Data structuring in the DÉRom (Dictionnaire Étymologique Roman)

Eva Buchi, Xavier Gouvert, Yan Greub

Data structuring in the DÉRom (*Dictionnaire Étymologique Roman*)

Éva Buchi, Xavier Gouvert, and Yan Greub

1. Introduction

The DÉRom (*Dictionnaire Étymologique Roman* or 'Romance Etymological Dictionary') aims to reconstruct the lexicon of Proto-Romance, that is the Latin language as it can be reached by applying the comparative method to the different Romance languages (see Buchi/Schweickard 2008; Buchi 2010)[1].

Any given DÉRom entry presents six constituent parts: (1) the lemma, (2) the documentation (the data section in a narrow sense), (3) the etymological commentary, (4) the bibliography, (5) the signatures, and (6) the publication date; in addition, most entries present (7) a footnote section. This paper considers in particular section (2) containing the documentation or data.

35 out of the 67 currently available entries of the DÉRom –i.e. about half of them– do not present any subdivision within the documentation. This is the case for instance in the entry */'ann-u/ m.n. 'year' (see Celac 2008– 2012 *in* DÉRom *s.v.* */'ann-u/): all Romance cognates directly represent a single Proto-Romance etymon: */'ann-u/ m.n. 'year' (which is widely documented in written Latin as *annum*). This entry does not need any subdivision within the data section.

When we began compiling the dictionary, we thought this would be the normal case. But it soon appeared that it was not: about half of the entries do indeed have subdivisions within the data section. Mostly, the linguistic reconstruction yielded in the compilation of the DÉRom entries requires more or less complex subdivisions of the entries. In what follows, we would like to show the kind of subdivisions we

[1] We are greatly indebted to the entire DÉRom consortium –about fifty scholars from thirteen mostly European countries–, who make this project a most gratifying scientific and human experience. We would like to thank in particular Julia Alletsgruber (ATILF) and Steven N. Dworkin (University of Michigan), who provided very interesting comments on a first draft of this paper.

adopted, and why we did so, and, to some degree, what these lexicographic choices say about the underlying linguistic analysis.

2. Data selection

Data selection in the DÉRom is quite different from data selection in the other major Romance etymological dictionaries, for instance in the *Französisches Etymologisches Wörterbuch* (FEW). The most obvious difference between these two dictionaries, of course, consists in their scope: the FEW is limited to Gallo-Romance, i.e. French, Francoprovençal, Occitan, and Gascon, whereas the DÉRom covers all Romance languages, from Romanian through Sardinian to Portuguese. But this is not what really matters. What really opposes the two dictionaries is the data selection for each Romance idiom: the DÉRom puts aside all data that are not relevant for reconstructing the common ancestor of the Romance cognates, which means for example that derivatives will normally be absent from its entries (they are only mentioned in a footnote if there is no direct testimony of their base). We can add that a large part of the dialectical material is put aside in the DÉRom, or more precisely typified in the form of a (for instance) French cognate.

This choice allows us to skip an important part of the data structuring necessary in the FEW: the need to provide and retrace a great amount of data. There is normally no discrepancy, in an FEW entry, between the linguistic history and the position of a form in the entry. For example, two identical compounds coined independently at different moments would be classified in different paragraphs. But, in most cases, there is no need to distinguish historically the different derivatives, so they can be classified together in a large ‘Derivatives’ section. This is a morphological division, which is subsidiary (not necessarily contradictory, but not necessarily in accordance either) to the etymological analysis.

Having eliminated any and all etymologically irrelevant data, the DÉRom is not confronted, contrary to the FEW, with the problem of classifying data along other criteria than sheer etymology: ahistorical morphological, semantic or conceptual data structuring has never to be used. The function of data structuring in the DÉRom is thus purely etymological.

This represents quite a change in comparison with other synchronic or even etymological dictionaries of Romance languages. In a synchronic one, data structuring is mostly done by criteria like verbal valency, semantic values (in the best cases) or referential context of use. The same goes for multidialectal dictionaries like the FEW or its Italo-Romance counterpart, the *Lessico Etimologico Italiano* (LEI): the need to present all the material of an entire lexical family leads to subcategories unified by referential or formal criteria. Otherwise it would be impossible to retrace the history of the lexico-semantic units. We would like to make clear that this structuring is in fact the best way of presenting the actual functioning of the lexicon in a given linguistic area.

