Sorin Paliga (Bucharest)
sorin.paliga@lls.unibuc.ro

Compiling dictionaries of defunct (?)
languages: Thracian elements in
Romanian

What thou loveth well remains, the rest is dross
Ezra Pound

Preliminaries

The study of Romanian, as the easternmost expression of the Romance world, still has an unresolved issue: the lack of a complete and reliable etymological dictionary. There is, in fact, a paradox. Romanian was approached from the etymological perspective at an early phase by Cihac (1870–1879, two volumes), especially if we consider the fact that similar etymological dictionaries did not exist at that time for a series of languages of the area. One should note, as an exception, the etymological dictionary of Albanian (Meyer 1891) and that both Romanian and Albanian have remained ‘etymology-resistant’ languages, in the sense that many words have remained unexplained, badly explained or
have been given several explanations, details, which tells us a lot about the specific situation of these languages. Immediately after its publication, Cihac’s work was vehemently criticized (to just use an elegant phrasing) by Hasdeu, an outstanding linguist and philologist, conversant with many languages, the initiator of the Indo-European studies in Romania, also the first editor of Slavic texts and, implicitly, the first Slavist as well. Cihac put into circulation some clichés, some of them still very persistent, despite the poor scientific quality of his work in general. On the other hand, Hasdeu’s reference studies are available in newer editions, e.g. Hasdeu (1988). Cihac’s method was simplistic and, of course, easily criticizable: any word of non-Latin origin was considered of Slavic, Turkic or Hungarian origin.

It is not the purpose of this paper to deny or underestimate the Slavic, Hungarian or Turkic, including Turkish (Ottoman), influence on Romanian. I would like to point out some facts which prove that the situation is, in reality, more complex than suggested by Cihac some 150 years ago and sometimes persistent until now. What Cihac really lacked, and unfortunately the situation has been so frequent in the studies dedicated to the Romanian etymology ever since, was a deeper insight into the substratum of Romanian, a field of investigation in which Cihac’s first critic, Hasdeu, initiated a brilliant series of studies, some outdated, some still valid from the etymological perspective. Hasdeu was the first to propose some interesting etymologies. A good example is offered by his analysis of the Romanian form *doină*, dialectal form *daină*, compared to the Baltic forms *daina* (Lithuanian and Latvian), all analyzed as archaic musical terms from a common prehistoric heritage. In Romanian, Lithuanian and Latvian the terms refer to specific folk songs.

The subsequent studies dedicated to these forms have generally agreed on the basic idea that they must be archaic, prehistoric terms. Various authors may have different views regarding their ultimate etymon, proposing Celtic, Iranian or even Basque etymologies, but Hasdeu’s hypothesis has been adopted by all those who have analyzed the Romanian and Baltic forms. Disregarding their ultimate etymon, Romanian *doină, daină* and Baltic *daina* must be indeed related and analyzed together. The author of this paper did not miss the opportunity to express his view too: a word of Pre-Indo-European origin, perpetuating a specific terminology of the languages spoken in the Neolithic, which would be fully in accordance with what we now know from investigating the Neolithic origin of European civilization and culture (Paliga 2007, which gathers many previous studies; *doină* is analyzed on pp. 95–112). Romanian *doină, daină* and Baltic *daina* are, without much doubt, survivals of prehistoric musical terms.
Enlarging the horizon

