Dirêjî Kurdî : a lexicographic environment for Kurdish language using 4th Dimension®

Gérard H. Gautier

Centre for Kurdish Studies

GAUTIER Gérard[1]

Paper presented at ICEMCO 96
5th International Conference and Exhibition on Multilingual Computing
London, UK, April 1996

Dirêjî Kurdî دریژی کوردی : a lexicographic environment for Kurdish language using 4th Dimension®

Abstract: *Keywords : Kurdish, multiscript, database.*

The Kurdish language presents specific problems for the lexicographer. Being the language of a people without a state, it has attracted less interest than the other languages of the Middle East, and it is only recently that lexicographic studies and bilingual dictionaries of Kurdish have been published. It lacks standardisation (division in dialects, and above all, use of three *different writing systems). It also presents further problems in the field of computing (encoding).*

Paradoxally, this complicated situation means the computer may be especially useful for lexicographic work on the Kurdish language. This paper deals with Dirêjî Kurdî*, a project for the development of a lexicographic software environment specifically geared towards Kurdish : the different phases of the project, the reasons for the technical choices made by the author, its current state of development, the problems encountered and the planned extensions (linking with CD-ROM dictionaries, corpora, compliance with TEI-SGML) are introduced.*

I) ORIGIN AND CONTEXT OF THE PROJECT

1- The "prehistory" of the project : dictionary on papier, dictionary on computer

Perhaps a word of warning is in order here : this is an applied research, whose aim is to do the best possible use of current text and database technology for lexicography in a particular case.

The first idea of this project came from my participation in a lexicographic work with Dr. Hakim in Paris, the result of which was a French-Sorani dictionary [Hakim & Gautier, 1993]. Dr Hakim had already been working since several years, and, using the traditional technology current up to the preceding generation of dictionaries, had built up an impressing amount of data on paper cards. I proposed to computerize the data, and we chose the Macintosh system as the most multilingual. The words were entered first in latin transcription and sorted in Microsoft Word. As there were neither citations nor didactic examples in the data, what we achieved was a lexicon, the wordlist of which had been determined by Mr. Hakim as a large subset of the *Dictionnaire fondamental de la Langue Française* [Gougenheim, 1958 (1993)].

[1] - After studies in Kurdish (Institut des Langues Orientales, Paris) and computer science (master, 1988), G. Gautier worked as a professional trainer and obtained a master (DEA) in educational science. After moving to Taipei to study chinese language needed for his Ph.D (anthropology, 1994), he worked there as a technical writer for MITAC. Currently assistant professor of French language and literature at Wen-Tzao Junior College, member of the Centre d'Etudes Kurdes (Paris), he has published on educational technology and anthropology of educational institutions and participated with Dr. Hakim in the making of the first comprehensive French-Kurdish dictionary (1993).

The data was then transfered over to the multilingual word-processor WinText (from the french company WinSoft[2] and the Arabic-Persian writing associated with Sorani was typed next to the Latin transcription.

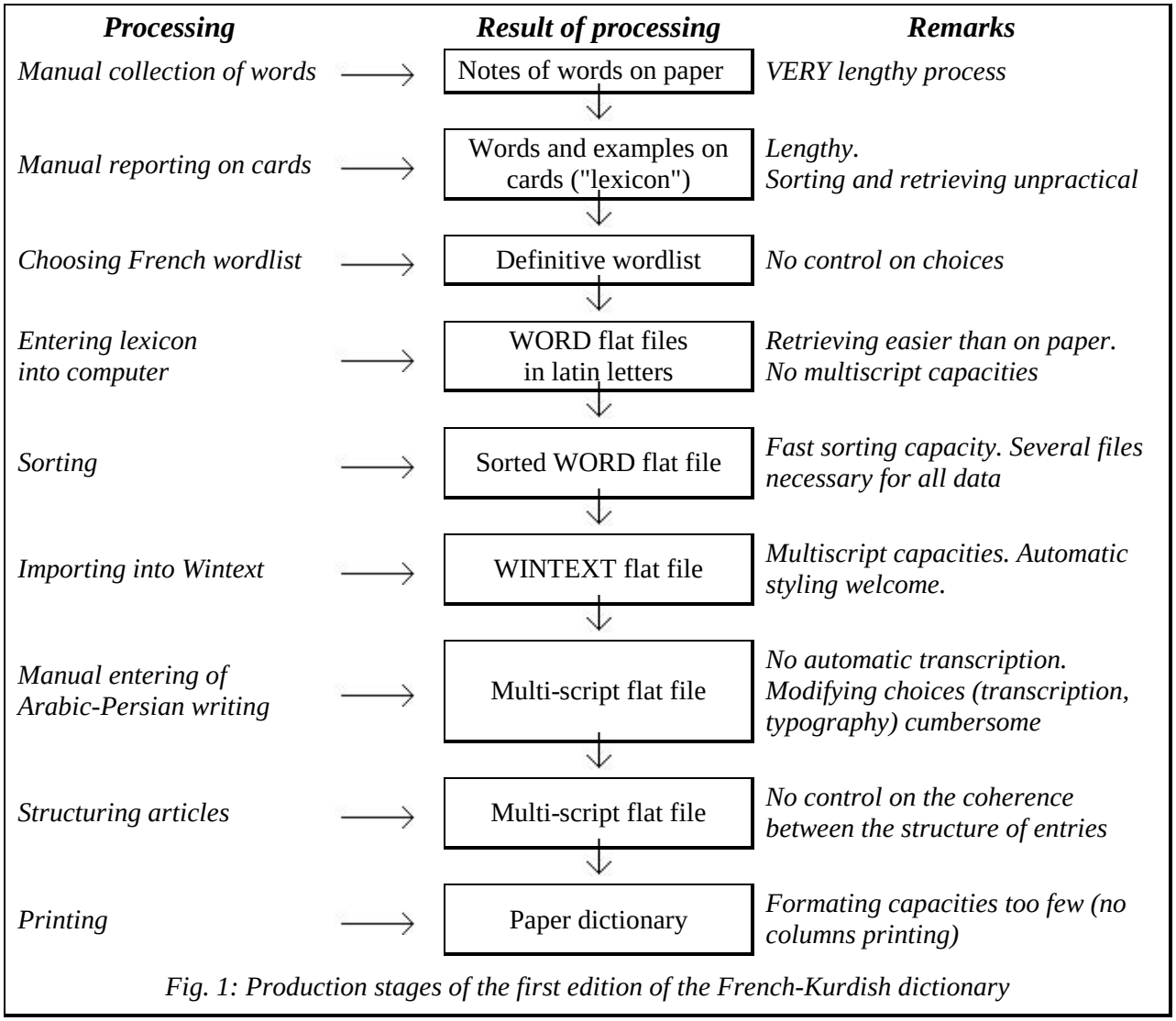


Fig. 1: Production stages of the first edition of the French-Kurdish dictionary

The use of a word-processing software was undoubtedly a big step forward from the more traditional way of working : using automatic "styles" in WinText make typographically formating parts of entries much quicker. The Kurdish letters too were produced in a easier way - although the "Diwan" Kurdish fonts, modified from Arabic ones, had some spacing problems.

But there were also obvious shortcomings : WinText did not enable formating in columns ! That necessitated a specially painful process when we had to print out a two-column dictionary. The data, stored in a proprietary format, could not be easily reappropriated for re-formating or new lexicographical work. Furthermore, the chart above shows the choices were all made in sequence and were then quasi-irreversible. What we had was in effect an *electronic* version of... a *paper* dictionary.

The data, although in electronic form, had no specific structure. A word processor did not offer enough in terms of reusability. What was obviously needed was some sort of database with capacities specifically geared towards linguistic data and the unique characteristics of Kurdish language.

[2] - The information about companies cited will be found in Appendix.

2- From a Sorani electronic dictionary to a Kurdish lexicographic environment

The aim of the project was first only to import the first edition of the dictionary into a database to build a truly versatile electronic version of it, which would then be used as a basis for an enlarged second edition. A prerequisite for such a work was the authorisation from Mr. Hakim to use the flat files data, which he willingly granted me.

Determining then which software to choose required a *technological survey* of the multilingual software market, as well as a general analysis of the specific situation of Kurdish language. So the project would have four phases : one year (which later turned out to be two) of *technological survey* to know the possibilities offered by current software, six months for *technical choice*, one year to assess the feasibility of the project through the *building of a prototype*, and then one year to finalise the product in its "first version".

But as soon as the first phase, the potential power of computerisaton caused me to modify the project : the transfer of the French-Sorani data on another support evolved into a more general purpose : to build a versatile *lexicographic environment* for Kurdish in general, easy to use for non-programmers, but powerful enough to help a lexicographer to generate new products after his own specifications.

Then, before prototyping, a (somewhat belated) reflection on the specific situation of Kurdish language itself showed me the problem was really very specific, and led me again to redefine the project - which at this time I decided to rename دریژی کوردی [Dirêjî Kurdî] "Kurdish Dimension", to put emphasis on the complexity of the situation (see fig. 2).

Phase	**Duration (Dates)**	**Aims**	**Remarks (Tools, choices...)**
1/ Technological survey	2 years (1993-1994)	Acquiring knowledge of: ▪ market ▪ multilingual tools ▪ lexicographic tools, basic needs and methods.	Sources: ▪ Specialised literature ▪ Reviews: *Sesame*, *Multilingual Computing* ▪ Discussion lists on Internet (PC-ARAB, READER, ITISALAT, INSOFT) ▪ Specialised CD-ROM databases (esp. ZIFF Computer Select)
First redefinition of the project : from electronic dictionary to lexicographic environment			
2/ *Technical choices*	6 months (beginning summer of 1995)	Determining which set of tools to use. Choosing a system (software and hardware) First tests of general feasibility	Choices and reasons: ▪ *Fourth Dimension*™, chosen january 1995. ▪ Base system: MacIntosh, but possibility of a port on PC. ▪ Versatile programming language.
3/ *Prototyping*	1 year (in progress at the date of writing)	Building a prototype to verify the feasibility of the project with the tools chosen.	▪ Specific study of Kurdish language problems ▪ Redefinition of aims
Second redefinition of the project : from Sorani-oriented to multi-script Kurdish			
4/ *Final realisation*	1 year (half 1996 to half 1997)	Extending the prototype into the final software tool	▪ Compilation ▪ Distribution as shareware

Fig. 2 : The evolution of the Dirêjî Kurdî *project*

II) ANALYSING THE SITUATION OF KURDISH LANGUAGE

1- The specific situation of Kurdish : different dialects and writing systems

Kurdish comprises a variety of dialects, just as Chinese does. But, contrary to the latter, as Kurds are divided among several states, there is no standardising body for Kurdish language. So there has been no emergence of a "national standard language" comparable to Chinese Mandarin. Although the different dialects of Kurdish are linguistically closer to each other than the different varieties of Chinese, contrary to Chinese (which is from the point of view of the reader "united" by the use of a common ideographic writing system), Kurdish dialects are further set apart by the different writing systems used in the countries among which Kurdish populations are scattered.

Although it is not the aim of this paper to dwell on the intricacies of different dialects, political situations, and choices of writing system and researchers' transcriptions for Kurdish, I will have to devote a minimum of space here to introduce the general context which led to the current definition of the *Dirêjî Kurdî* project.

2- The different dialects

Kurdish, an Indo-Iranian language, is divided into two main dialects, Northern Kurmanji (thereafter "Kurmanji") and Southern Kurmanji (thereafter "Sorani") allowing intercomprehension between educated speakers and in fact between any Kurds after some days of practice, but still grammatically very different [Kreyenbroek, 1992, pp. 68-83] [Izady, 1992, pp. 167-175]. Sorani and Kurmanji are themselves subject to local variations ("Mehabadî", "Sulîmanî"...). Two other, less widely spoken dialects, linguistically related together, are Dimîlî (sometimes called "Zaza" in a somewhat derogatory way) and Gorani. They can be found respectively in the far northern and southern part of Kurdistan.

