Hans Geisler

Dialect Data Processing with Personal Computers

With the arrival of efficient personal computers, high resolution graphics cards and laser printers in the last few years it has become possible to create a special environment for dialect data processing without having to buy additional equipment which is expensive and hard to manage. The environment which we want to present consists of several independent programs which have had to be newly devised and programmed.[1] In addition, some third party programs can be used to deal with standard situations involved in dialect data processing (see table 1).

I) **Dialect-PC** is a totally menu-driven program which incorporates a set of utilities to make available some dormant hardware capacities of a PC and a laser printer for dialect data processing. The utilities have been designed to store transcribed data in a structured data-base-like form, to select data with a query language based on linguistic concepts and to show the topographic distribution of the data by printing dialect maps with raw data, symbols or isoglosses.

II) **Font-Editors** are necessary to create all the special characters to be used by various transcription systems. Unfortunately there are no readily available fonts for these systems and the missing characters and diacritics have to be designed by the dialectologist. For a laser printer there are already a lot of well-established programs like Publisher's Type Foundry or Font Solution Pack which help to create new characters for a high resolution proportional font.

III) **Word Processing:** A specially adapted version of the word processor Word Perfect with variable font integration, keyboard remapping and key macros is used to incorporate the transcribed dialect data in a very elaborate word processing environment (with all sorts of data input and formatting facilities like overstriking, templates, merging and printer management; see tables 1-6 and composite characters done with this version).

IV) **Data conversion:** In order to allow free data interchange between the word processor and Dialect-PC a program has to be devised to convert the data encoded in the word processor format to a format used by Dialect-PC (and vice versa). With this program it is also possible to make changes in the transcription systems used with the data: for example, to reduce a very narrow transcription to a wide transcription or to switch to another transcription system. If data from different sources with variable transcriptions are used,

[1] The programs are written in Turbo Pascal and Assembler by Dieter Strehle.

ZDL-Beiheft 74: Verhandlungen des Internationalen Dialektologenkongresses Bamberg 1990. Band 1. Wolfgang Viereck (Hrsg.).

this is an obligatory task for getting data consistency for queries (a problem which is also latent with traditional work on dialects; see map TPPSR with original transcription and conversion to IPA). The conversion is based on equivalence tables between transcription systems (which are sometimes very problematical to establish, see table 5).

All the programs run on a normal PC-configuration with an EGA or VGA-card and dialect maps can be produced with a Laser printer (HP Laserjet). A 24-dot-matrix printer (NEC P 6) can be used, but the representation of transcription systems with multi-level diacritics has given unsatisfactory results.

The following is a short account of some problems with dialect data processing and how they are solved with Dialect-PC (see the program structure in table 2).

1. Transcription system

The transcription system for the data used with Dialect-PC can be freely chosen if the corresponding screen and laser fonts are available. At the moment there is a screen and printer font for a subset of the revised IPA-alphabet as represented by the 1989 Kiel convention (see *Journal of the International Phonetic Association* 19/1, 1989: 67-80) which should be sufficient for a description of most of the European languages (in accordance with the IPA-Font used by the *JIPA* the characters are based on a Times Roman 10pt proportional font, see table 3 for diacritics and suprasegmentals). There are also screen and printer fonts for the Rousselot-Gilliéron and Ascoli type transcription systems which are very widespread in Romance dialectology (see, for example, the transcription used with the *ALF, ALF par régions* and the *AIS*; the fonts are based on a Times Roman Italic 10pt font).

Whereas the IPA-system uses a relatively high number of different basic characters and diacritics are not abundant, the older systems are characterized by an extensive use of diacritics and superposed characters (denoting 'intermediate sounds'). It is even possible to combine diacritics to build composed characters with 5 levels. The famous *ALF* shows more than a thousand different types, which has been a great challenge for type-setters (see table 4 for some combinations of *e* + diacritics and superposed vowels). The *ALF par régions* have reduced this complicated system to make transcriptions more manageable. Using the built-in 8-bit code of PC's with only a restricted number of 256 code values there is no means of encoding all possible combinations in a 1:1 relation. For this reason all diacritics and superposed characters are typed in and stored in serialized form (that means basic character followed by one or more superposable signs; f.e. $\acute{\bar{e}}$ is typed in as $e\acute{}\bar{}$). The original multi-level composite character is only reconstructed in print by using an overstrike technique. This method seems to work well for

the input and management of dialect data, and above all for data selection, whereas the output shows the original appearance of the characters.

The encoding of diacritics with this system is further complicated by the fact that several instances of the same diacritic have to be designed, in order to obtain optimal results by combining the diacritic. For example tilde and macron normally take the same height relative to a basic sign (ē and ẽ); if tilde and macron are combined, the tilde has to be raised for some pixels; if there is also an acute, the other two diacritics have to be placed over the acute ($\tilde{\bar{e}}$), a macron without a tilde is put beneath the acute ($\bar{\tilde{e}}$).