After a series of studies dedicated to other aspects of the Romanian etymology, some of them dating from the mid 1980’s, dealing with the Romance, Slavic and substratum heritage, an interim conclusion was reached: a global, ample etymological dictionary is needed. Such an ambitious project would require a more determined, perhaps also a collective, effort, even if there are outstanding etymological dictionaries written by one author only, e.g. Machek (1971) for Czech; Skok (1971–1974) for Serbo-Croatian; also the author of an unparalleled study dedicated to the place-names of the Adriatic coast (Skok 1950); Vasmer (1953–1958) for Russian; Brückner (1970) for Polish, etc. Given these circumstances, it was clear that Romanian etymological lexicography needed, and still needs, a more careful approach. The successive attempts towards an ample, global and complete etymological dictionary of Romanian have been a repeated series of failures: Cihac’s attempt, so fiercely criticized by Hasdeu, was below an acceptable level, even for the second half of the 19th century. A second attempt, technically of a high level and characterized by a good approach, the dictionary of Candrea and Densusianu (1914), was initiated but unfortunately not completed; only letters A through L were written, and only with reference to the Latin heritage. Ciorănescu’s (2002) dictionary may be elegantly labelled as just another attempt, not to use more radical terms. Other attempts, in the general dictionaries, were catastrophic, with obvious errors and incomplete analysis (e.g. the much quoted explicative dictionary known as DEX\(^1\)). In 2010, the Institute of Linguistics of the Romanian Academy initiated the publication of an etymological dictionary, only one volume of which has been published so far, that is letters A and B. It is technically well done, using simple terms, but it appears to repeat the older errors, a fact that will probably lead to its not being completed by the present team of authors. Even if it were to be completed, repeating the constant errors of the last 150 years may not recommend it as a good solution. Since it is an uncompleted work, it will not be analyzed here.

In these circumstances, after publishing some limited studies in the field of Romanian etymology, I attempted to take a more coherent step: to compile an etymological dictionary of the indigenous substratum elements of Romanian, i.e. of those words for which the substratum origin is certain, probable or, at least, possible. I was aware of the difficulty, but am also aware of the fact that only such an attempt may push things forward, present data

\(^1\) For years, also available online: dexonline.ro, together with other dictionaries of Romanian.
in a coherent way, make things interconnect, and thus make forms explained. Only by amassing a large amount of data can the standard principles of analysis so well explained by Skok (1950), amongst which are the principle of repeatability and of identical phonetic evolution, be effectively used to validate the etymologies proposed. Additionally, common vocabulary should be put together with place-names, as many forms are mutually relevant. Place-names are, in fact, heirs of the same linguistic strata present in an etymological analysis, therefore they illuminate some vocabulary items and, vice versa, vocabulary items are relevant to an analysis of place-names. I dare say only a good analysis of the Romanian and southeast European substratum (represented mainly by Thracian and also by Illyrian along the Adriatic coast) may lead to a good etymological dictionary not only of Romanian and Albanian, but of all the languages grouped in the so-called *Balkansprachbund* or *Union linguistique balkanique* – the term is not entirely correct, but is mentioned here for the sake of convenience.

This is not a polemic attitude, as it may possibly seem. It really describes a certain bitter reality, the result of a complex cultural situation in the Romanian intellectual world, preoccupied with the “noble” Romance origin of the Romanian language, and less with the “barbarian” influence, a rubric under which both Slavic and indigenous (Thracian or substratum) influence were consigned *de plano* or *de facto*. It is a clear fact that, unlike in the neighbouring countries, Romanian linguists have largely abandoned the issue of the indigenous heritage, and have placed it in a peripheral position, using derogatory terms like “thracomania” or the newer “dacomania” to refer to such interests. The internet has indeed led to a flourishing pseudo-scientific cult of thracomania, which seems entirely logical (to me at least) as long as the matter continues to be neglected in “academic” circles. Ignorance has always been comfortable.

A final note: beside the scattered studies referring to the Thracian (or substratum or indigenous) heritage of Romanian, beginning, as mentioned above, with Hasdeu, there are in fact only two substantial lists or glossaries or mini-dictionaries of such etyma: Poghirc (1969) – in fact the last publication under the aegis of the Romanian Academy, note the year of its publication and the only attempt to include place- and river-names) and Russu (1981), in which the central part of the volume refers to a list of the indigenous elements. Given the size limitations of the present paper, I cannot extend their analyses. Certainly, each of these two authors (and other authors as well) have produced case studies and other limited studies on the topic but, by and large, such attempts are very rare and in most cases are now quite outdated.
Step I: The substratum influence in South Slavic