A consequence of this - and also of a general political situation where research and teaching in/of Kurdish has been either forbidden or not given enough means to develop - is a problem of standardisation of the language itself. However, Kurdish seems lately to evolve in the direction of a *bi-standard* situation [Kreyenbroek, op. cit.], with Kurmanji (written in Latin letters) becoming the language of numerous publications almost totally appearing in emigration, and Sorani (written in arabic letters), becoming a de facto litterary standard, a trend which began at the turn of last century, after Gorani had lost its preeminence. *But* :

- This trend towards a "bi-standardisation" doesn't imply the disappearance of local variations inside main dialects,
- According to the State which administrates a given area in Kurdistan, the spoken language there receives a heavy influence from the dominant language (either adoption of words or resistance to it),
- The way of writing the language itself is evolving to reflect phonological changes. As an example, there is a trend to no longer write some double wovels (i.e ووڕچ `wurch` tends to become وڕچ `wirch`) [Hakim, 1993], *Introduction* to [Hakim & Gautier, 1993].

3- Three different writing systems in Kurdish lands

The lines of divide between writing systems match the political boundaries, but those in turn only match the linguistic borders between dialects very approximately : if it is true that, in a first approximation, Kurmanji mainly uses a Latin script and Sorani an Arabic-Persian one, it is only a generic correspondance : Kurmanji in ex-Soviet Armenia uses a Cyrillic alphabet, in Turkey a Latin alphabet, in northern Iraq and northwestern Iran an Arabic-Persian alphabet, and in Syria (and probably so in Ex-Soviet Azerbaijan) another Latin transcription (although I have no precise information about its difference with the one used in Turkey) (see map fig. 3).

If the situation of Sorani seems at first easier, as it is mainly written in Arabic-Persian script, then from a *computerisation* point of view, it is in fact not better at all : encoding standards or practices for

Persian and Arabic, apart from being themselves rather fluid[3], are vastly different from each other, so the choice of encoding for Kurdish is not at all self-evident.

Fig. 3 : Kurdish : map of political (= writing systems) and dialectal borders
Simplified map. Sources : [Kreyenbroek, 1992, pp. 68 sq] *and especially* [Izady, 1992, carte p. 170]

The would-be Kurdish lexicographer has also to take into account the *historical* evolution of the writing system : before 1929, the date for the reform of the alphabet in the Turkish Republic[4], Kurdish (as Ottoman Turkish) was written in arabic letters, so some of the early lexicographical works on Kurmanji [Jaba, 1890], are using an "Ottoman-like" Arabic system of writing, although very different to the one now in use for Sorani...

4- Transcription problems

Either in its Arabic, Latin, or Cyrillic writing, depending on the dialect, Kurdish script is almost totally phonetic. In Sorani, unlike in Arabic, all vowels are written on the line (corresponding to the "long vowels" of the arabic alphabet) and cannot be omitted. It may be linked to the fact that, not being a semitic language, Kurdish ignores the consonantic structure and internal vowel variations necessary in Arabic (if the word "book" in Kurdish is indeed "kitêb" (کتێب), its plural form is a regular "kitêban" (کتێبان) and no "kutubun" (کُتُبٌ).

A consequence is that automatic transcription would be easer to program if the need arises. We will see that this characteristic influences another area, very important for lexicography : the building of words lists properly collated after the order defined by the grammarians of the Kurdish Academy in Baghdad. The only exception to the full writing of vowels is the mute e (schwa [ə]), which is written only at the initial of a word as ئ and (except for very few words mainly coming from Arabic), mainly before another vowel, with which it forms a compound as ئا or ئه. In Kurmanji, this same sound is expressed by a dotless i [ı], as in Turkish (the opposition being between [ı] or [i] or, sometimes, between [i] and [î], depending on the publication).

5- Distinguishing different levels of variations in the language

If, compared to French or English, Kurdish writing may be said to be fairly phonetic, it is not to the point that an IPA transcription would be made unnecessary. Apart from the problem of the

[3] - Iran has just defined a new standard, and in the Arabic world, the Microsoft Arabic Windows proprietary encoding modified the physionomy of the real-life practice profoundly last year...

[4] - The law introducing the new Turkish (Latin) letters was passed on november 1928, and all publications started to use them in 1929 [Feroz, 1993, pp. 80-81].

schwa (mute e) already mentioned for Arabic-Persian writing on the Sorani side, there are also variations in pronunciation on the Kurmanji side, for several consonants which may or may not be accented. Accented letters will usually be unaspirated, and unaccented letters will be aspirated. This is a phonological problem, on the same level as, say, the softening of the spanish [d] in nada, which pulls it towards the english soft [th].

This characteristic prompted [Rizgar, 1994] in his Kurdish-English Dictionary, which is centered on Kurmanji, but also contains words of other dialects, to alter the usual writing system of Kurmanji as practiced in Turkey to underline the *accented* sounds for ç, k, p, r, t. Rizgar notes that the Cyrillic writing used in Armenia distinguishes the accented letters by a postscript quote. This information have been confirmed by the data kindly supplied to me by M. Chyet [Chyet, personnal communication, 1995, © 1990].

From this information, I tried to build a general chart of correspondence between Kurdish writing systems and sounds, hence including an IPA reference (see Appendix A). With reference to the phonological contextual modifications of pronunciation, however, such a chart should be regarded with a great deal of caution. The "standard" pronunciation it describes is likely to be altered - and also according to locality. Specific problems also arise from the fact that Kurdish is a *bundle* of dialects, not a unified language. So the "same" word, when going from Kurmanji to Sorani, may at the same time go through several levels of change : writing system, phonology, morphology, and semantics...

For instance, on the Kurmanji side, the letter x is used as corresponding sometimes to the Sorani خ as in xode (God), which transcribes خـوّده but also sometimes to غ, as in kexaz (paper), which transcribes کەغاز (no morphological difference between dialects, but a script problem). The Sorani word مـاڵ (family, home, house) has a Kurmanji equivalent which is mal : the sound corresponding to the Sorani letter ڵ does not exist in Kurmanji (phonological difference between the two dialects, reflected in the writing system). Another way to tell a house (with emphasis on the *building* side, not the family) is in Kurmanji the word xanî, which becomes in Sorani خانوو (xanû). It seems to me that here we are before a *dialectal* difference which is to be recorded as such. A lexicographic program should obviously be able to distinguish clearly between those different levels.

6- The numerous "lexicographers' transcriptions" of Kurdish

Since the 20s, researchers or lexicographers in Kurdish, westerners as Kurds, even if they were concerned with Sorani, have tended to use different Latin transcriptions concurrently with the original writing.The adoption of a Latin alphabet by the new Turkish Republic has played in this direction as well as the fact that many of those works were published in the West.

The transcriptions used in those works vary according to the country of origin and the academic background of the authors (Arabic or Persian studies) as the current transcriptions for those languages differ. The choice may also depend on the scope and type of the work. [Wahby & Edmonds, 1966 (1971)] use a latin transcription which differs from the one used for Kurmanji in Turkey (ch instead of ç for ج, sh instead of ş for ش etc...) This is also the transcription used at the *Institut National des Langues Orientales* of Paris [Blau, 1980], [Hakim & Gautier, 1993]. In contrast, in his reference work, which goes further than lexicography, [Izady, op. cit.] uses â for the Arabic letter ا and a for the Arabic letter ە, for the sake of consistency of transcriptions between all the languages concerned, throughout a period of several thousand years and across a continent[5]...

7- A lack of typesetting tools only lately addressed

The lack of proper facilities for typesetting Kurdish of whatever dialect has consistently been a problem for authors since the beginning of publications on/in this language. In the last century, the

[5] - A table showing differences between usual transcriptions for some words will be found in Appendix.

situation was quite the same for any non-western language, and the Jaba Kurdish dictionary is no different in quality (in this matter) that the Wade edition of the chinese *Yi-Jing* (周易)... But if the situation has somewhat improved since then for many other non-european languages, that is only true since very lately for Kurdish : [Mc Carus, 1967] as well as [Blau, 1980] had to add by hand the Sorani diacritics, and if [Rizgar, 1994] underlines the h̲ to transcribe the arabic letter ح, to differenciate it from the arabic letter ه, whereas the common transcription for ح is ḥ, it is most probably due to the lack of computer facilities and/or of a good typographic system. This problem often pushed authors to use the underlining[6] as a substitute for an underwritten dot [Hakim & Gautier, 1993], or even a quote instead of a little overwritten ᶜ for ع .

8- Kurdish as an Arabic "add-on" ?

Lately, specific Kurdish fonts have been produced both for PC (Windows) and Macintosh platforms : *Diwan* company has bundled "Goran" and "Sulîmanî" fonts with its pagesetting software *Al-Nashir al-Maktabi* (and probably with the *Al-Sahafi* version). But with the late disappearance of this company, it is difficult to know what those fonts (which were however not without defects) will become, and *Decotype* company just released a *Naskh Kurdish* font for use under *Arabic Windows 3.1*.

But Kurdish appears often on the computer more as a *modification* of another language – mainly Arabic – than a language in its own right : on Macintosh as well as Windows, Kurdish is a *font* under an Arabic *system*. Kurdish users may so be misled by the Arabic keyboard and indeed type Kurdish as a modification of Arabic : for instance Kurdish letter ه / ھ is often typed improperly as the <ARABIC LETTER HEH> ه / ھ / ه, which therefore must be followed by a non-joiner before any connecting letter to prevent for instance the contextual transformation of گەرم germ into گهرم ghrm or of دەهۆل dehol into دەهھل dehhl. In contrast, *Diwan* and *Decotype* fonts, following [UNICODE / ISO 10646, 1991] which provides a separate codepoint (06D5) for the Kurdish under the name of <ARABIC LETTER AE>, separate completely those two letters, which in my opinion the way to go[7]. But a "Kurdish" physical keyboard should reflect that unambiguously by distinguishing clearly on the keys the two glyphs (ه vs ه), as the Kurdish HEH seldom appears in end position (ie as ھ), but rather as an initial followed by a vowel in current digraphs as هه ها هو هی etc.

III) ANSWERING THE (LINGUISTIC AND TECHNICAL) CONTEXT

1- Is the computer useful as such ?

From the preceding section, one can get an overview of the situation : the researcher has to deal with what I will call a bundle of dialects, three different writing systems, writing standards and software standards either non existant or "in becoming", and to top everything, a quasi-impossibility to do on-field lexical research should be mentionned ! It is not surprising then that numerous different choices have been made by former researchers.

Any software designed for Kurdish should be able to allow the user to choose between those possibilities, and keep the choices *open* (i.e mainly : reversible to the last moment). A good example is the choice of transcriptions used in a printed work on Kurdish : I just showed how several different choices had been made by different authors, each with a potent logic. The fact is that, in a paper publication, the author has to choose *one* solution only, even if he thinks (and that is usually the case) that it is not a perfect one. He has to do some sort of bargaining with the situation. But the computer could enable determining which transcription to use for a given set of data according to the specific parameters of use : not only the dialect (obviously, it is more a temptation

---

[6] - The problem is that underlining is not really encoded in the text as a specific letter, but as a typographic property prone to disappear with exportation to another software, if precautions are not taken to ensure its survival...

[7] - The Kurdish letter ه / ھ uses the codepoint of the <marbutta> ة, which does not exist in Kurdish. This does not prevent typing Arabic, as a simple change of font does the trick if needed.

to present a Sorani word in the Persian-Arabic script), but also the intended user : will the reader use primarily this script ? Would not a Kurdish reader from Armenia benefit to be able to look to Sorani data in the cyrillic system of writing he uses daily ? The decision could be made on-the-fly by the user according to his / her own needs.