2. Keyboard layout

To facilitate data input with such complex transcriptions as is the *ALF*-system, it is of great help to redefine the keyboard in such a way as to have the basic characters and diacritics of the chosen transcription system ready to hand. With such a keyboard remapping facility integrated into Dialect-PC a four-level keyboard ("Key", "Shft+Key", "Ctrl+Key", "Alt+Key") can be created to contain all signs used in the dialectological work done by a particular user. As is the case with normal keyboard layouts the redefinition is guided by the frequency of the characters in an application.

With IPA-transcription some proposals for remapping PC-keyboards have been advanced (see Wells 1987 with a summary of older key reassignment techniques). Because the redefinition of the keyboard is independent of the encoding strategy (which has to be standardized to facilitate data interchange), it should be optimized with regard to the subset of the transcription characters needed in a restricted domain like Romance dialectology (as is already the case with 'national keyboards').

3. Map management

Dialect-PC has a module with a drawing program which offers the elementary functions needed to trace maps containing the geographical information (borders, rivers, mountains, towns, etc.) used as a background for the dialect area. In a second step the program makes it possible to set the coordinates for the locality points and for placing the locality name and the corresponding data. All three coordinates are correlated but can be edited separately to get an optimal map layout (for example, it is often necessary to adjust the position of data with variable length which would otherwise get in the way of other data, etc.).

Besides geographical background, locality and data positions there has to be a means of freely positioning all kinds of map text not connected to a single locality but pertaining to the map as a whole (e.g. map captions and legends, map titles, further information like the indication of variants, etc.).

4. Data management

Dialect data are stored by Dialect-PC in a data-base-like form. A DBMS system allows the input and editing of dialect data, locality data and map text independently ("input", "edit", "save", "load", "search data", "merge data with geographical background"). The input and editing of data is done on a screen template which connects a data record (that is the data of one map) and the correlated coordinates of the locality, the locality name and the data. The information of an entire map is thus available on screen at the same time, despite the fact that the data and the coordinates are stored in different files. The merging of this heterogeneous map information for editing and printing is done by associating data and coordinates by means of the locality names. As yet another information type the map text is not directly dependent on the other information and has to be related, resp. loaded separately.

5. Data selection

A separate component is built into Dialect-PC to analyse the raw data by using a kind of query language which incorporates linguistic concepts like "sound group", "sound pattern" and "neighbour to a locality". This component enables us to go beyond the mere management of dialect data and to proceed to data classification techniques. So far, it is possible to select data by 1) identity with the data of neighbouring points or 2) conformity to a sound pattern (this sound pattern is composed of up to three sound groups which can be freely chosen from the set of sounds used with the data). The queries are implemented as logical expressions, the truth functions of which can be combined with the output functions "symbol" or "isogloss". If the output function "symbol" is chosen, data values which fulfill the query condition are substituted by symbols in a previously determined way; if "isogloss" is chosen, there is an activation of the borderline between two adjacent localities. In this way it is possible to create symbol and isogloss maps without manual selection.

The isoglosses are based on a procedure called Thiessen-polygons which divide a whole area into polygons with a locality point as their center. Thus each locality is divided from its neighbours by a line which constitutes part of an isogloss if activated by a query. The concept of "X is neighbour to Y,Z.." is automatically derived from polygon division of the area and need not be determined in advance.

6. Printer management and map layout

Dialect-PC has a printer manager to interactively change printer defaults. This is necessary to get an optimal map layout regardless of the type of map or data. Dialect maps must normally show great flexibility with regard to map scale and correlated symbol heights and font attributes in order to present the map information in an optimal way (for example in map 1 a middle graphics

solution has been chosen with a special *ALF*-type font for the data (10pt italic), a font for the map text (Times Roman 10pt normal) and a font for the map title (Times Roman 12pt normal). In addition there is a special code for switching from normal font to the *ALF*-type font in the map text. These and some other defaults can be associated with a certain map type which is then stored for further use.

7. Sorting sequences

Data in transcribed form normally gave unwanted results with the built-in sorting algorithm based on ASCII-values. The user-defined characters are therefore associated with ASCII-values which do not correspond to the characters' sorting values. Therefore there has to be a utility to determine the sorting order freely, in order to get a sorting order which is for example adequate to the characters' sound values (by using an IPA-system all *a*-like vowels should be sorted together in sequences like [a]-[æ]-[ɐ]-[ɒ] or *d*-like sounds in sequences like [d]-[ð]-[d']-[ɖ]).

8. Applications of Dialect-PC

Dialect-PC makes it possible to store and process the data of existing dialect atlases, dictionaries and monographs in a consistent machine-readable form, and to create new or updated dialect atlases (see table 6 with original data of the *TPPSR* and their topographic distribution in map 1; map 2 shows the same data in an IPA-transcription; map 3 and 4 show all locality points of the *ALF* + *FEW* and of the *AIS* + *LEI*).[2] The data can be made more flexible by using indices and registers, so that a map-independent access to single data can be achieved, opening the way to new research on dialect data.