The DÉRom, happily, can avoid these problems by eliminating the vast majority of the data which constitute an FEW entry. This rids the data structuring of a lot of noise[2], as in the entry */plan't-agin-e/ f.n. ‘plantain’, and allows the compiler to concentrate on a more subtle structuring according to really relevant questions. In this case –see Delorme 2012 *in* DÉRom *s.v.* */plan't-agin-e/–, the lexicographer, Jérémie Delorme, distinguishes three different types for French: *plantain* (under III. */plan't-agin-e/), *plantaine* (under II. */plan't-agin-a/) and *plantage*, this third item being put aside as a learnèd form in footnote 8. Thus, data structuring is made clearer not only for the dictionary user, but also for its compiler, who will be compelled to examine the central problem, i.e. reconstructing the protolexeme.

The aim of the DÉRom, a multilingual dictionary itself, does not lie in describing the lexicon of Romance dialects, but in reconstructing the lexicon of their common ancestor, i.e. Proto-Romance. This is made obvious in the entry structure itself, since it is not done accordingly to the best possible description of the different data

[2] However, this noise is not lost for the DÉRom reader: loanwords and inherited lexemes presenting phonetic or morphological irregularities are quite often mentioned in footnotes.

(for example the simplest one), but according to the degree of relevance the different categories have for the reconstruction.

3. Morphological data structuring

Compiling the first 67 entries of the dictionary revealed three types of subdivided entries or, rather, three kinds of structuring problems: morphological problems, phonological (or morphophonemic) problems, and semantic problems. We would like to comment on three significant examples of these problems, without going into factual details, but in order to illustrate the very practical, concrete difficulties the lexicographer may experience in his daily practice.

Let us consider the entry */ˈpɔnt-e/ m.n. 'bridge' of the DÉRom (see Andronache 2008–2013 *in* DÉRom *s.v.* */ˈpɔnt-e/).

The Romance cognates fall into two distinct categories: firstly, masculine forms (Sardinian *pónte*, It. *ponte*, Fr. *pont* etc.). Secondly, feminine forms (Rom. *ponte*, Sp. *puente* etc.). If our dictionary classified the Romance materials in a purely morphological way, in a synchronic view, we would have just two subdivisions: (1) masculine cognates, (2) feminine cognates. But the DÉRom does not simply classify the data: it intends to *explain* them in an etymological, that is a genetic perspective. We do not forget that the purpose of such a dictionary is to reflect the internal dynamics of the protolanguage, i.e. the *succession* of synchronic stages.

The entry */ˈpɔnt-e/ poses the following problem: the situation of the Sardinian masculine cognate (*pónte*) may not be compared to the situation of Dalmatian, Istrian, Italian, French, Catalan, and Occitan cognates, which are nevertheless masculine as well.

In fact, the geographic distribution of cognates plays a crucial role in the diachronic interpretation. If we project our data on a map, we can see immediately that there are not two, but three separate areas. The feminine type corresponds to lateral, peripheral, isolated areas: Balkans, Iberia, plus some alpine regions. There is little chance that, in these lateral areas, the morphological change from masculine to feminine could be a common and late innovation (these areas being separated by hundreds or thousands of kilometers).

Therefore, our assumption , following de Dardel 1976, is that the feminine is not an innovative feature, but a residual one. It is some vestige of an older state of the protolanguage. The masculine gender instead, which covers a central, compact region of Romania (Italy and Gallia, mainly), a region whose center of gravity seems to be Rome, appears as an innovation which spread from Italy, but could not reach Iberia, nor the Balkans.

But if we follow this interpretation (a residual and marginal feminine, an innovative, central masculine), the position of Sardinian *pónte* can no longer be explained. We know with certainty that Sardinian is the first language to have split from (continental) Proto-Romance (Klinkenberg 1999: 135-136), and that this linguistic separation of Sardinia dates back to the second century at the latest. It is highly improbable (in a geohistorical view) that the Sardinian masculine *pónte* represents the same historical stage as Italian *ponte* and French *pont*.