So the proper environment for Kurdish lexicography should allow for changing methods of transcription and fonts and choosing the most suitable way to present data to the user – *whatever internal encoding is used* . Similarly, in regard of the instability of standards, a good lexicographic environment should be able to *change encodings* when needed for importation / exportation, hence combining versatility with respect of standards[8]. With the versatility of modern software, which allows a great deal of manipulation of the way in which data is represented and presented, the same software may now suit different transcription needs according to the user's request, and export/import data from/to files or other databases using different internal encodings and logical formating (TEI...).

If will probably make the development more difficult. But, at the same time, because of the openness of the computer regarding transcription problems, its ability to retain complex information such as the origin of a word, its variants in spelling and using, it can be argued that a work on Kurdish done on computer following these lines will be more re-usable than a one done by traditional means. As challenging as it is, I would want to make the point here that the situation of Kurdish language also lends a real interest to the use of the computer, but at one condition : any solution proposed to the problem of defining a software tool for kurdish lexicographic purposes has to be open and susceptible to modifications.

2- The reasons for the choices of Fourth Dimension® on Macintosh

After a two-year technological survey about theoretical issues and application software for multilingual and linguistic computing, the following choices were made :

- to develop a multiscript lexical database which could import former work (flat-files data in general), manage a corpus of Kurdish texts and generate dictionaries according to precise specifications ;
- the chosen platform was the MacIntosh, and the chosen software was ACI "4th Dimension™" (hereafter "4D"), for the following reasons :
  - at the time of choice, the only real multilingual system was the Macintosh (the Windows-based PC being at best only bilingual, and no database system allowing Sorani was available on PC[9]). The MacIntosh *System 7* with its WorldScript extension enabled problem-free mixing of English, French (with its accented characters), Cyrillic-like and (Latin) Turkish-like transcriptions of Kurmanji, and Arabic-Persian-like transcription of Sorani. Furthermore, defining customised keyboard drivers for all those different fonts is much easier on a Mac than under Windows.
  - 4D has a "two-level" structure : a programming environment, and a user environment. It is possible to use the programming environment to build a friendly interface for the non-programming user ;
  - the 4D development environment gives access to a powerful, compilable language, and ACI (the developper of 4D) allows license-free distribution of any compiled product as a (non-modifiable) runtime, which opens the perspective of a shareware-based research tool;

[8] - While it must be acknowledged that only a general revamping such as the one proposed by UNICODE / ISO-10646 will give a solution to this problem, the seemingly longer and longer delay before operating systems developpers begin to think about supporting a "real life" UNICODE makes waiting solutions inevitable...

[9] - This choice was however not thought as totally binding for the future : 4D was available on Mac, but a PC version ("4D Universal") was expected to be developped by ACI for Windows (it is in beta version at the time of writing). As some UNICODE tools were being developped (GAMMA), it opened the perspective of a future port to PC.

- while not being cheap, prices for 4D products do not go to the heights of some professional relational database softwares, and there are substantial academic discounts.

IV) *DIREJÎ KURDÎ* : A GENERAL DESCRIPTION OF THE PROTOTYPE

The development started in January 1995. In its current state (November 1995), the software is still oriented mainly towards using Arabic-Persian script as the main writing system, but it already has Latin transcription capacities and Cyrillic is being implemented. It may be used either in the *development environment* (that is, by using the functionnalities supplied by the 4D engine) or in a first version of the *user environment*, in which the following functions have been added :

1. Importation / exportation of data from /to a flat file dictionary with control/edition by user, automatic creation of links upon importation ;
2. Search/display of Kurdish (Sorani) and French words in a list, or alone with all links (translations, synonyms, linked expressions) in the script of choice for Kurdish ;
3. Exact transcription for Kurdish from Latin/Cyrillic to Arabic script, near-exact from Arabic to others.

The data of the dictionary flat files is being used to validate those functions and the adaptation of this first version of the data structure to lexicographic work.

1-The prototype database structure

From the lexicographic point of view, this structure is still rather rough. It has, however, enabled testing the system, especially for importation, automatic transcription, and retrieving of linked data. The principle choosed is that each language wordlist (French, Kurdish) is separated and structured in its own way. Then links are created between fields of different / same wordlists.

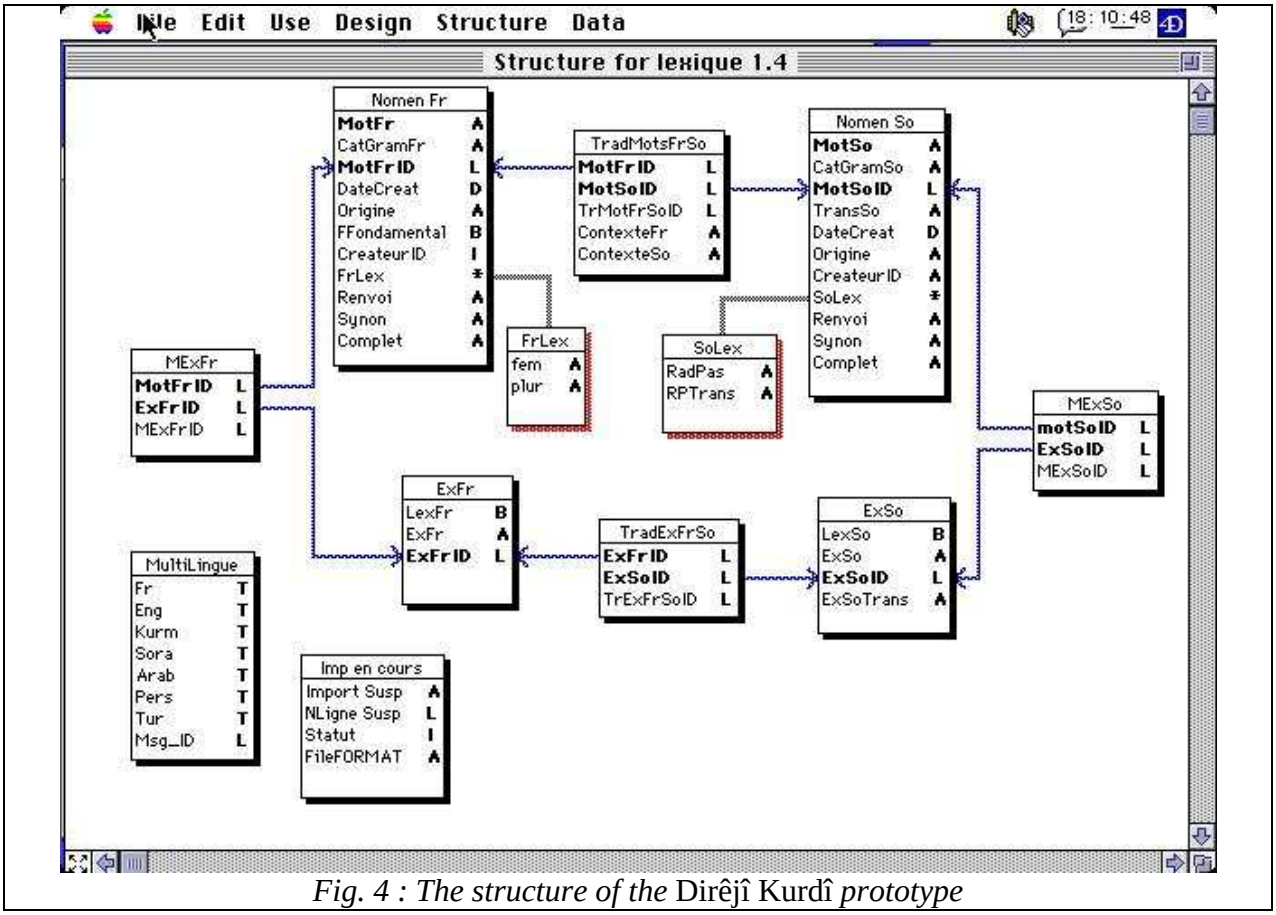


Fig. 4 : The structure of the Dirêjî Kurdî *prototype*

So, according to those principles, French and Kurdish wordlists are stored each in a file, Nomen Fr and Nomen So, whose structure reflects the specifics of the language : as Sorani verbs have a different

radical form for past and present tenses, two fields in a Lexical sub-file, called SoLex, have been created, linked to the infinitive form. In French, morphological modifiers depending of gender or number (*belle*, *belles*, *beau*, *beaux*) are kept in a similar FrLex file[10]. Several indicators also allow to record : who registered the data, from which document the data came, at which date it was created etc...

Possible translations of a word are *links* between the two wordlists, mediated by a file named TradMotsFrSo (= translation words French Sorani), which also keeps the *context* of the linking : domain of use ("physics", "art") or an indication distinguishing several meanings. This intermediate file allows to link *one* Kurdish word to *several* French words or/and the opposite. Similarly, two separate files for French and Kurdish *expressions* are related together and to the words files through linking files (an indicator records it if an expression is lexicalised).

The 2 files without links on lower-left are *service* files. Imp en cours (Current Importation) stores the status of the importation in progress in case of an accidental interruption, allowing the system to propose the user to resume the work exactly where it was left. Multilingue (Multi-lingual) is scheduled to store online Help messages in several languages when this facility will be implemented.

2-Functions for consultation : visualising wordlists

In 4D, the data contained in the base is presented to the user through *layouts*. Each layout has specific characteristics, which may be defined by the programmer who builds them. The *list-oriented* layouts of 4D provide a convenient way to visualise a wordlist, as seen below.

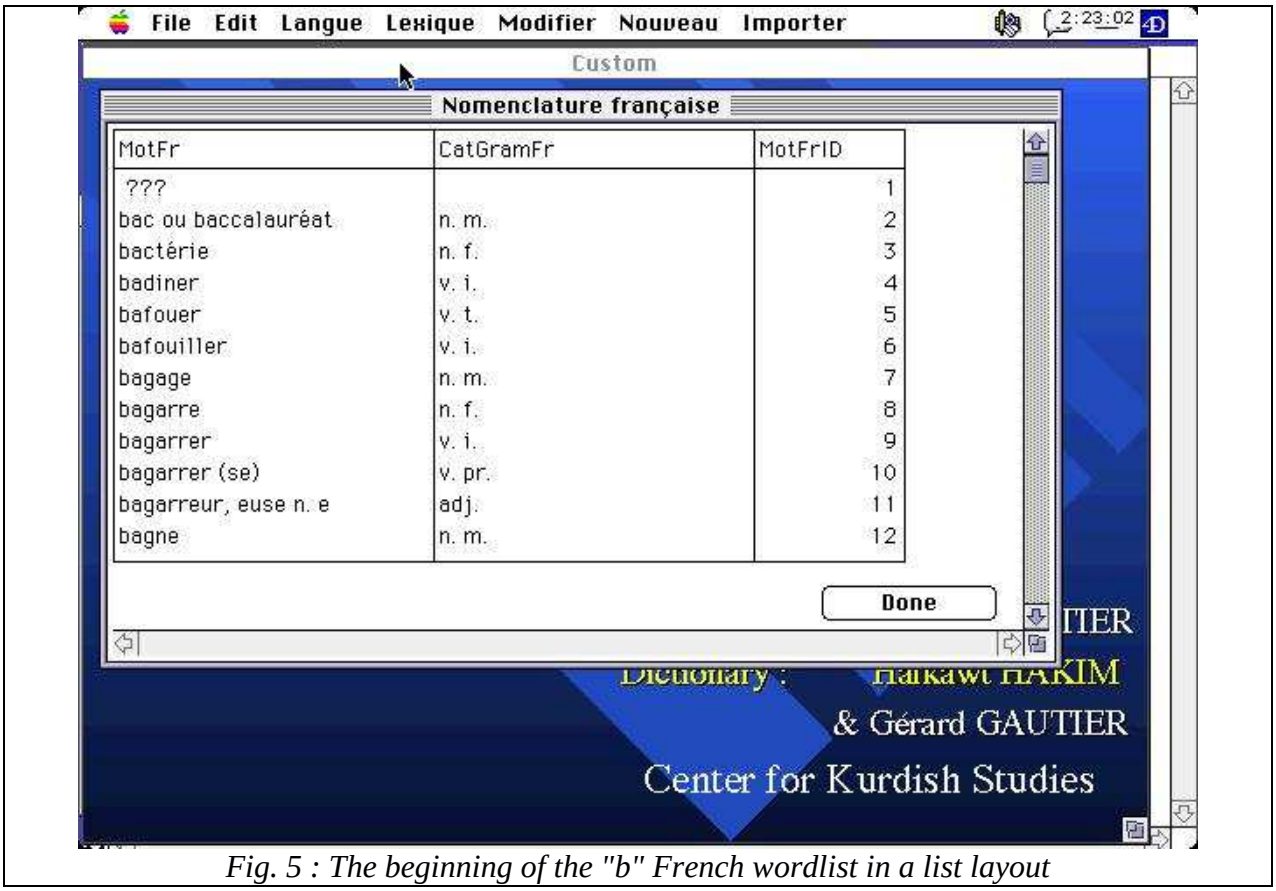


Fig. 5 : The beginning of the "b" French wordlist in a list layout

Upon creation of a new wordlist (i.e at the first use of an empty database), two special French and Kurdish words are automatically created, written "???". This is an "Orphanage" system : if, when importing from a flat file a word of the "A" language, the software is not able to create a link

[10] - The 4D system allows to modify a structure without risk for the data (another reason for its choice), so increasing the number of fields of any file or sub-file as the need is felt may be done ar any moment, thus making room for other lexical markers.

to any word of the "B" language, it will automatically link it to the "???" word of the "B" wordlist file. So the "orphaned" will be easy to find later, and could even be automatically put into a specific list always available...