[2] A more detailed description of applications is given in Geisler (1990/1991).

Table 1

DIALECT DATA PROCESSING

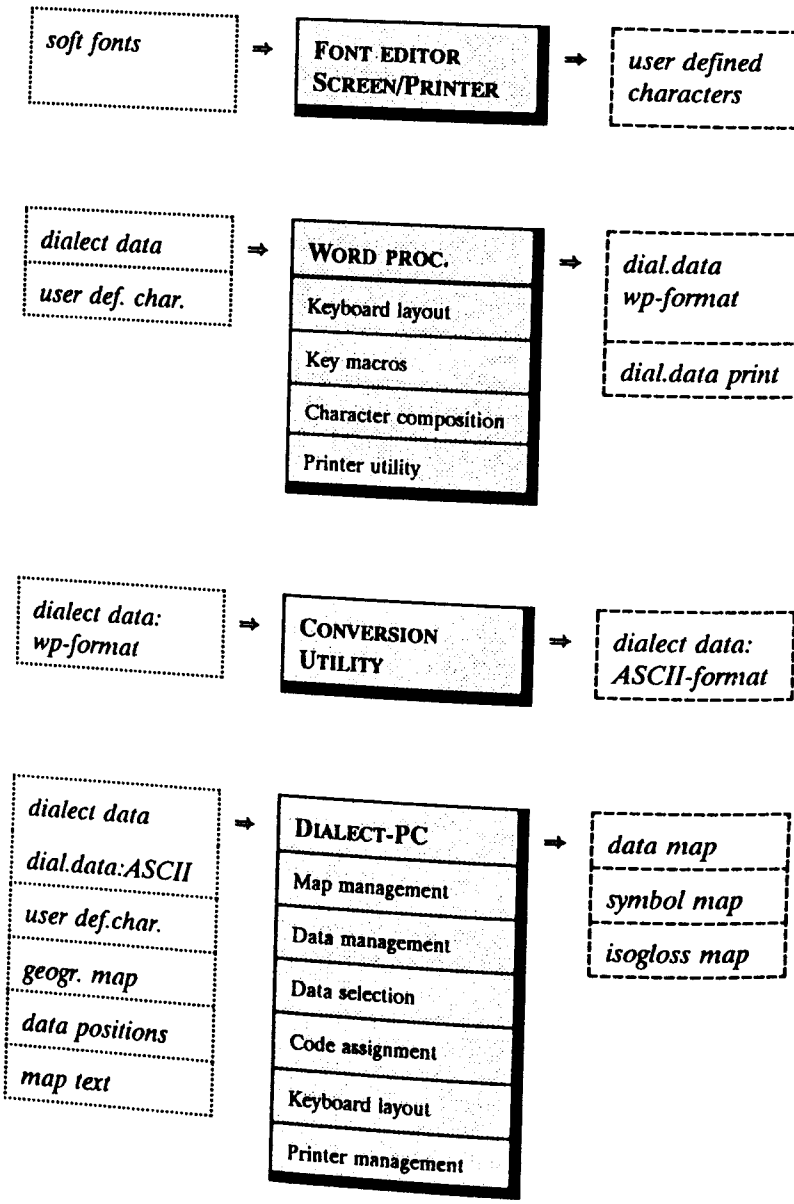


Table 2

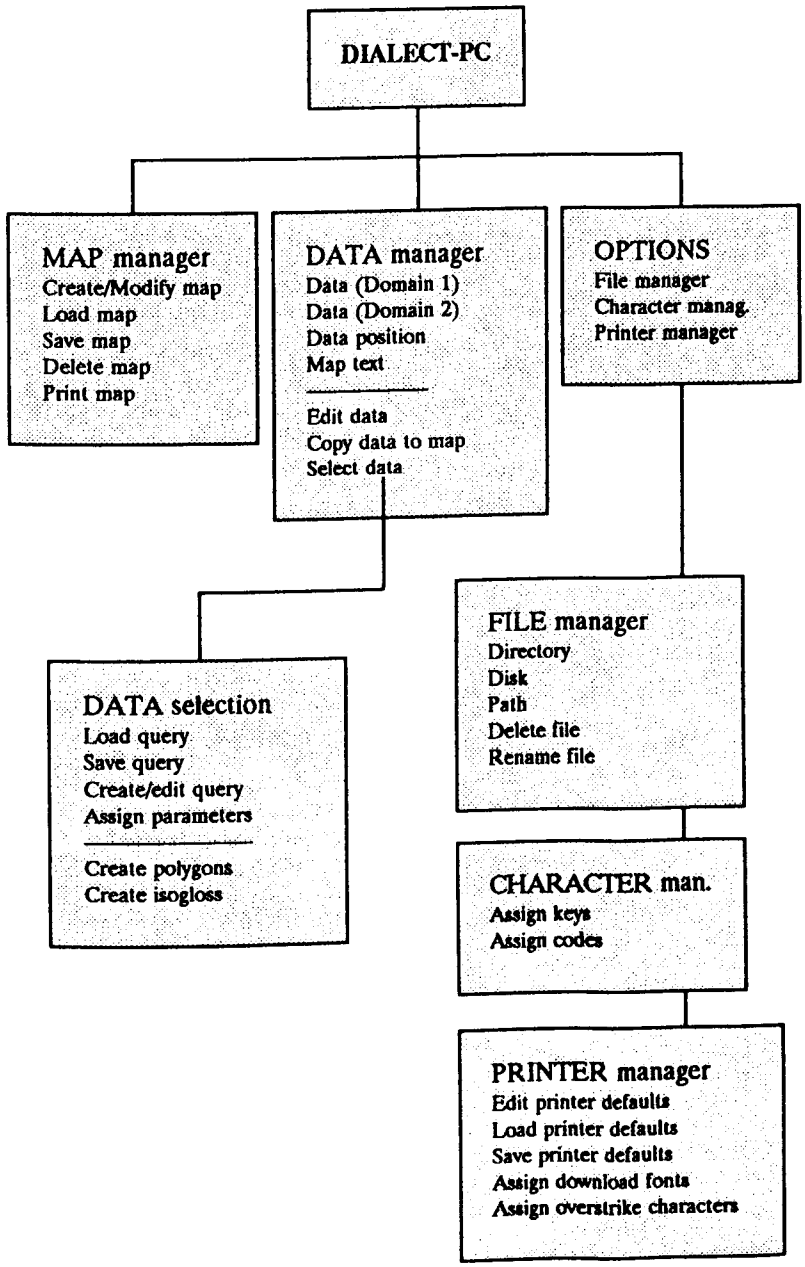

DIALECT-PC
MAP manager
Create/Modify map
Load map
Save map
Delete map
Print map
DATA manager
Data (Domain 1)
Data (Domain 2)
Data position
Map text
Edit data
Copy data to map
Select data
OPTIONS
File manager
Character manag.
Printer manager
DATA selection
Load query
Save query
Create/edit query
Assign parameters
Create polygons
Create isogloss
FILE manager
Directory
Disk
Path
Delete file
Rename file
CHARACTER man.
Assign keys
Assign codes
PRINTER manager
Edit printer defaults
Load printer defaults
Save printer defaults
Assign download fonts
Assign overstrike characters

Table 3

IPA 1989 Kiel Convention

Diacritics

̥	Voiceless	n̥ d̥	̟	More rounded	ɔ̟
̬	Voiced	s̬ t̬	̠	Less rounded	ɔ̠
ʰ	Aspirated	tʰ dʰ	̟	Advanced	u̟
̤	Breathy voiced	b̤ a̤	̠	Retracted	i̠