In fact, we must not only reconstruct a masculine and a feminine noun in Proto-Romance, but three different types: masculine$_1$, feminine and masculine$_2$. Therefore, our entry */'pɔnte/ should contain three historically relevant subdivisions, which are: original masculine noun (I.), feminine noun (II.), and innovative masculine noun (III.). While the feminization in II. can easily be explained as part of a general tendency for nouns of the third declension like */'lakt-e/, */'mar-e/, or */'mel-e/ to join the class of feminine nouns (see de Dardel 1976), the masculinization in III. analyses itself as the result of a learnèd influence of the acrolect, *pons* being masculine in classical Latin.

The DÉRom's analysis was contradicted by Möhren (2012: 6-9). We would like to take the opportunity to answer briefly the interesting arguments and the new data proposed there. Möhren's contribution is twofold: 1° he adds ancient and medieval Latin attestations; 2° he invokes Peter Stotz's authority.

Möhren adds to DÉRom's material some Latin attestations of the feminine, mainly from the medieval period. According to him, they contradict the gender dominating in their areas. Actually, the contradiction is very weak, since all the medieval attestations he adds originate from non-Romance speaking regions (Germany and Ireland): they are perfectly coherent with the description given by the DÉRom, which expects the presence of the feminine in marginal and residual areas. If the German and Irish attestations of the feminine noun are in some way a testimony of the Romance situation of a remote past, they fit suitably well in DÉRom's scheme.

Möhren adds one antique testimony, in Gallia (near Le Puy). Since this attestation is very old (probably from the end of the third century), it does not contradict the DÉRom's reconstruction: our assumption was that the wave of remasculinization was too late to reach Dacia before the separation of this space from the *Romania continua*, which is supposed to have taken place in the third century. It is thus perfectly plausible (and, in fact, to be expected) that in the same third century the feminine would not have been eliminated yet by this wave, in Gallia or elsewhere.

Any argument based on Stotz's opinion is obviously strong, and we do not intend to contest it, as we agree perfectly with the general consideration Möhren extracts from it: that medieval Latin is not necessarily a reflex of the vernacular languages, that it can document continuing traits of Latin (Möhren 2012: 8). This is precisely why the DÉRom does not put into its data any Latin written testimony, and why, anyhow, the Latin attestations discussed here cannot have a preponderant weight in the etymology of French *pont* and its cognates. The members of the DÉRom project are obviously not the first, nor the last, to think that the Latin attestations (in the Middle Ages as well as in late Antiquity) do not represent truly the Romance situation to be.

As we are discussing Möhren's arguments, we take the opportunity to insist on the fact that Proto-Romance features can be, in some (probably rather rare) cases, the product of educated people, as it was probably the case with the remasculinization of */'pɔnte/. This only means, in historical terms, that the school succeeded in imposing the masculine gender on the (main part of) the Romania. It is hardly a surprise that such a standardised language as Latin has known successful learnèd influences.

4. Phonological data structuring

But the case of */ˈpɔnte/ is still very simple compared to other cases with which we had to deal. With the example of */ˈakuɪl-a/ f.n. 'eagle' (see Greub to appear *in* DÉRom *s.v.* */ˈakuɪl-a/), we would like to examine now a slightly more complex case, where data structuring is based on phonetics. This is the general structure of DÉRom's */ˈakuɪl-a/ entry:

I. */ˈakuɪl-a/

II. */ˈa**i**kuɪl-a/

III. */ˈakul-a/ ~ */ˈa**u**kul-a/

IV. */ˈakul**i**-a/

V. */ˈa**ug**uɪl-a/

The unity of each subdivision is assured by, firstly, its phonetic (and of course semantic) starting point, and, secondly, its geographical cohesion.

It is impossible to reconstruct directly one etymon, the common Proto-Romance lexeme */ˈakuɪl-a/, from the different Romance cognates. The only way for going about it is to reconstruct various direct etyma, which constitute, in lexicographical terms, the titles of each subdivision of the entry. But from this point, it is possible to understand what links the forms of all these direct etyma together: it is the many ways of interpreting the problematic group */kui/, which caused difficulties that were then given different regional solutions.

