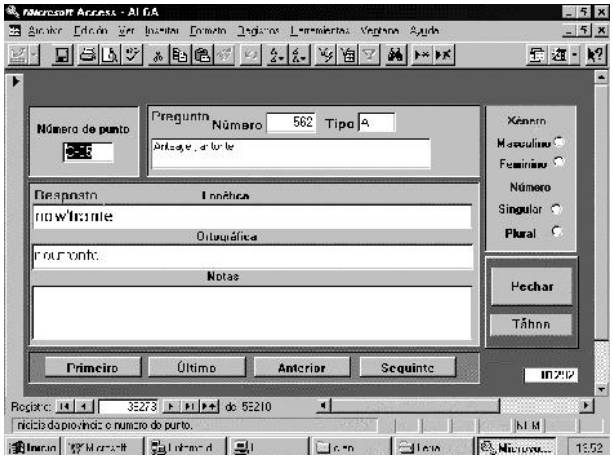

Figura 4, Formulario *Respostas*

A INTRODUCCIÓN DOS DATOS

Para a introducción dos datos dos cuestionarios na BDALG acordouse distribuí-lo traballo por provincias. O método máis proveitoso e lóxico sería proceder pregunta por pregunta; aínda que nun principio foi considerado, axiña o rexeitamos polas dificultades que supoñía no manexo dos cadernos. Nestes momentos cada un dos manipuladores encárgase da transcrición de tódolos cadernos dunha provincia, ó remate desta fase engadiranse os datos dos 15 puntos situados fóra do territorio administrativo galego.

O formulario que se emprega para a introducción de datos é o denominado *Respostas*. Para comeza-la transcrición dun caderno o manipulador debe abri-lo formulario desde o módulo principal (Figura 5). O primeiro campo que debe cubrir é o do número de punto; ó inicia-lo traballo cun cuestionario fíxase un valor predeterminado neste espacio, deste xeito nos seguintes rexistros non será preciso enchelo. A continuación debe escribi-lo número da pregunta, despois desta acción aparece no recadro inferior o texto da pregunta segundo reza no cuestionario. Os dous seguintes espacios que cómpre completar son os correspondentes á resposta. Se é pertinente deben cubrirse tamén os valores das categorías gramaticais. O valor *tipo* móstrase sempre por defecto coa letra A; no caso de que para a mesma pregunta se dea no cuestionario máis dunha resposta será preciso abrir un novo rexistro coa mesma numeración pero asignándolle un tipo

diferente (B, C ou D). No recadro *notas* apúntanse tódalas informacións adicionais que atanguen á pregunta, no primeiro rexistro se fan referencia a tódalas respostas ou no correspondente se pertencen a calquera das outras.

Na actualidade aínda non se rematou o traballo de introducción dos datos. A distribución do traballo dos tomos do *ALGa* correspondentes ó léxico obrigou a introduci-las preguntas por grupos temáticos. Os redactores e directores do proxecto do *ALGa* dividiron as preguntas en tres grandes grupos: *Terra*, *Agricultura* e *Home*. As preguntas do grupo *Terra* están hoxe xa todas introducidas na BDALG, dos outros dous grupos introduciuse o 50%. A supervisión e revisión do traballo correspondente a cada grupo está da man dos coordinadores dos próximos tomos do *ALGa*, os profesores: Manuel González González, Rosario Álvarez Blanco e Francisco Fernández Rei.

O DESENVOLVEMENTO DA BDALG

Como ben se pode observar na descrición precedente, o deseño da base de datos actual é moi elemental. A explicación desta sinxeleza está, como apuntei antes, na necesidade perentoria e primaria de asegura-la conservación dos datos que ata hoxe envellecían e perdían cor nos cuestionarios. Os encargados do proxecto esperamos no futuro arrequentar e completa-lo deseño da BDALG con novos engadidos que a perfeccionen: inclusión de campos gráficos para os debuxos e fotos, introducción de novos campos con información tirada da análise das respostas (segmentación morfolóxica, caracterización semántica detallada, vinculación de respostas relacionadas, etc.), desenvolvemento e engadido de informacións provenientes doutras fontes (etimoloxía, sinonimia, referencias bibliográficas, etc.), etc. Tamén forma parte do proxecto a automatización dos volumes publicados para a edición de toda a obra en CD-ROM e o desenvolvemento de aplicacións que permitan a consulta da DBALG e do *ALGa* a través das redes de datos[6].

A CONFECCIÓN DE MAPAS DO *ALGA*

Como exemplo do primeiro rendemento que se está a tirar da BDALG describirei de xeito breve as aplicacións desenvoltas para a confección dos mapas do *ALGa*.

Na parte esquerda do menú principal da BDALG sitúanse catro botóns de comando (Figura 5). Os dous botóns superiores (*Listado por punto* e *Listado por pregunta*) utilí-

[6] Un exemplo desta aplicación pode consultarse en *The Empirical Linguistics and Linguistic Atlas Page* (http://hyde.park.uga.edu/home.html), proxecto dirixido polo profesor William A. Kretzschmar, Jr. Nesta páxina poden examinarse os datos e os mapas de distintos atlas lingüísticos dos Estados Unidos (LANE, LAMSAS, LANCS, LAGS, LAUM, LAO, LAPW, LAPNW, LARMS).

zanse para crear informes impresos dos cuestionarios. Estes informes empréganse como borradores para corrección e como primeiras guías para a confección dos mapas (selección de formas que se van representar, cómputo das variantes de cada mapa, etc.). Se picamos *Listado por punto* conseguimos unha relación de tódalas respostas do punto escollido; ó seleccionarmos *Listado por grupo* obtemos un repertorio de tódalas respostas diferentes e ó seu carón o número de veces que aparecen; con *Listado por pregunta* temos como resultado un catálogo de tódalas repostas á mesma pregunta nos distintos puntos enquisados.