̰	Creaky voiced	b̰ a̰	̈	Centralized	ë
̼	Linguolabial	t̼ d̼	̽	Mid centralized	e̽
̪	Dental	t̪ d̪	̩	Adv. t. root	ę
̺	Apical	t̺ d̺	̯	Retr. t. root	ę
̻	Laminal	t̻ d̻	˞	Rhotacity	ɚ
ʷ	Labialized	tʷ dʷ	̃	Nasalized	ẽ
ʲ	Palatalized	tʲ dʲ	ⁿ	Nasal release	dⁿ
ˠ	Velarized	tˠ dˠ	ˡ	Lateral release	dˡ
ˤ	Pharyngealized	tˤ dˤ	⁻	No audible release	d⁻
̴	Velarized or pharyngealized	ɫ			
̝	Raised	ę ɻ (= voiced alveolar fricative)			
̞	Lowered	ę β̞ (= voiced bilabial approximant)			
̍	Syllabic	n̩	̯	Non-syllabic	e̯

Suprasegmentals

ˈ Primary stress ˌfoʊnəˈtɪʃən

ˌ Secondary stress

ː Long eː

ˑ Half-long eˑ

˘ Extra-short ĕ

. Syllable peak ɹi.ækt

| Minor (foot) group

‖ Major intonation group

‿ Linking (absence of a break)

↗ Global rise

↘ Global fall

Table 4

Transcription system ALF (= Rousselot-Gilliéron)

Diacritical marks: Vowels

1 closed: ´ 2 opened: ` 3 shortened: ˘ 4 long: ¯
5 nasalized: ˜ 6 slight nas.: ˜ 7 stress: ,
8 intermediate: å 9 weak articulation: a
10 centralized: ·