In this particular case, it is not possible to understand this complex reorganization directly from the Romance lexemes, and there is no way either of classifying every single form under the same etymon, distinguishing between borrowings from other dialects or from Latin and competing types. The five etyma represent forms which really existed, and distinguishing and connecting them is probably the only way of understanding their underlining historical dynamics.

5. Morpho-semantic data structuring

The case of */'ɸamen/ 'hunger; famine; desire', which was compiled in the context of the European Master in Lexicography (EMLex; see Buchi/González Martín/Mertens/Schlienger 2012 *in* DÉRom *s.v.* */'ɸamen/), is far more complex. By comparing Romance lexemes meaning 'hunger', 'famine', and/or 'desire', one has to reconstruct, strictly speaking, not one Proto-Romance etymon, but at least five different bases (see Buchi/González Martín/Mertens/Schlienger submitted):

– the prototype */'ɸam-e/, a feminine noun, which is reflected in the majority of the Romance idioms: Romanian *foame*, Italian *fame*, French *faim*, Catalan *fam*, Portuguese *fome* etc.;

– the type */'ɸamen/, apparently a masculine noun, reflected by the Sardinian masculine noun *famen*;

– the type */ɸa'min-a/, a feminine noun, reconstructible from French (*famine*), Francoprovençal and some Northern Italian dialects;

– the type */'ɸamin-e/, also feminine, reflected by Occitan, Catalan, and Gascon;

– finally, another feminine type, */'ɸamit-e/, represented by Romanian *foamete*.

How should the DÉRom organize and present these data? The structure of the entry (and the very choice of a single entry) is based on the hypothesis that these five types ultimately go back to one single etymon. But which one? And in what order should the material be presented?

This example clearly shows that the process of data structuring is not only a lexicographical task, but represents the central part of the lexicological analysis: structuring means writing the history of words. Lexicographic data cannot be represented without prior etymological analysis, but at the same time, the etymological analysis presupposes that the material has already been classified. What looks like a vicious circle is actually a dialectical movement between data description and data analysis –which is the natural movement of etymological work. The question here is how to build gradually a plausible scenario that reflects the diversity of cognates, starting from their original unity. In the case of */'ɸamen/, the different steps of reasoning are the following:

(1) */'ɸam-e/ and */'ɸamen/ very probably result from a regular phonetic phenomenon. All Romance languages except Sardinian have lost the final nasal consonant */-n/. As we admit that Sardinian early split from the core of Proto-Romance, this explanation is very satisfactory.

(2) Between */'ɸamen/ and */'ɸamin-e/, there is no phonetic, but a morphological relationship. */'ɸamin-e/ apparently results from some remorphologization, i.e. an alignment of the quite rare, residual type */'ɸamen/, with its final consonant, on the imparisyllabic */'ɔmo/ 'man' (nominative singular) ~ */'ɔmin-e/ (accusative singular) type declension, which displays regular stem alternation. This alignment with a type presenting a regular thematic vowel engaged probably through the plural form –*/'ɸamin-a/ if it was a neuter noun, */'ɸamin-es/ if it was a masculine or feminine one– of the primitive inflectional type */'ɸamen/.

(3) Therefore, the type */ɸa'min-a/ can receive a fairly plausible explanation. Rather than assuming a derivative with the suffix */-'ina/ (whose function and value would remain unexplained), we may suppose a remorphologization from an earlier */'ɸamin-a/, with simple accent shift. Moreover, other lexical types support this assumption: French *vermine*, versus Italian *verme*, Gascon *bermi* etc., also go back to an alternating type */'βɛrm-e/ ~ */'βɛrmen/ ~ */'βɛrmin-e/.

(4) This explanation, in turn, allows us to deduce the original gender of the Proto-Romance lexeme. If */'ɸame/ is a feminine type, if */'ɸamen/ is a masculine one,

and if a plural form */'ɸamin-a/ did exist (with a characterized neutral plural ending */-a/), it is logically required to reconstruct an original neuter lexeme.

A third remorphologization, restricted to Romanian and therefore assigned to a later stratum, produced the type */'ɸamit-e/, formed by analogy with the regular inflectional type */'lim-e/ ~ */'limit-e/.