Below is the equivalent layout for visualising the Sorani wordlist, with exactly the same "Orphanage" system. As the font used to display a field of textual data may be determined by the programmer when he builds a layout, from the user point of view, changing a font may be as easy as choosing a different layout to look at the same data.

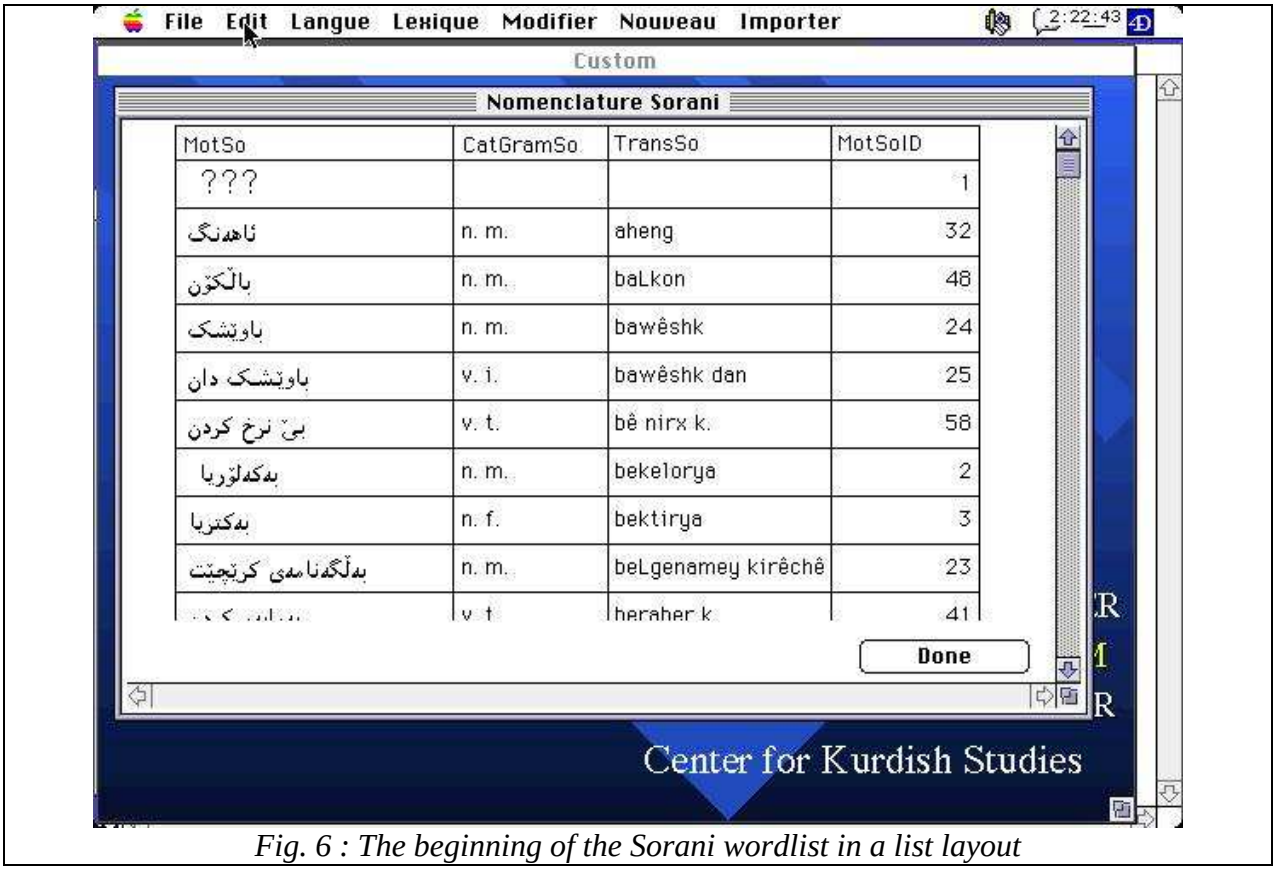


Fig. 6 : The beginning of the Sorani wordlist in a list layout

Here, the Sorani data is sorted after its latin transcription, which is hardly satisfactory for a Sorani-speaking user ! The Baghdad Kurdish Academy defined a collating order for Kurdish, different from both Persian and Arabic. As there is no Kurdish Macintosh system, a Kurdish collating function has been built *inside* 4D. For each Sorani word, it calculates a sorting value which is stored into a (hidden) field. If the user chooses to sort the data after this field, Kurdish words appear in the correct order[11].

3- Functions for consultation : visualising data linked to an entry

Another type of layout allows the visualisation of a *single entry* of a file (hereafter, "main file"). Fields from other files may also be displayed in the layout. When there is data in those fields which is linked to the entry of the main file, the 4D engine will retrieve them automatically. This provides a convenient way to see all the links of any word or expression. The user may search explicitely for a word, or double-click on any word in a list layout.

If several there linked words, they are placed into a *selection*, through which the user may skim by clicking on the arrows at left. If only one word matches the research, the word layout will be brought up. The two next illustrations show word layouts for French and Sorani words, with their possible translations, linked expressions and the Kurdish/French translations of them.

[11] - In fact, the very function generating the latin transcription of Sorani words also builds the sorting number.

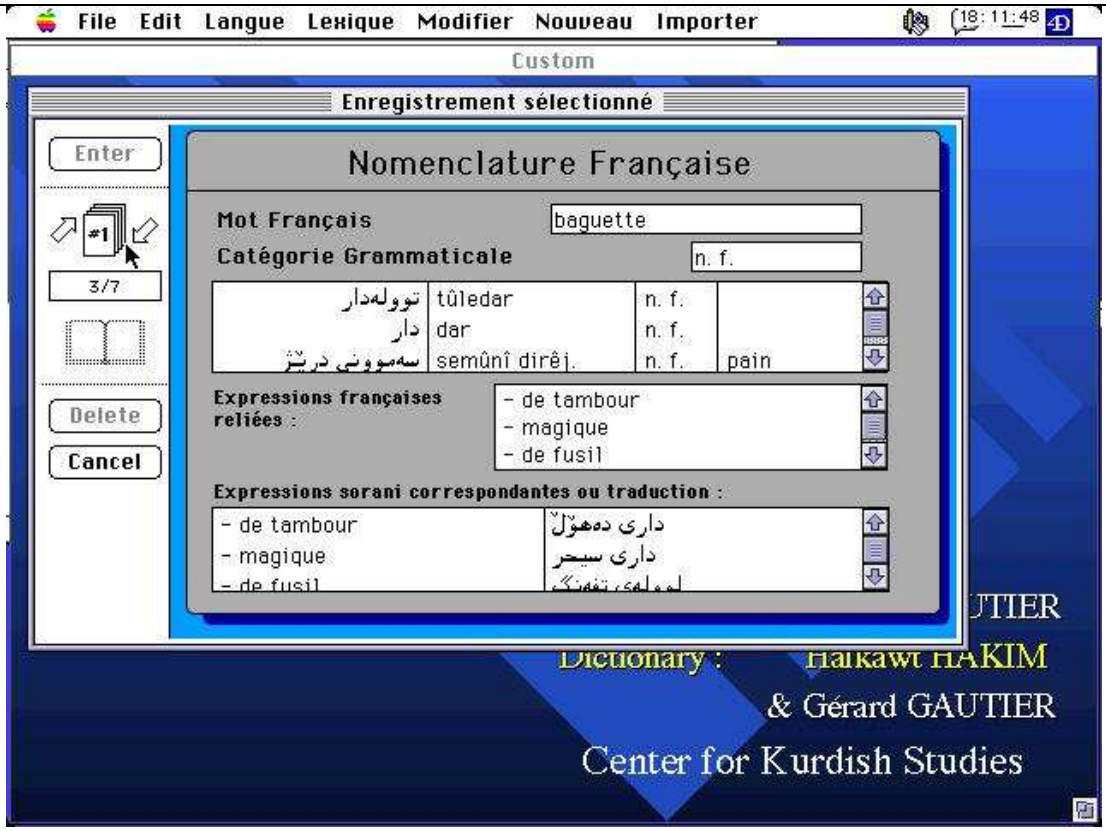


Fig. 7 : A French entry with its links displayed.

The french word “pain” (bred) above, at the right of the Sorani expressions, is a “context” word imported from the flat file, used to distinguish between several meanings of the french “baguette” (stick, drumstick, chopsticks, and a kind of long bred).

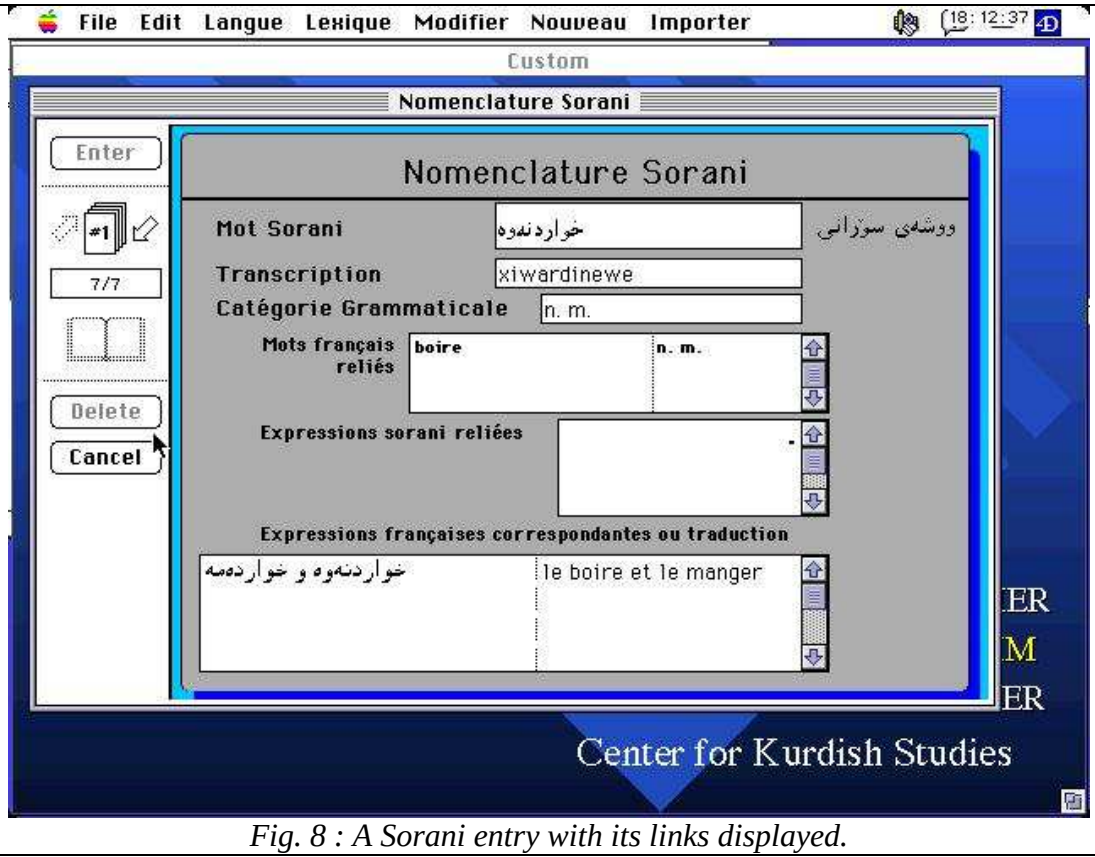


Fig. 8 : A Sorani entry with its links displayed.