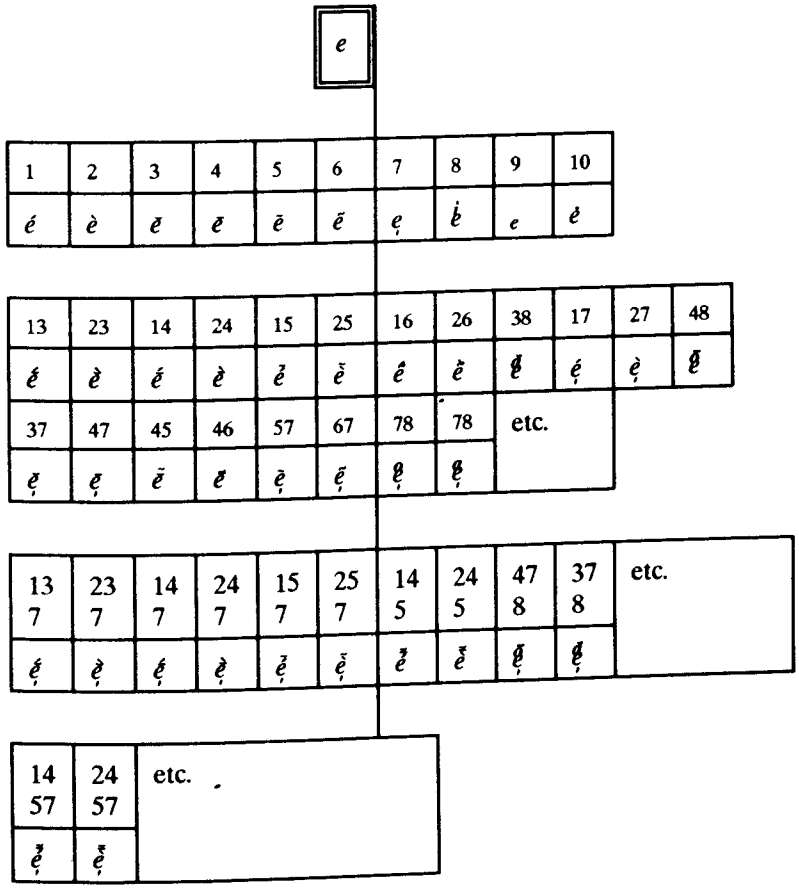


Table 5

Transcription systems

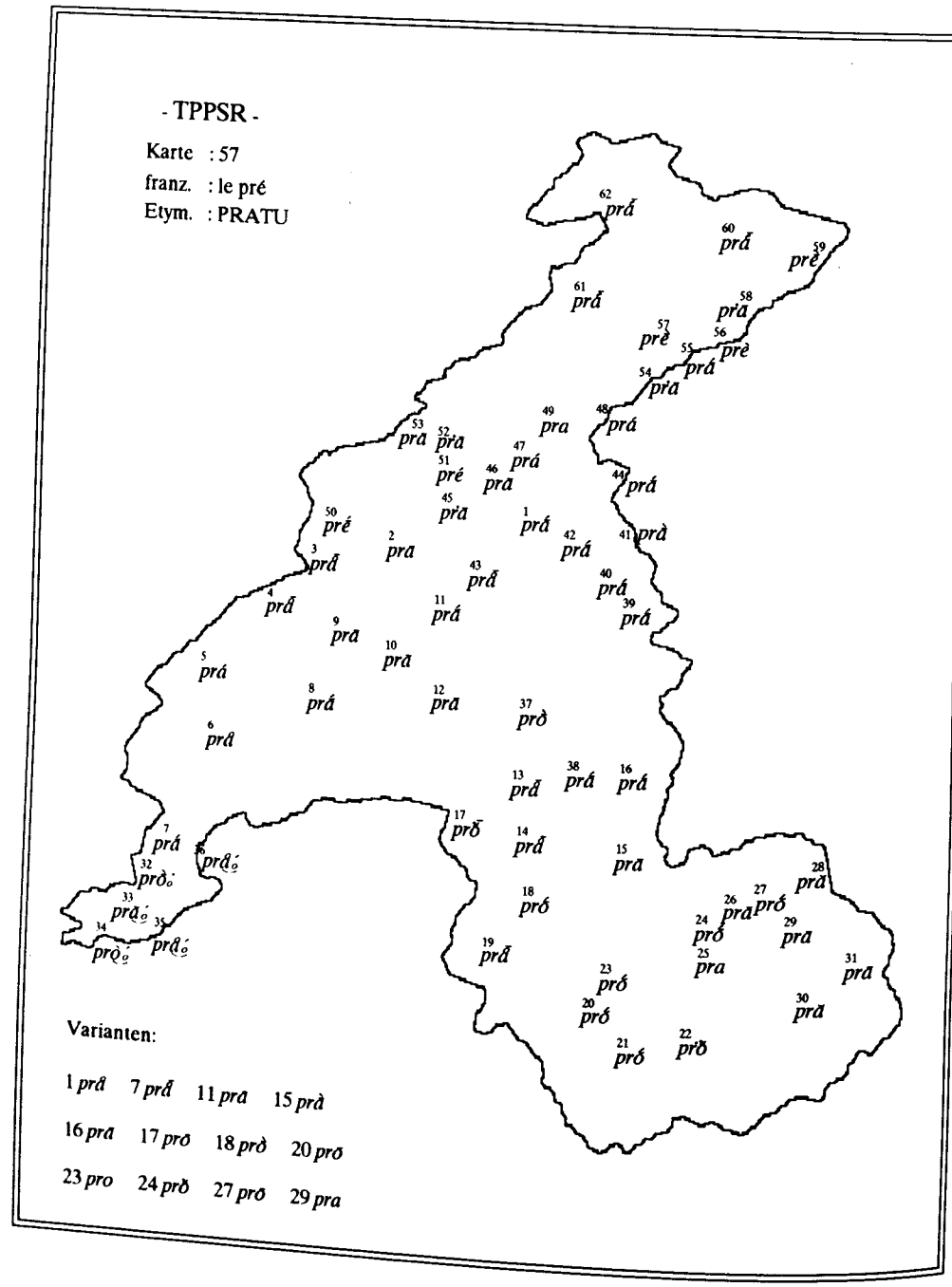
Equivalences - Vowels

IPA	AIS	FEW	LEI	ALF	...	...	...	...
i	i̯	i	i	ɩ́				
ɪ	i̯	i̯		ì				
e	e̯	e̯	e̯	é				
ɛ	e̯	e̯	e̯	è				
æ	a̯	a̯	a̯	á̯				
a	a	a	a	à				
ɑ	ɑ̯	a̯	α	á				