Unlike Romanian, the Bulgarian and former Yugoslav linguists were not at all embarrassed or otherwise limited by preconceived ideas of the value of substratum investigations, and did study substratum influences with outstanding results. Out of the numerous attempts, I would first of all cite Skok’s *Slavenstvo i romanstvo na jadranskim otocima* (1950) [Slavic and Romance elements in the Adriatic islands], a book seemingly forgotten meanwhile. Skok identified and defined the linguistic strata in southeast Europe, and also clarified some principles of analysis, of which *repeatability* and *phonetic evolution* are the most important. Also, he gives the contours of the FOUR major linguistic strata of the Adriatic area: 1. Pre-Indo-European (or “Mediterranean”); 2. Indo-European, specifically Illyrian in the area in question, and corresponding to the Thracian stratum in more eastern and north-east areas; 3. Latin (postclassical Latin); 4. Romance, which is divided into Venetian and East-Romance. To the list one might add some notable studies published in Bulgaria, where *Thracology* has become a recognized as a field of academic investigation in many universities.

Starting from Skok’s study, and some other studies, my first step was to analyze the substratum influence in South Slavic. The result was the doctoral thesis *Influențe romane și preromane în limbile slave de sud* [Slavic and Pre-Slavic influences on South Slavic], a first edition of which appeared in 1997 (the book having been finalized for publication in March 1996) and a revised edition in 2013. The changes were, in fact, few in number, just some correction of errors, and some pages dedicated to the long-debated form *boier*, of a peculiar importance in east- and south-east Europe. The red thread of the discussion there is that the substratum (Thracian and Illyrian) influence is crucial in understanding the emergence of south Slavic languages and cultures, a fact that is apparent in both common vocabulary and place-names. There are many clear, obvious examples and, of course, some other less clear examples, which is to be clarified later.

Based on the material gathered for this area, mainly during my stay in Ljubljana in 1993, I extended the analysis to the archaic, substratum (i.e. Pre-Slavic) place-names in the Czech and Slovak area, some of which have obvious correspondence in the south Slavic area. The results of this study were added to the book’s etymological lexicon appendix.
Step 2: The substratum influence in Romanian

With this rich material, to which some other data were added, in 2006 I published a tentative etymological lexicon of the indigenous (or Thracian) elements in Romanian (Paliga 2006, now available as a free PDF download), which I consider the clue to the very essence of Romanian etymology. The book is based on my previous studies, and includes other examples analyzed by some scholars, e.g. Poghirc (1969), Russu (1981, 1995), Hasdeu (1988) or, more recently, Bonfante (2001) whose basic studies regarding Romanian are now easily available. The dictionary thus includes:

- The “classical” 180–190 words present in Poghirc and Russu, almost all of which may be labelled as “certain”, i.e. of substratum (Thracian) origin, with very few exceptions (see below).
- The forms which were explained by others as of substratum origin, including those added by the author.
- Place-names thus analyzed in Poghirc (1969), with many additions, derived from my previous studies.
- Many other forms for which the label “substratum origin” is justified at least as a possibility, if not a probability. Obviously, the boundary between certainty, probability and possibility is relative in any scientific analysis or debate. The reader is asked to accept these expressions for convenience, as a reference point for further discussions, not as a final answer to complex problems.

I was forced by circumstances to include some forms which cannot be of substratum origin, but were considered as such by other authors. One, perhaps famous, example is the verb a vătăma ‘to affect, to hurt,’ insistently analyzed by Russu as indigenous. The word is also essential to Romanian legal terminology. Nevertheless, the “classical” explanation, from Latin victimare, in fact from a post-classical form *vittimare, is the only reasonable one. It is true, Latin sequence ct is usually reflected in Romanian as pt, as in lac(t), lactis > lapte ‘milk’ or pectus > piept ‘chest; bosom’, etc. Nevertheless, a language is a living organism, and many examples in Romanian prove their peculiar evolution, including isolated forms; post-classical *vittimare is a fairly reasonable etymon. Note, as a similar example, the peculiar position of Romanian semantic changes like castigare > a câștiga ‘to gain, to win’ v. ‘to punish’ in West Romance. This evolution is so peculiar that some may suggest that the word’s apparent Latin origin is just accidental formal similarity, and consider the word as indigenous; or bucă ‘buttock’ < bucca ‘mouth’, whose meaning
is preserved in West Romance and in Romanian also in the compound *a îmbuca* (*in + bucc* *a*) ‘to eat’, i.e. ‘to put food into mouth’. Many other examples could be quoted.