4- Functions for lexicographic research and creation : importation

As it was necessary to import a minimum of data into the database to experiment, the first function realised has been an importation function. To quickly create a minimum database, a proprietary formating necessitating only a minimum preparation of the files has been adopted. An analysis of the structure of the paper dictionary found it fairly repetitive, following the formal pattern below, and it was decided to implement a mini-parser in 4D language for the corresponding formal grammar, recursively defined below :

<Entry>	⟶	<Headword>**[tab]** <Article> **[return]**
<Headword>	⟶	<FRENCH HEADWORD> (**[comma]** <MODIFIER>) **[space]** <PART-OF-SPEECH>
<Article>	⟶	<Word Descriptor> (**[tab]** <Expression Descriptor>)
<Word Descriptor>	⟶	(**[opening parenthesis]** <CONTEXT OF USE> **[closing parenthesis]**) **[space]** <KURDISH WORD> **[space]** <TRANSCRIPTION> (**[comma]** <Word Descriptor>)
<Expression Descriptor>	⟶	<FRENCH EXPRESSION> **[semicolon]** <Kurdish Expression> (**[tab]** <Expression Descriptor>)

(xxx) :	Optional object	<FRENCH HEADWORD> :	Lexical data (final vocabulary)
[tab] :	Separator	<Word Descriptor > :	Intermediate vocabulary

Fig. 9 : Formal Grammar describing the structure of the flat files.

The weakness of this method is obviously that it is only adapted to specifically prepared files. In the compiled version, this function will accept parameters to suit any other parsing system, and specifically according to a Text Encoding Initiative (TEI) compliant SGML DTD [see *Text Encoding Initiative* in bibliography]. The files to import were semi-automatically prepared to conform to this structure, through the use of *Quickeys* macros :

1- Replacement of the underlined characters (h̲, r̲) by capital letters (H and R) ;
2- Copy of the WINTEXT files into strictly text files (hence loosing all the typographic information, as fonts and size, but also "underlined") to allow for importation into 4D ;
3- Separation of the contents of each article by markers corresponding to the formal grammar.

The importation screen called by this function is shown in the next page.

4- The process of controlled importation

The data is not immediately "poured" into the database. It is first stored in a variable mirrored in the upper field of the importation window (hereafter, "Entry field") :

bagage n. m. شت و مەکی سەفەر sHit u mekî sefer. ; plier - هەڵهاتن. مردن

The user may hence modify the data at will, either by editing it directly in the Entry field, or by clicking on one of the two tables below, which propose ways to structure the data in the database which will be carried out when the "Importation" button on the window will be depressed :

Mots kurdes correspondants	Transcriptions	Contextes d'emploi
شت و مەکی سەفەر	sHit u mekî sefer.	

The little window on top shows the current action (“Looking up next Kurdish word” etc...).

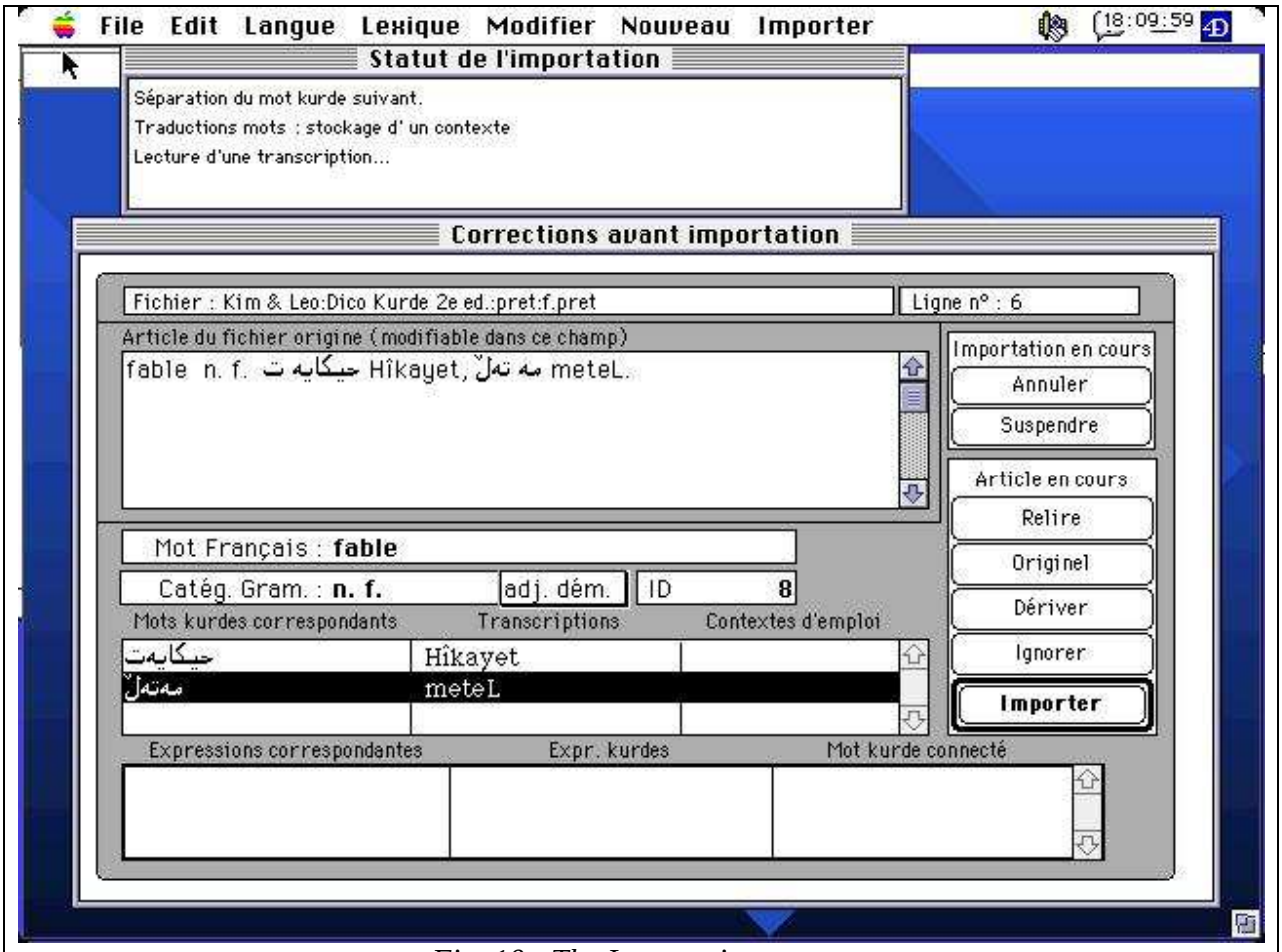


Fig. 10 : The Importation screen

For instance, if the user wants to correct the first line of the first table, he may click on it to bring up this “Word correction window” below :

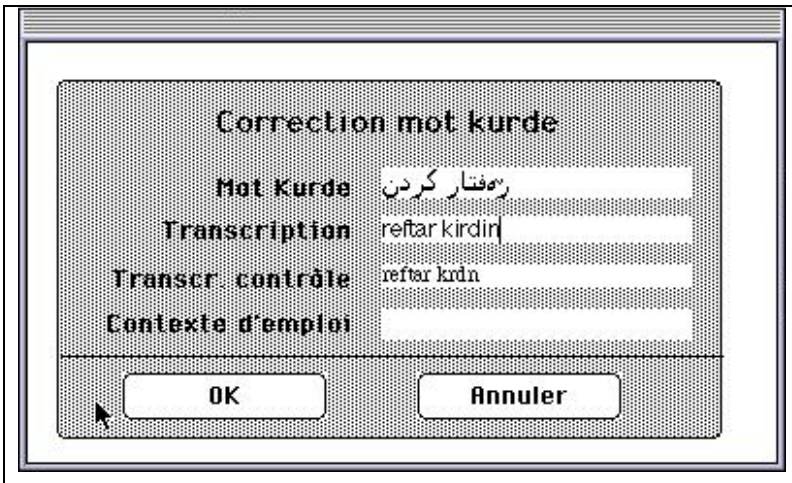


Fig 11 : The “Word Correction” pop-up window

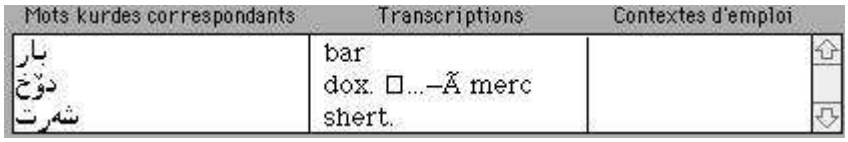
As it is sometimes difficult to be sure of the real order of letters when Arabic and French text are used together, the window provides an automatic transcription control field. This transcription is not designed to be 100% accurate, but gives a way of verification. If the control transcription is already accurate, it may be directly copied into the Transcription field which is the one which will be copied into the database.

If the user copies or cuts anything, for the same reason, a message window containing the chunk of data selected in Arabic font will appear to him.

Finally, in the importation window, seven buttons allow to fine-tune the processing of the entry which appears in the field. In the compiled version of *Dirêjî Kurdî*, in order to obtain a less cluttered user interface, these buttons will be done away with and replaced by a menu linked to the importation layout.

1. "**Cancel**", allows the user to totally stop the importation process.
2. "**Suspend**", stops the process and stores its status in the "Importation" file, which allows to resume it at will and at the same point.
   The buttons below concern the current entry.
3. "**Read again**" re-reads the Entry field (and *not* the flat file : may be used after corrections if a first parsing of the entry yielded errors).
4. "**Original**" cleans the entry field of any unsatisfactory modification by retrieving the unmodified text from the flat file.
5. "**Derive**" imports the data from the Entry field into the base, then re-reads the original entry from the flat file (allows for instance to split an original entry into several entries, imported separately).
6. "**Ignore**" allows the user to jump over this particular entry (for instance, if the data is already in the base, as signaled by the popping up of a "Doublon" (double) marker on the window).
7. "**Import**" allows to import and go to the next entry in the flat file.

By combining the capacities of edition of the original data with the functions of those buttons, the user may suit the data to its wishes. For exemple, below, an error in the preparation of the flat file translated into a wrong separation between transcription and Kurdish word :



The garbage in the "Transcription" window (which is in fact part of the Sorani word mistakenly journeying into the next field) can be corrected by one of the two methods :

1- Modifying the Entry field, then clicking "Read again" (if the user can see what caused the error);

2- Clicking on the second line of the table and correcting it manually - if the cause of the error remains unclear (here, a lack of a *Latin* space - and not an *Arabic* one...- between word and transcription).