ɒ	a̯	a̯	a̯	a̯				
ɔ	o̯	o̯	o̯	ò				
o	o̯	o̯	o̯	ó				
ʊ	u̯	u̯		ù				
u	u̯	u	u	ú				
y	ü̯	ü	ü	ú				
Y	ü̯	ü̯		ù				
ø	œ	œ	œ	œ́				
œ	œ̯	œ̯	œ̯	œ̀				
ə	ə	ə	ə	ə́				
ɐ	α			ɑ̀				

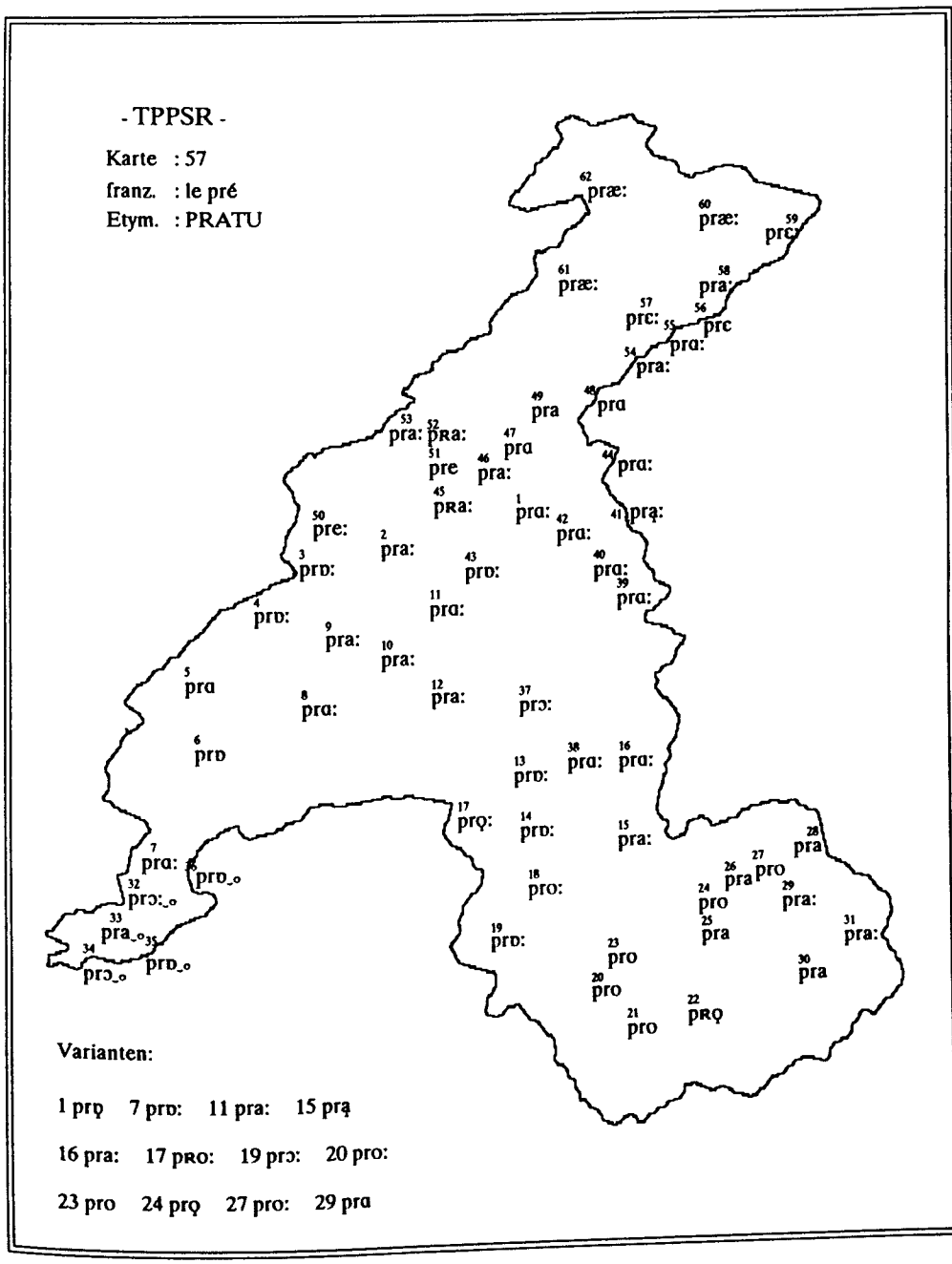
Table 6: *TPPSR*, tabl. X, col. 57: lat. PRATU, fr. *pré* (cut-out)