Also, a form like *boier*, of major importance as an administrative and social term in the Middle Ages, cannot be of Turkic origin, as assumed by almost all linguists, but from Latin, i.e. from *bos > bu* *ou*, hence *boier* ‘owner of cattle’ = ‘owner of land = rich man’; cf. *ovis > oaie* ‘sheep’ – *oier* ‘shepherd’. The initial meaning of *boier* must have been ‘owner of cattle’ and – together with the shepherds, the owners of sheep – formed the essential elements in a classical society based on cattle and sheep breeding, as clearly shown by the evolution from *pecu, pecus to pecunia*, i.e. from the meaning ‘herd, flock’ to ‘money’, i.e. wealth. The hypothesis that *boier* is due to Turkic-Altaic influence must be abandoned, as it is not supported by any argument. It shows the shallowness of etymological and historical analysis, rather than a serious approach to the topic. It is most probable that Romanian *boier* is etymologically related to the French personal name *Boyer*.

Therefore, eliminating from the “classical” list several elements that do not seem indigenous, but adding many others, the contour of the substratum heritage of Romanian proves significantly broader than usually accepted, perhaps around 450 basic elements, which means more than twice as many as suggested by the lists of Russu and Poghirc. Because there are, of course, many derivatives and many variants, an accurate estimation is, for the time being, impossible. Numbering the elements is, perhaps, less important than a good, really good, etymological analysis. To compare, the number of the original Latin elements of Romanian is approx. 2,200 including some doubts regarding certain forms. The number of the Slavic and other elements must be reconsidered, as many are, in fact, pseudo-Slavic, a topic I would like to approach more seriously on another occasion, as it has vast implications in understanding the making of southeast Europe in the first millennium and after. (See below just one case study, that is the numeral 100, and the addendum on the relation *zmeu* ‘chthonian male figure of the Romanian tales’ ~ *zmeură* ‘raspberry’).

The basic list of the Romanian words in the dictionary is followed by lists of place-names and other forms analyzed on other occasions, including the place-names in South Slavic, in Czech and Slovak, some mythological forms in Thracian, etc. All in all, there may be more than 3,000 forms which support each other in their origin, their meaning and their phonetic evolution.

There is no serious argument for considering such attempts as “thracomania”, which is a simplistic and comfortable cliché, but unsupported by a genuine, serious analysis.
Linguistics and archaeology

Having completed remarks on the etymological dictionary, two books should be briefly analyzed. One is co-authored with an archaeologist and thus represents an interdisciplinary approach to the emergence of the early Slavs (Paliga, Teodor 2009). The other is a broad introduction to the comparative study of the Slavic languages (Paliga 2012). There are solid arguments proving that late, unassimilated north Thracian groups survived until at least the 6th century AD and that these groups had clear contacts with what one might label “the early Slavs”, moreover, that these late Thracian groups had a certain influence on Proto-Slavic. These must have been the Carpians (Carpi, Korpiloi) of post-classical times. In these circumstances, it is obvious that the issue of the substratum heritage of Romanian is getting new contours, because some elements prove to be pseudo-Slavic. A correct etymological analysis may clarify these situations too.

One example of this sort, and a relevant case study, is Romanian numeral sută ‘one hundred’, which – in its turn – also clarifies the borrowed character of Slavic *sъto, which proves to have a completely different phonetic evolution if compared to the numerals ‘10’ and ‘1000’, desętь and ty-sęšta, ty-sоšта respectively, with the normal, expected nasal though, true, with an unexplained epenthetic š in the case of the numeral ‘1000’, which is less problematic than ţ in sъto. I have provided arguments against the idea that this case falls under ‘exception in the phonetic evolution due to frequency’ (see Mańczak 1971, where a general analysis of this issue is presented).