V) TOWARDS THE FINAL SOFTWARE

This work was undertaken with in mind the idea that it should emulate on a computer what lexicographers usually do with paper : collecting data from (electronic) texts (including other lexicographic works), structuring it, (construction of articles), manipulating it (choice and change of the mode of representation), and exporting it (to a word-processor, another database, a pagesetting software) as a whole or as a selected subset. This is with this orientation in mind that the extensions described below have been chosen. I will focus here on the aspects of those extensions which are the most related to the orientation of this conference.

1- Increasing user-friendliness

If *Dirêji Kurdî* must be of use to other people - and not only to its author - giving them the means to make good use of it is an integral part of the project. A general "user environment" should encompass :

- the *software environment proper*, which should allow a non programming user (but already familiar with, say, a word-processing software) to enter new data into and retrieve data from the base,
- an introductory *User Manual,*
- for the advanced user, a *Programmation Report,* including the code of all the functions realised and the necessary technical comments.

It is obvious that, if *Dirêjî Kurdî* has to be used by people not skilled in computers, there is also room for improvement in the the friendliness of the user interface in general, as well for the consultation of data as for its editing and importation. Several third-party adds-on for 4D exist that could allow for that. One specific area of possible extention would be to add hypertext features to the data fields... On the multilingual level, it is planned to let the user *choose the language* of the interface at startup. It is possible, because a 4D application is able to retrieve its menus, layout titles and explanation texts from its own files, and can dynamically change their fonts. A specific file will be used to store on-line help messages in several languages. The language the user will select for the general interface will be at the same time selected as the help language.

The last phase of the realisation will be the *compiling* of the software, which will increase its speed (another way to increase user-friendliness) and will allow its distribution as a shareware.

2- Stabilizing overall structure

The most important problem in the development of a lexicographic software is the definition of the overall structure of the database. But as the main subject of this paper is the specifics of multiscript - specially as Arabic type is concerned - I will give only a cursory description of it, and will focus on the transcriptions and importation/exportation problems.The first phase of the project, the “technological survey”, enabled to look not only at multilingual systems and relational database engines, but also at ongoing research and terminological and lexicographic software of the market . At the same time, several Kurdish dictionaries were analysed[12]. This is from this phase of the work that was decided what the final software should look like.

A lexical database should store not only words but the *relations* that take place between them. [D.A. Cruse, 1986, pp 86 sq]. Those relations may be relations between the meaning of words. Inside a language : synonymy, hypernymy, hyponymy ; across languages : possibility of translation. In turn, a possibility of translation of one word into another often exists only through a *context of use* (field of knowledge, specific sentence), which should also be recorded.

There may also be relations of *proximity*, constitution of new meanings through assembling : words composed from other words, expressions, lexicalised or not, current collocations [Rey, 1977, pp. 188sq].

Each element must first be uniquely identified and stored [Mannila & Räihä, 1992, p. 71]. Once each entry of the database has been given a unique ID number, storing a relation between one entry and another becomes as simple as manipulating those ID. There are basically two methods :

- recording a cross-reference from one entry to another as an ID in a field; for example, all synonyms of an entry may have their ID stored in a Synonym field for this entry ;
- recording a type of relation by writing related IDs in a specific Synonymy file, that is a “joining” file as has been already used for the recording of translation possibilities.

Another problem is to build a structure which allows for future extension : adding another relation, or even another language, must be possible with a minimum of interference in the already existing structure. In other words, the structure must be of maximum modularity. That is why the second approach : [one relation = one file] has generally been chosen.

A good structure modularity would allow, for instance, for the introduction of English, simply through the adding of another "linking file" with another database structure. This other database structure could very well be on CD-ROM (from the market or from a research project).

Another area of extension would be links to *text corpora* in 4D or even external (text) files. A specifically interesting area of development in this respect would be *the constitution of a corpus of electronic Kurdish text*, as (to my knowledge) no one exists yet.

---

[12] - Consulted Kurdish dictionaries will be found in bibliography. Software is in a specific section of bibliography.

On the other hand, a *property* specific to an entry (as, for an expression, if it is or not lexicalised) should be stored as an *attribute*, together with the entry in a field, a boolean or other. An exemple of such a property is etymological data. Among the Kurdish dictionaries which have been consulted in the first phase of the project, only [Wahby & Edmonds, 1966] gives etymological information, and this information is limited only to the rough origin of the word when it is not Kurdish : **A** for Arabic, **P** for Persian etc. Certainly a specific field for etymology should be reserved in the Kurdish database, and if it is to store corresponding words in several dialects of Kurdish, this data should be very clearly marked. A list of possible origins and dialects (possibly hierarchical) should be first defined in the basis version of the software, but the user should be able to expand it at will.

I already gave some examples of inter-dialectal variations. I think the area of inter-dialectal research, particularly specific to Kurdish, could greatly benefit from *Dirêjî Kurdî.*

3- Rationalising the transcription system

The main writing system in the database is now still the Arabic-Persian one, with an automatic system of transcription to one latin, specific, system of writing, the one used in [Wahby & Edmonds, *op. cit.*].

But the Arabic-Persian writing, as already explained, lacks the schwa. That means that a word as كردن is susceptible of two renderings in latin letters : (1) `kirdin`, or (2) `kridin`. It is obviously not only a matter of writing, as those two writings would not be pronounced the same way. A Kurdish speaker will know that the good answer is (1), but not a 4D function... What we need is a more complete transcription system, a "mother" transcription from which any other writing system could be deduced.

As there are sounds which exist in some Kurdish dialects, but not in others, none of the transcriptions elaborated for one single dialect will do (Although the closest would certainly be the one used by Rizgar). A more general system is needed.

The International Phonetic Alphabet (hereafter IPA) sounded at first like a good candidate for such a task. From it, given precise rules, it is certainly possible to generate any other system of writing. But it encodes *pronunciation*, not writing, and is in a way much too precise. It would already be impossible to write English in IPA, because it has as many pronunciations as places where it is spoken. For the multi-dialect Kurdish, the problem is certainly bigger, as it is well examplified by what Rizgar writes in his introduction :

> *"Underlining [the accented letters and the letters e, h, and x when they are pronounced like Arabic ع , ح and خ has not been used for words I have only seen in texts and therefore not heard them being pronounced."* [Rizgar, op. cit., p. 9]

Another problem is when *one* IPA character matches *several* characters in a Kurdish writing system (s → س or ص ? t → ت or ط ? Significantly, this problem arises in the case of Arabic "imported" words). Hence, it will be probably better for clarity to keep pronunciation (in IPA), and a general transcription (which could look like a *modified* IPA) *separately* to distinguish these cases.

That does certainly not mean that the user will have to type new words in this "modified IPA" ! He must still be able to type in the transcription he wants, but the system should then present him with a proposal in IPA and ask him to correct it (more on this problem in the next section).

The user should be able to define a transcription through intervention at two levels :

1- to define a *filter* which will possibly transcode the data before presentation in a layout,

2- to choose the *font* he wants to use to visualise data.

The system should come with several prepared filters and fonts, but the user should not be restricted to them. The play of different filters could be used to choose between several scripts systems (i.e. Arabic-Persian, Latin, Cyrillic), and the choice between fonts could allow the user to choose (for instance) between several variants of latin writing etc.

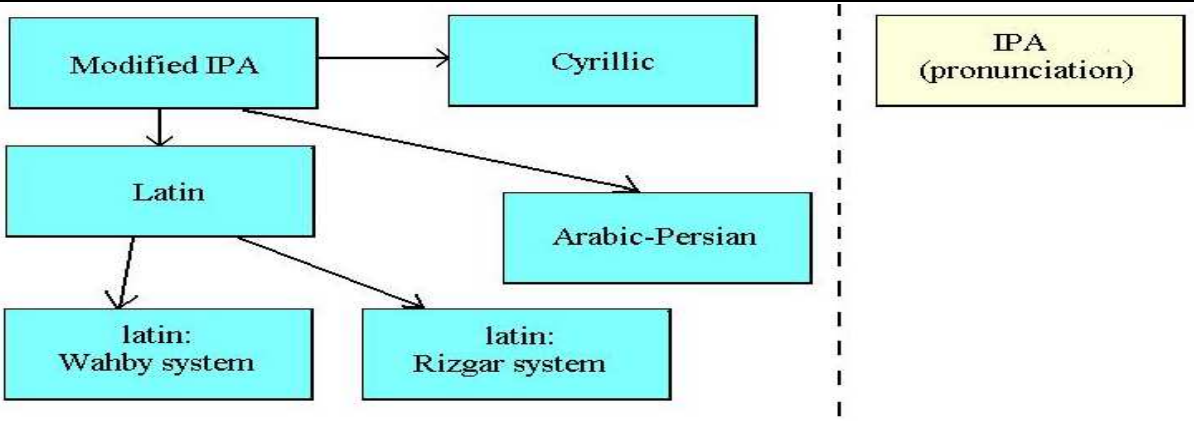


Fig. 12 : An example of the possible choices for transcription of Kurdish data

The way a word will be presented to the user should depend of two things :

1- Parameters linked to the word : dialect, sub-dialect etc... This set of relations between fonts and dialect should constitute a default, recorded in the setup file of the software, and modifiable by the user,

2- The user choices, which could override the default for any single word, and should be recorded at word level.

But there are two points that must be stated very clearly :

1- As I discovered in writing the Sorani collating order procedures and the procedures for transcription from Arabic to Latin writing, character-to-character filters are enough to perform automatic transcription only if an unambiguous writing is first chosen as the basis for every presentation of the data. So the user will modify the default settings of the system at his own risk.

2- In the mind of the user, there should be a very clear separation between the *internal encoding* and the *presentation level* . But, in conformity with the TEI prescriptions, a very definite separation must also be made between the manner a word is processed and presented *inside Dirêjî Kurdî* on one side, and the encoding under which it is *imported or exported* to / from other systems of representation (other lexical databases,word - processors, pagesetting systems etc.).

A consequence of this complexity is that the system should probably be protected from the manipulations of users for whom those distinctions are not clear, and who would be at risk of harming the validity of the data. Fortunately, 4D provides a “login” facility and security features which enable restricting the access to whatever functionalities the “owner” of the database deems insecure (that was also another reason for the choice of 4D).

4- Recording the state of the work in progress

I mentionned that the user may be asked by the software to correct the IPA transcription. He may well be (as Rizgar would be in some cases) unable to do so. That brings up a problem characteristic of any work of the scope of a dictionary : *time* . There are parts of the research which will stay in a "in-between" stage for an long period of time, before being finally completed when the necessary data becomes available. That means the database will certainly need some sort of indicator encoding the degree of completion of the work on an entry, a transcription, a semantic relation etc...

This is the type of situation which is highly problematic on paper, because an unfinished job must be stored in a specific stack - or in human memory - to be easily retrieved. But in a database, this state of being “in progress” may be easily stored in an “indicator of completion”. The final product should be then able to record the degree of completion estimated by the user for any of the

database components. As already described for others parts of the software, a list of "basis" indicators should be prepared, which will be loaded at startup. It should also be extendable by the user.

5- In conclusion : exchanging data with the outer world

Any work done in *Dirêjî Kurdî* must not be locked inside, otherwise, the very purpose of the project would be defeated. So, giving the user versatile ways to import and export data must be a very important part of the final version.

In 4D, there are several ways to import or export data :

1- through customised procedures, directly from or into text files. This way was chosen to import data from the flat files of the French-Kurdish dictionary ;

2- through specific importation and exportation layouts (hereafter, "exchange layouts").

At encoding level, 4D allows the use of transcoding *filters* during exchange. The user should be allowed to choose a filter in a list – or to build a custom one – before exchange. For importation from external text files, the software should be able to recognize any TEI-compliant header placed in the file and if necessary load the necessary transcoding filter referenced in the header.