Ortsname	Nr.	TPPSR-Beleg	TPPSR-Variant.	IPA-Notation
Chevroux	1	*prå*	*prå*	prɑ:
Vaugondry	2	*prɑ̊*		pra:
L'Auberson	3	*prɑ̈*	*prɔ̈*	prɒ:
Vallorbe	4	*prɑ̈*		prɒ:
Le Sentier	5	*prá*		prɑ
Longirod	6	*prɑ̊*		prɒ
Commugny	7	*prá*	*prɑ̈*	prɑ:
....				
Lourtier	22	*pɾʊ̈*		pʀǫ
Fully	23	*prɔ́*	*pro*	pro
Conthey	24	*prɔ́*	*prʊ̈*	pro
Nendaz	25	*pra*		pra
Savièse	26	*prɑ̆*		pra
Ayent	27	*prɔ́*	*prɔ́*	pro
Miège	28	*prɑ̆*		pra
...				
Vermes	59	*prɛ̀*		prɛ:
Develier	60	*prɑ̈*		præ:
...				

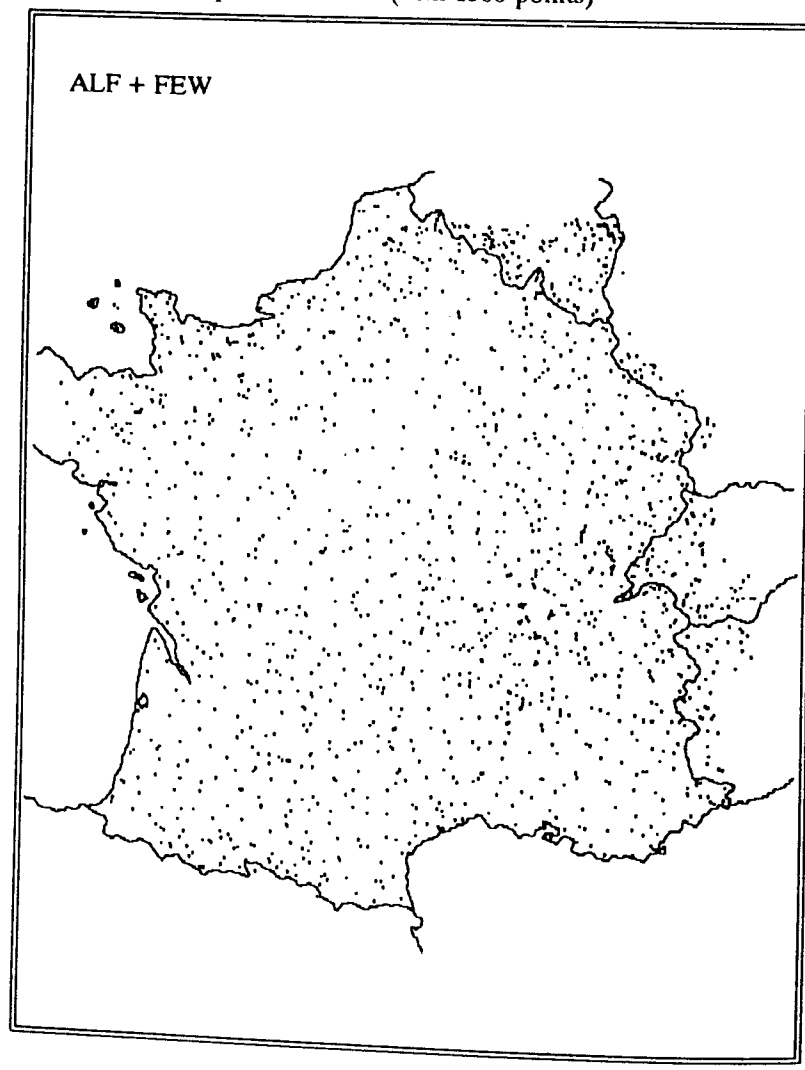
Map 1: *Tableaux phonétiques des patois suisses romands* (original transcription)