Figura 5, Menú principal

Os outros dous botóns son os que se usan para confeccionar un mapa. Ó seleccionar *Mapa* aparece unha ventá na que se solicita o número da pregunta que se desexa cartografar. Despois de teclea-lo número e confirma-la acción execútase unha consulta que ten como resultado a creación dunha táboa. Esta nova táboa contén datos de dúas táboas descritas arriba: número da pregunta, clave do punto, resposta en alfabeto fonético, tipo da resposta e coordenadas de cada resposta (Figura 6).

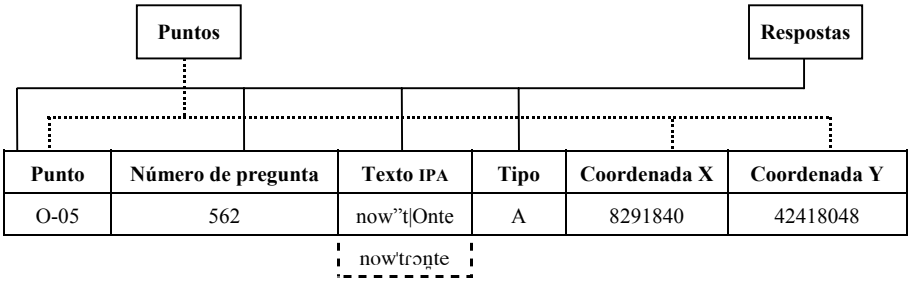

Figura 6, Táboa creada para a realización dun mapa.
O formato do texto de tódolos rexistros da táboa é o mesmo; por iso no campo *Texto IPA* o que se le é a correspondencia na fonte Times New Roman da transcrición fonética anotada debaixo.

Automaticamente, esta relación de datos convértese a un formato compatible co programa que creará o mapa; para abrilo abonda con pica-lo botón inferior da columna esquerda. O motor que se encarga da asignación de datos a puntos xeográficos nun mapa é un programa de *GIS* (*Geographic Information Systems*), en concreto o sistema creado por MapInfo[7]. Este é un programa de tratamento de datos xeográficos que converte a información dun campo da base de datos en elementos gráficos de representación sobre un fondo gráfico. Para realizar esta función precísase un espacio xeográfico delimitado por coordenadas (neste caso, o territorio lingüístico galego) e unha base de datos na que cada rexistro conteña a información correspondente ó par de coordenadas. O sistema de representación que nós utilizamos é o temático: a cada rexistro diferente do arquivo asígnaselle un símbolo gráfico diferenciado ben pola forma, ben pola cor. Unha vez aberto o programa, cárgase a base de datos, sitúanse os puntos no fondo gráfico e convértense as respostas en símbolos de representación. O resultado final é un mapa coa lenda das atribucións de cada un dos símbolos empregados. Tanto o mapa como a lenda poden ser editados para realizar correccións e modificacións (Figura 7).

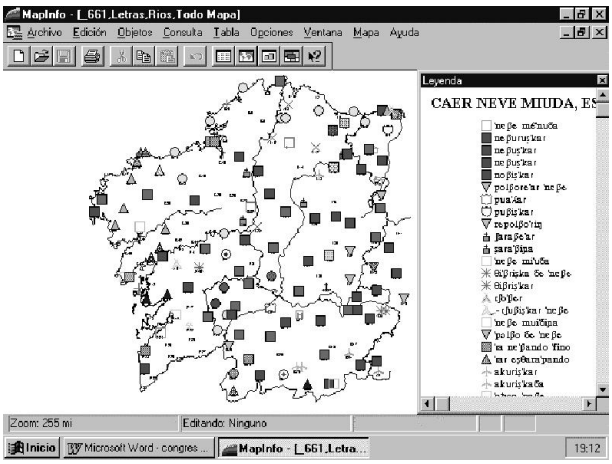

Figura 7, Exemplo do mapa creado con MapInfo para a pregunta 661 "Caer neve miúda"

O programa ofrece outras posibilidades de representación dos datos amais da temática. Por exemplo, cunha base de mapa en mosaico poligonal será posible crear mapas cos que realizar estudios de dialectometría[8]; cambiando o símbolo de representación por unha lenda poden facerse mapas semellantes ós dos primeiros estudios de xeografía

[7] Aínda que non coñecemos nin o deseño nin o desenvolvemento, sabemos da existencia doutros proxectos nos que se está a utiliza-lo mesmo sistema: o *Atlas Lingüístico del Caribe*, desenvolto na Universidade de Saint Rose, Albany, N. Y., e un atlas do inglés medieval dirixido por Keith Williamson dentro do *Older Scots Project* do Institute for Historical Dialectology de Edimburgo.

[8] Un exemplo da aplicación deste tipo de estudios ó *ALGa* atópase en Bernhard Pöll, "Zur Verbreitung zweier Analogiephänomene des galiscischen Verbums anhand des ALGa (Atlas Lingüístico Galego)", in *Sprache, Literatur und Kultur Galiciens*, Serie 1, tomo IV, 1993, pp.35-52.

lingüística, nos que ó lado dos puntos se colocan as respostas en transcrición fonética. Tamén permite o programa que sobre o mapa dos datos lingüísticos se superpoñan mapas que conteñan outro tipo de informacións (divisións eclesiásticas, orografía, poboamentos, migracións, datos toponímicos, etc.).

Nin a BDALG nin o *ALGa* son proxectos pechados. Cando se remate coa introducción dos datos poderase falar dunha etapa concluída, pero o traballo de aproveitamento e interpretación da información e o desenvolvemento de novas aplicacións continuarán por bastante tempo.