At the structure level, the choice of the exchange layout allows to choose a specific structure. The simplest importation procedure should be able to import a wordlist (or to export a wordlist or part of it after various selection criterions), but more complicated functions as importing / exporting already structured articles shoud be accessible.

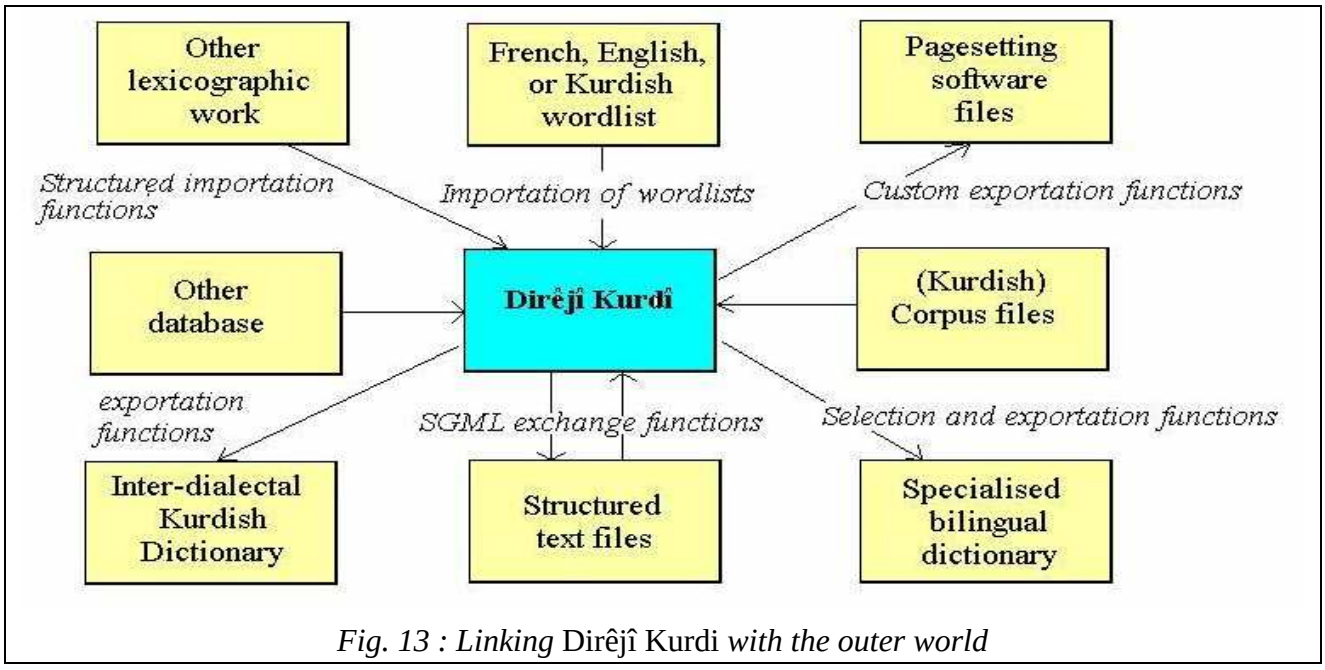


Fig. 13 : Linking Dirêjî Kurdi *with the outer world*

By linking each exchange layout with a specific DTD (Document Type Definition), it would be rather easy to make the exchange of data compliant with the TEI proposals. In turn, with more and more word processors developpers releasing SGML-compliant interfaces, the possibilities of exchange should grow steadily.

The same layout mecanism would also allow to insert into the exportation flow custom control characters specific to a typesetting software, but that would again lock the data in a proprietary format, the very situation to avoid which the development of *Dirêjî Kurdî* was started.

Appendix A

A tentative chart of corresponding Kurdish letters and IPA

IPA	Arabic-Persian (used in Iraq / Iran)		Latin (used in Turkey)[1]		Cyrillic (used in Armenia)	
	Glyph	UNICODE name / code	Glyph	UNICODE name / code [2]	Glyph	UNICODE name / code
aː	ا / ئا	Arabic letter alef / 0627	a A	Latin letter a	а А	Cyrillic small letter a / 0430 *and* capital / 0410
b	ب	Arabic letter baa / 0628	b B	Latin letter b	б Б	Cyrillic small letter be / 0431 *and* capital / 0411 [3]
ʤ	ج	Arabic letter jeem / 062C	c C	Latin letter c	щ Щ	Cyrillic small letter shcha / 0449 *and* capital / 0429
ʧ	چ	Arabic letter haa with three middle dots downwards / 0686	ç Ç *ch*	Latin letter c cedilla / 00E7 and capital / 00C7	ч Ч	Cyrillic small letter che / 0447 *and* capital / 0427
ʧʼ	چ	Arabic letter haa with three middle dots downwards / 0686	*ç' Ç'* *ch*	Latin letter c cedilla+ non spacing comma above right / 0315	чʼ Чʼ	Cyrillic small letter che + non spacing comma above right / 0315
d	د	Arabic letter dal / 062F	d D	Latin letter d	д Д	Cyrillic small letter de / 0434 *and* capital / 0414
d [4]	ض	Arabic letter dad / 0636	*ḍ Ḍ*	Latin letter d + non spacing dot below / 0323		
a	ه	Arabic letter ae / 06D5	e E	Latin letter e	ə Ə [5]	Latin small letter schwa / 0259 *and* capital / 018F
ʕ	عه/ع	Arabic letter ain / 0639	*e' E'*	Latin letter e + non spacing comma above right / 0315	əʼ Əʼ	Latin letter schwa + non spacing comma above right / 0315
ɛː	ئێ/ێ	Arabic letter ya with small v / 06CE	ê Ê	Latin letter small e circumflex / 00EA *and* capital letter / 00CA	е Е	Cyrillic small letter ie / 0435 *and* capital / 0415
f	ف	Arabic letter fa / 0641	f F	Latin letter f	ф Ф	Cyrillic small letter ef / 0444 *and* capital / 0414
g	گ	Arabic letter gaf / 06AF	g G	Latin letter g	г Г	Cyrillic small letter ge / 0433 *and* capital / 0413
h	ھ	Arabic letter knotted ha / 06BE	h H	Latin letter h	һ Һ [6]	Cyrillic small letter h / 04BA *and* capital / 04BB
ɦ	ح	Arabic letter haa / 062D	h H *ḥ Ḥ* *ḧ Ḧ* *h' H'*	Latin letter h + non spacing dot below / 0323 + non spacing diaeresis / 0308 *or* + non spacing comma above right / 0315	һʼ Һʼ	Cyrillic letter h + non spacing comma above right / 0315

Notes:

1- *Italics* denote "orientalists' transcription" (for instance used in dictionaries) and not the standard Kurdish writing.

2- For Latin letters, only codes for *non-ASCII* characters (extended ASCII) have been written (ç, ê etc).

3- The Cyrillic small letter <be> glyph differs from Unicode chart.

4- This Arabic letter is not currently used in Sorani. I cannot ascertain its pronunciation in "imported" Arabic words.

5- Does not exist in standard Cyrillic fonts, nor in Unicode Cyrillic codepage.

6- Does not exist in standard Cyrillic fonts. In Unicode, it is in "extended Cyrillic".

A tentative chart of corresponding Kurdish letters and IPA (part 2)

IPA	Arabic-Persian (used in Iraq / Iran)		Latin (used in Turkey) [1]		Cyrillic (used in Armenia)	
	Glyph	UNICODE name / code	Glyph	UNICODE name / code [2]	Glyph	UNICODE name / code
ɨ	Ø[7]/ئ	Arabic letter hamza on ya / 0626	ı I i İ	Latin small letter dotless i / 0149 *and* cap. letter i *or* Latin small letter i *and* Latin cap. letter i dot / 0130 [8]	ь Ь	Cyrillic small letter soft sign / 044C *and* capital / 042C
i:	ی	Arabic letter dotless ya / 06CC	i İ î Î	Latin small letter i *and* cap. letter i dot / 0130 *or* Latin small letter I circumflex / 00EE and capital letter / 00CE [9]	и И	Cyrillic small letter ii / 0438 *and* capital / 0418
ʒ	ژ	Arabic letter ra with 3 dots above / 0698	j J	Latin letter j	ж Ж	Cyrillic small letter zhe / 0436 *and* capital / 0416
k’	ك	Arabic letter caf / 0643	k K [10]	Latin letter k	к К	Cyrillic small letter ka / 043A *and* capital / 041A
	ک	Arabic letter open caf / 06A9	*k’ K’*	Latin letter k + non spacing comma above right / 0315		
l	ل	Arabic letter lam / 0644	l L	Latin letter l	л Л	Cyrillic small letter el / 043B *and* capital / 041B
ɬ	ڵ	Arabic letter lam with small v / 06B5	[11] *ɭ ɭ* *ɬ L̂*	Latin letter l + non spacing dot below / 0323 Latin small letter barred l / 019A *and* Latin cap. letter l + non spacing circumflex / 0302	л’ Л’	Cyrillic letter el + non spacing comma above right / 0315
m	م	Arabic letter meem / 0645	m M	Latin letter m	м М	Cyrillic small letter em / 043C *and* capital / 041C
n	ن	Arabic letter noon / 0646	n N	Latin letter n	н Н	Cyrillic small letter en / 043D *and* capital / 041D
o	ۆ	Arabic letter waw with small v / 06C6	o O	Latin letter o	о О	Cyrillic small letter o / 043E *and* capital / 041E
œɛ	وێ	*(written as two separate letters)*	wê [12] *ö Ö*	Latin small letter o diaeresis / 00F6 *and* capital / 00D6		*(Probably also written as two letters in Cyrillic)*

Notes:
7- **Ø** denotes the null character (not written inside words).
8- It must be noted here that the corresponding small and capital letters are not the usual ones, as the capital "I" must have a dot... (see also note 9).
9- In publications where there is a dotless i, the î is not used and the normal dot i is used instead. Otherwise, the opposition is between i and î.
10- The difference between ك and ک is in style of writing only and may be processed at the *font* level, whatever UNICODE chooses to say. Chyet [*op. cit.*] gives two Roman transcriptions for this letter.
11- This sound does not exist to my knowledge in Kurmandji.
12- The “ö” notation is used only in the works using the [Wahby & Edmonds, op. cit.] Latin transcription.

A tentative chart of corresponding Kurdish letters and IPA (part 3)

IPA	Arabic-Persian (used in Iraq / Iran)		Latin (used in Turkey) [1]		Cyrillic (used in Armenia)	
	Glyph	UNICODE name / code	Glyph	UNICODE name / code [2]	Glyph	UNICODE name / code
p’	پ	Arabic letter taa with 3 dots below / 067E	p P	Latin letter p	п П	Cyrillic small letter pe / 043F *and* capital / 041F
			p’ P’	Latin letter p + non spacing comma above right / 0315	п’ П’	Cyrillic letter pe + non spacing comma above right / 0315
q	ق	Arabic letter qaf / 0642	q Q	Latin letter q	q Q	Latin letter q [15]
r	ر	Arabic letter ra / 0631	r R	Latin letter r	р Р	Cyrillic small letter er / 0440 *and* capital / 0420
r	ڕ	Arabic letter ra with small v below / 0695	*r̄ R̄* *ṛ Ṛ* *r̲ R̲*	Latin letter r + non spacing macron / 0304 *or* + non spacing dot below / 0695 *or* non spacing macron below / 0331	р’ Р’	Cyrillic letter er + non spacing comma above right / 0315
	ڑ	Arabic letter ra with small v / 0692 [13]				
s	س	Arabic letter seen / 0633	s S	Latin letter s	с С	Cyrillic small letter es / 0441 *and* capital / 0421
s	ص	Arabic letter sad / 0635	*ṣ Ṣ* *s̲ S̲*	Latin letter s + non spacing dot below / 0323 *or* + non spacing macron below / 0331		
ʃ	ش	Arabic letter sheen / 0634	ṣ Ṣ	Latin small letter s cedilla 0323 *and* capital / 015F	ш Ш	Cyrillic small letter sha / 0448 *and* capital / 0428
t	ت	Arabic letter taa / 062A	t T	Latin letter t	т Т	Cyrillic small letter te / 0442 *and* capital / 0422
t’			t' T’	Latin letter t + non spacing comma above right/ 0315	т’ Т’	Cyrillic letter te + nonspacing comma above right / 0315
(14)	ط	Arabic letter tah / 0637	*ṭ Ṭ* *t̲ T̲*	Latin letter t + nonspacing dot below / 0323 *or* + non spacing macron below / 0331		
u	ئو/ و	Arabic letter waw / 0648	u U	Latin letter u	ö Ö [16]	Latin letter o diaeresis
u:	ئوو/وو	Arabic letter waw	û Û	Latin letter u circumflex / 00FB and capital / 00DB	у У	Cyrillic small letter u / 0443 *and* capital / 0423

Notes:

13- The difference between this and the above is only a matter of writing, and may be decided at the *font* level. The sound ڕ also exists in Kurmanji, but is not normally differenciated in writing from the ر.