We see that this way of reasoning looks like a jigsaw puzzle, where each element enlightens the others and reinforces the coherence of the whole.

Data structuring in a DÉRom entry must logically reflect this process of reasoning: therefore, it is a hierarchical, tree-branching structuring. We do not just classify types, we integrate them as a part of a comprehensive explanatory scheme. In the entry */'ɸamen/, section I. (the original etymon) contrasts with sections II., III., IV. and V., which correspond to remorphologizations. But, as each remorphologization is historically independent from the others, it takes place on the same level of branching (the authors of the entry did not create a section called *Remorphologizations*). Within each of these first level subdivisions, we provided subsections corresponding to the semantic developments: II.1. 'hunger', II.2. 'famine', II.3. 'desire'. The status of the first division level and that of the second one are different, because the divisions I., II., III. etc. have a chronological relevance, while the divisions II.1., II.2. etc. belong to the same synchrony and do not reflect any chronological sequence. The meaning 'famine', for instance, is not a secondary semantic development –at least not on a level Romance linguistics could reach–, but one of the original meanings of the protolexeme.

6. Conclusion

With these examples of data structuring in the DÉRom, we have tried to emphasize the fundamental interaction between linguistic comparison and data structuring in a historical-reconstructive dictionary. According to our point of view, compiling an etymological dictionary of an entire language family is not, and cannot be, just cataloguing lexical units. When producing such a dictionary, the lexicographer necessarily produces a history and a dialectology of the protolanguage. We hope to have shown, through this short paper, that the new approach developed by the DÉRom project in Romance etymology is able to renew usefully lexicological theory as well as lexicographical practice.

7. Bibliography

Buchi, Éva 2010: „Where Caesar's Latin does not belong: a comparative grammar based approach to Romance etymology". Brewer, Charlotte (ed.): Selected Proceedings of the Fifth International Conference on Historical Lexicography and Lexicology held at St Anne's College, Oxford, 16-18 June 2010. Oxford: Oxford University Research Archive: http://ora.ox.ac.uk/objects/uuid%3A237856e6-a327-448b-898c-cb1860766e59.

Buchi, Éva/González Martín, Carmen/Mertens, Bianca/Schlienger, Claire (submitted): „L'étymologie de FAIM et de FAMINE revue dans le cadre du DÉRom". *Le français moderne*.

Buchi, Éva/Schweickard, Wolfgang 2008: „Le *Dictionnaire Étymologique Roman* (DÉRom): en guise de faire-part de naissance". Lexicographica. International Annual for Lexicography 24, 351-357.

Dardel, Robert de 1976: „Une analyse spatio-temporelle du roman commun reconstruit (à propos du genre)". Varvaro, Alberto (ed.): XIV Congresso internazionale di linguistica et filologia romanza, Napoli 15-20 aprile 1974. Naples/Amsterdam: Macchiaroli/Benjamins, 14/2, 75-82.

DÉRom = Buchi, Éva/Schweickard, Wolfgang (eds.) 2008–: Dictionnaire Étymologique Roman (DÉRom). Nancy: ATILF: http://www.atilf.fr/DERom.

Möhren, Frankwalt 2012: „Édition, lexicologie et l'esprit scientifique". Trotter, David (ed.): Present and future research in Anglo-Norman: Proceedings of the Aberystwyth Colloquium, 21-22 July 2011/La recherche actuelle et future sur l'anglo-normand: Actes du Colloque d'Aberystwyth, 21-22 juillet 2011. Aberystwyth: The Anglo-Norman Hub, 1-13.

FEW = Wartburg, Walther von et al. (eds.) 1922–2002: Französisches Etymologisches Wörterbuch. Eine darstellung des galloromanischen sprachschatzes. 25 vols. Bonn/Heidelberg/Leipzig-Berlin/Basel: Klopp/Winter/Teubner/Zbinden.

Klinkenberg, Jean-Marie, 1999^2 [1994^1]: Des langues romanes. Brussels: De Boeck/Duculot.

LEI = Pfister, Max/Schweickard, Wolfgang (eds.) 1979–: Lessico Etimologico Italiano. Wiesbaden: Reichert.