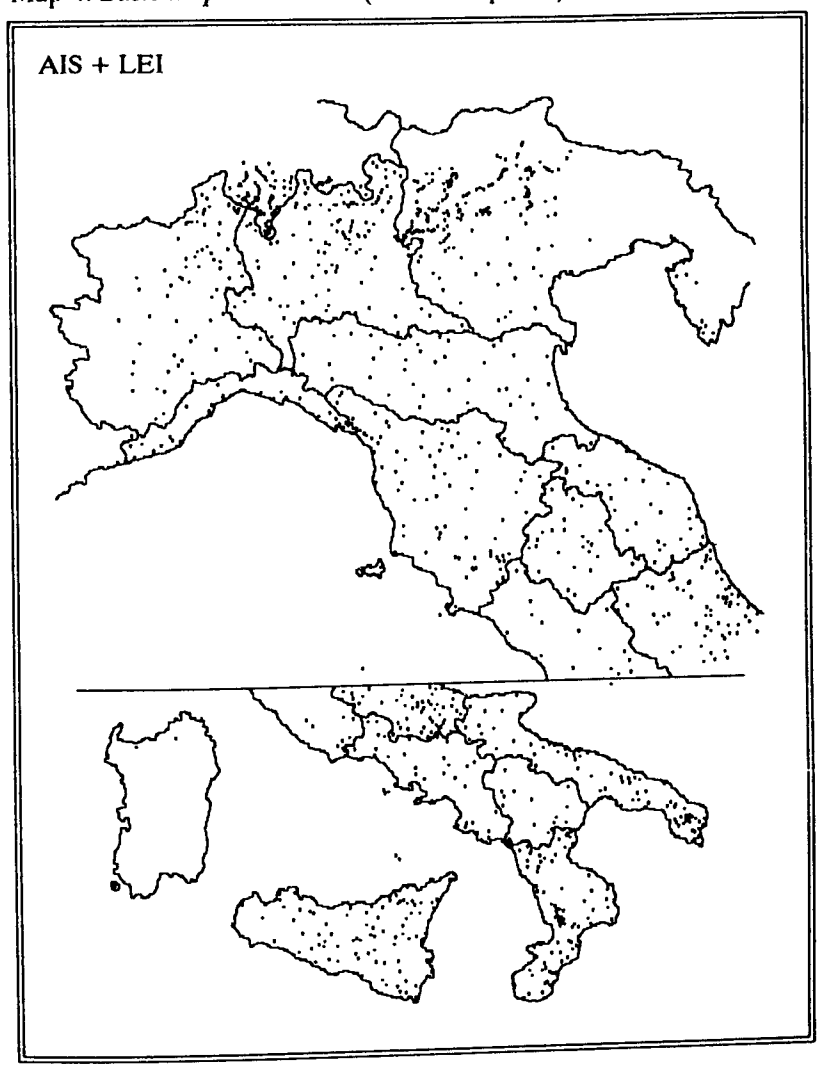
Map 2: *Tableaux phonétiques des patois suisses romands* (API-transcription)



Map 3: *Basic map ALF + FEW* (with 1560 points)



Map 4: *Basic map AIS + LEI* (with 1330 points)



Bibliography

AIS = Jäberg, K., and J. Jud. 1928-40. *Sprach- und Sachatlas Italiens und der Südschweiz*. 8 vols. Zofingen.

ALF = Gilliéron, J., and E. Edmont. 1902-04. *Atlas linguistique de la France*. 10 vols. Paris.

FEW = Wartburg, W. v. 1922-. *Französisches Etymologisches Wörterbuch. Eine Darstellung des galloromanischen Sprachschatzes*. Bonn - Leipzig - Tübingen - Kassel.

Geisler, H. 1990. "Aufbereitung lexikalischer Information und deren Umsetzung in Sprachkarten (aufgezeigt anhand romanischer Beispiele)." In: B. Schaeder and B. Rieger, eds. *Lexikon und Lexikographie: maschinell - maschinell gestützt. Grundlagen - Entwicklungen - Produkte.* Sprache und Computer, 11. Hildesheim, Zürich, New York, 233-43.

-----. 1991. "Erstellung und Auswertung von Dialektkarten mit Personal Computern." In: J. Rolshoven and D. Seelbach, eds. *Romanistische Computerlinguistik. Theorien und Implementationen.* Linguistische Arbeiten, 266. Tübingen, 209-29.

International Phonetic Association. 1989. "Report on the 1989 Kiel Convention." *Journal of the International Phonetic Association* 19/2: 67-80.

LEI = Pfister, M. 1979-. *Lessico etimologico italiano*. Wiesbaden.

TPPSR = Gauchat, L., J. Jeanjaquet and E. Tappolet, eds. 1925. *Tableaux phonétiques des patois suisses romands, relevés comparatifs d'environ 500 mots dans 62 patois-types*. Neuchâtel.

Wells, John C. 1987. "Computer-coded phonetic transcription." *Journal of the International Phonetic Association* 17/2: 94-114.