14- This letter being normally not used in Kurdish words, no transcription is provided – ط is probably assimilated to the ت.

15- This letter does not exist in the standard Cyrillic font, nor in Unicode (unless if it is the Cyrillic letter "o hook" / 004A8 - 004A9, which I do not believe)).

16- This letter does not exist in the standard Cyrillic font, nor in Unicode (as a Cyrillic letter).

A tentative chart of corresponding Kurdish letters and IPA (part 4)

IPA	Arabic-Persian (used in Iraq / Iran)		Latin (used in Turkey) [1]		Cyrillic (used in Armenia)	
	Glyph	UNICODE name / code	Glyph	UNICODE name / code [2]	Glyph	UNICODE name / code
v	ڤ	Arabic letter fa with 3 dots above / 06A4	v V	Latin letter v	в В	Cyrillic small letter ve / 0432 *and* capital / 0412
w	و	Arabic letter waw / 0648	w W	Latin letter w	ѡ Ѡ	Latin letter w [17]
x	خ	Arabic letter khaa / 062E	x X	Latin letter x	х Х	Cyrillic small letter kha / 0445 *and* capital / 0425
ɣ	خ	Arabic letter khaa / 062E	ẍ Ẍ	Latin letter x + non spacing diaeresis / 0308	г᾽ Г᾽	Cyrillic letter ge + modifier letter non spacing comma above right / 0315 [18]
	غ	Arabic letter ghain / 063A				
j	ى	Arabic letter dotless ya / 06CC [19]	y Y	Latin letter y	й Й	Cyrillic letter short ii / 0439 *and* capital / 0419
z	ز	Arabic letter zain / 0632	z Z	Latin letter z	з З	Cyrillic small letter ze / 0437 *and* capital / 0417

Notes: 17- This letter does not exist in the standard Cyrillic font, nor in Unicode (as a Cyrillic letter).

18- The Unicode modifier "Letter non spacing comma above right" / 0315 has been chosen among several glyphs which looked alike to differenciate it from the apostrophe / 0027, and because the character "modifier letter prime" / 0B29 is specified, among other uses, to be used as a transliteration of the Cyrillic soft sign, which could lead to ambiguities.

19- The behaviour of the "Arabic ya" - ie being or not "dotless" - may be seen as a font attribute (the normal ya behaves on the Macintosh Diwan fonts as a Kurdish ya).

Sources: For the IPA correspondance : [Blau, op. cit.], for the Cyrillic correspondance : [Rizgar, op. cit.] and [Chyet, op. cit.]. For the Latin : several dictionaries and texts, including [Rizgar] and [Chyet].

Appendix B

Some transcriptions from Kurdish dictionaries and publications

According to the transcription of [Wahby & Edmonds, op. cit.], also adopted in [Blau, op. cit.], [Hakem & Gautier. op. cit.], the name of the city of مەهەباد must be rendered as Mehebad.

In [Izady, op. cit.], the transcription Mahâbâd is used, out of concern for a general consistency with transcriptions from other languages.

In [Eagleton, 1991], the French text contains the transcription Mahabad.

[Nikitine, 1956], in his reference work about the Kurds, illustrates another problem, which is the word division, as he writes (p. 202) Mâh-Abâd, and (in his Index) Mah Abad.

It would be difficult to decide if someone is right or wrong : those four works are neither comparable nor do they aim at the same type of readers.

The general problem of faulty transcriptions arise as well for Chinese or Arabic, but here (if we except Nikitine, in the case of whom the date of the first publication is also an explanation for the apparent lack of consistency...) – each choice follows a consistent logic of its own.

BIBLIOGRAPHY

and Sources used for Technological Survey

4D-DIGEST, Internet discussion list on 4D, majordomo@isig.mit.edu

A.C.I. (1994) *4th Dimension Design Reference, 4th Dimension Language Reference, 4th Dimension Tutorial, 4th Dimension User Reference*

Amindarov, Aziz (1994), *Kurdish-English, English-Kurdish Dictionary,* New York : Hippocrene

Apple Computer Inc. (1991) *Inside MacIntosh,* vol. VI Addison-Wesley.

Blau, Joyce (1980) *Manuel de Kurde, Dialecte Sorani,* Paris : Klincksieck

Blau, Joyce & Hakem, Halkawt (1981) *Perles d'un Collier, Textes Kurdes,* Paris : Institut National des Langues Orientales

Chyet, Michael L. (1990) *Standard Kurdish Romanization Table* (personal communication 1995)

CORPORA, Internet discussion list on corpus linguistics, listserv@hd.uib.no

Cruse, D. A. (1991) *Lexical Semantics,* Cambridge : Cambridge Univ. Press

Eagleton, William Jr. (1991) *La République Kurde de 1946,* Bruxelles : Complexe

Feroz, Ahmad (1993) *The making of Modern Turkey*, London & New-York : Routledge.

Gautier, Gérard (1994) *Typing Kurdish on a Computer ?,* unpublished.

Gougenheim, G. (1958) *Dictionnaire Fondamental de la Langue Française,* Paris : Didier

Grogan, Denis J. (1991) "Dictionaries of English : a decade of development" *Journal of Librarianship and Information Science,* 23 (1) March 1991, pp 37-50

Hakim, Halkawt & Gautier, Gérard (1993) *Dictionnaire Français-Kurde,* Paris : Klincksieck.

HUMANIST, Internet discussion list on computers and the humanities, information at : http://www.princeton.edu/~mccarty/humanist or mail to : owner-humanist@lists.princeton.edu

Ilson, Robert (1980) *Dictionaries, Lexicography and language Learning,* Oxford : Pergamon Press

INSOFT, Internet discussion list on software internationalisation, listserv@magellan.iquest.com

ITISALAT, Internet discussion list on Arabic on computer, listserv@huvm.ccf.georgetown.edu

Izady, Mehrdad, R. (1992) *The Kurds - A Concise Handbook*, Washington & London : Crane Russak

Jaba, Auguste (1879) *Dictionnaire Kurde-Français*, publié par Ferdinand Justi, Saint-Pétersbourg : Académie Impériale des Sciences, reed. 1975, Osnabrück : Biblio Verlag

Kreyenbroek, Philip G. (1992) *The Kurds - A contemporary overview*, Routledge, London.

Kurdoev, K.C. (1983) *Kurdsko-Russkye Slovar (Sorani),* Moskow : Russkiie Yazk (Курдско – руский словарь (сорани) москва „руский язык")

LINGUIST, Internet discussion list on linguistics, listserv@tamvm1.tamu.edu

LN, Internet discussion list on natural Language Processing, listserv@pollux.cnusc.fr

Mannila, H. & Räihä, K.-J. (1992) *The design of Relational Databases,* Addison-Wesley

Mc Carus, Ernest (1967) *A Kurdish-English Dictionary - Dialect of Sulaimania, Iraq,* Univ. of Michigan, Ann Arbor.

Multilingual Computing, Sandpoint, ID, USA

Nikitine, Basile (1956) *Les Kurdes*, Paris, Imprimerie Nationale, reed. 1975 Plan de la Tour : Editions d'Aujourd'hui, France

Perlmann, Geoff (1993) *Inside 4th Dimension*, Alameda, CA : Sybex

PC-ARAB, Internet discussion list on Arab on computer, listserv@sakfu00.bitnet

READER, Internet discussion list on Arabic on computer, administritivia : iskandar@ee.tamu.edu

Rey, Alain (1977) *Le lexique : Images et modèles - Du dictionnaire à la lexicologie* Paris, Colin

Rizgar, Baram (1994) *Kurdish-English / English-Kurdish Dictionary, Kurdî-Ingîlîzî / Ingîlîzî-Kurdî,* M.F. Onen, 1993, 45 Wilmot Close, Peckham Hill St., London SE15 6TZ.

Roberts, Diana C. "Integrating an electronic dictionary into a natural language processing system", *Hewlett-Packard Journal*, June 1992, v. 43 n. 3 p. 54 (12)

Sesame Bulletin, Language Automation Worldwide, Sesame Computer Projects, Harrogate, England

Sinclair, John (1991) *Corpus, Concordance, Collocation,* Oxford Univ. Press

TEI-L, Internet discussion list on Text Encoding Initiative, listserv@uicvm.cc.uic.edu

Text Encoding Initiative - texts published in :

- *Computers and the Humanities*, 1995, Dordrecht : Kluwer Academic Press, vol. 29, no 1, no 2, no 3 : "The Text Encoding Initiative, Background and Context"
- *Text and Technology - The Journal of Computer Text Processing,* Dakota State univ. Special issue : Electronic Texts and the Text Encoding Initiative, vol. 5, no 3, Autumn, 1995

Guidelines - ftp://ftp-TEI.uic.edu/pub/TEI *or* ftp://info.ex.ac.uk/pub/SGML/TEI/p3

Unicode Consortium, 1991 *The Unicode Standard*, Version 1.0, vol. 1 & 2, Addison-Wesley.

Wahby, Tawfiq & Edmonds, C. J. (1966 [1971]) *A Kurdish-English Dictionary*, Oxford Univ. Press

ZIFF Computer Select, Database on CD-ROM, Ziff-Davies Inc., Accessible in numerous libraries.

COMPANIES AND PRODUCTS

Products mentionned in this text :

Wintext	is a word processor and a product of the company Winsoft, based in Grenoble (France).
Diwan	refers in the text to the fonts of the Diwan company, bundled with the software Al Nashir al Maktabi.
Decotype	is a company from Netherlands.
4D	(or 4th DIMENSION) is a product of ACI, Paris and ACI-US.
WorldScript	refers to the multilingual system extension for Macintosh by Apple Company ;
Windows	refers to the system software for PC by Microsoft.

Some of the products reviewed during the technological survey :

Gamma Unitype	Unicode engine under Windows, by Gamma, Inc., San Diego, CA
LexPro	Terminological database system for PC and Macintosh, by LCI, Jouy-en-Josas, France.
MultiTerm	Multilingual terminological database for Windows, by Eurolux, Gonderange, Luxemburg.
Mercury-Termex	Terminological and lexicographic database system for PC by LinguaTech, USA.
Lexibase	Tool for the creation of dictionaries under Windows or DOS, distributed in France by Softissimo, Paris.

All the names of products mentionned in this appendix or in the text are acknowledged as Trade Marks of their respective owners.

ACKNOWLEDGEMENTS

Perry BEAM and Simon WHITE, teachers at the Department of English of the Wen-Tzao School of Foreign Languages, reviewed the text of this paper through pages 1 to 18 for English mistakes. Any error left is my own !

Michael L. CHYET provided me from the US with precious information about the Cyrillic Kurdish system of writing.

Halkawt HAKIM allowed me to use the data of the first edition of "our" dictionary for this project, and sent me from France the dictionary of Baram RIZGAR.