The simplistic approach to those historical periods must be abandoned. The situation was more complex than it seemed in the 19th century, and even later. In a larger horizon, if we admit that Albanian is NOT a neo-Illyrian idiom (with a substantial, preponderant Romance vocabulary) but rather a neo-Thracian idiom, then we might safely assert that Thracian is NOT a defunct language but, with a long and tortuous history, the continuant of a late form of Thracian, spoken by those unassimilated Thracian groups who, around 500 and later, were pushed south by the major movements of the first Slavic groups in their way to the Balkans. In this case, the very title of this brief note should be modified: the attempt is not to construct a dictionary of a defunct language, but of a still living language, revitalized and continuing the lost groups of late Antiquity, a language preserved in its scattered, but sometimes very important elements in Romanian, and also in the structure of Albanian. But things are still complicated, and considering Albanian a Neo-Thracian, not a Neo-Illyrian, idiom is already another subject, touching on very sensitive issues in the Balkan studies.
References

Paliga S. 2013. Influente române și preromâne în limbile slave de sud. [Ed. a 3-a, definitivă, cu indice]. București.
Appendix

*zmeu*, *zmei* s.m. An essential figure of Romanian folk beliefs and folk mythology, with various beneficial and malefic attributes, usually represented as a subterranean male in search of a human wife. The form has been held to be a Slavic borrowing, even if many details belie this; closely related to *zmeur*, *zmeură* (see), which is NOT the result of chance. Sl. *zmьjь* ‘snake, serpent, dragon’, from a radical *zm-*, in its turn reflecting the root for ‘earth’, is not easy to explain, as IE *ḥ(h)d(ḥ)em-* ‘earth’, zero grade *ḥ(h) (d)hm-* resulted in Old Slavic *zem-* ‘earth’, whereas the root for ‘snake, dragon’, *zmьjь*, raises major questions, as Machek correctly observed. • The archaic meaning must have been ‘being living on EARTH’, i.e. ‘human’ (in its etymological meaning, from humus), hence ‘man’, as in Lithuanian, where – from the same etymon – there is *žmuo* ‘man, human’ (related to Prus. *smoy* ‘man’). The similarity between Romanian, which preserves a Thracian (substratum) element, and Baltic (Lithuanian and Prussian) is expected. The indigenous character of *zmeu* is also supported by some attested Thracian forms: Zimi-, Ziemî-, Zemo- in compound forms like Zimi-kenthis, Ziem-ices, Zemokontes, Zym-drenos, Zym-zdrenos etc. As in other similar cases (see under *sută*), we surmise that Sl. *zmьjь* seems to be a borrowing from either a northern Thracian dialect, before expansion, or from Proto-Romanian, during the first phase of expansion (as definitely *sută* and *kьmotra* are). See also *zmeur(ă)*. The relation between *zmeu* and *zmeur(ă)* is essential in understanding the archaic origin of these forms. • See also *rezema*, *răzema* (re-/ră-zema).

*zméur* -i s.m. and *zméură*, -e s.f. The plant *Rubus idaeus* (*zmeură*); ‘raspberry; hindberry’. Der.: *zmеurиš* ‘raspberry/hindberry bush’. Indigenous, from IE *ḥ(h)d(ḥ)em-* ‘earth’, zero grade *ḥ(h) (d)hm-* > Thr. *zmeur-* > Rom. *zmeur-*. Similarly Lith. *žem-uoga* ‘strawberry’ (the plant *Fragaria*) (from *žemė* ‘earth’, cf. *zmeu* above) and Germ. *Erd-beere* ‘id.’, from *Erde* ‘earth’, lit. ‘earth-bery’. These examples show that both strawberries and raspberries were initially associated with ‘earth’; they also support the association *zmeu* – *zmeur(ă)* in Romanian as substratum elements. Probably most linguists have assumed that the relation *zmeu* – *zmeură* is the result of chance; the etymological analysis clearly shows that they are indeed related as Urverwandtschaft.

Sample Entries: *zmeu* and *zmeură* (Paliga 2006: 213)