The pre-Roman elements of the Sardinian lexicon

Swanenvleugel, C.

2 Methodology

The study of prehistoric substrate languages comes with several issues. By definition, these languages are often not directly attested or they did not leave any attested descendants. As a result, the existence of any non-attested substrate language, including its properties, can only be inferred from the traces it left through contact with the languages that survived it. Expectedly, this leaves ample room for speculation concerning phonological and grammatical features of the substrate, not to mention its origins and affiliation. The main problem is that no direct linguistic evidence exists to either falsify or verify any claims made. All evidence is transmitted through the filter of contact with at least one other language, rendering the comparative method, a historical linguist's most trusted tool, only partially applicable (cf. § 2.1.1.2). Moreover, the indirect attestation of a hypothetical substrate language makes claims regarding its lexicon, grammar, and genetic affiliation difficult to evaluate. However, suggestions on how to address these methodological difficulties have been made at least since the late nineteenth century. For in-depth overviews of the field of linguistic substrate research, cf. Craddock (1969: 18–47; with a focus on Romance) and Kroonen (2024a; with a focus on Indo European).

In the present chapter, the most prominent challenges in linguistic substrate research are discussed, along with the methodological approaches adopted to address these challenges. In § 2.1, approaches to the identification and collection of the relevant data are discussed. Next, § 2.2 treats the analysis of the gathered data in terms of phonological and morphological features. Finally, § 2.3 explores the potential of what the interpretation of the linguistic data reveals about language relationships, and which processes and events they imply.

2.1 Identifying substrate elements

The identification of substrate elements in attested languages, as opposed to elements inherited from an ancestor language, is notoriously challenging. Especially when the ancestor of the language under research is not attested, discerning inherited from non-inherited formations is not always straightforward. The reliability with which a non-inherited element in a language can be identified inevitably depends on the degree of understanding we have of the ancestor of that language. Even for a well-researched language family like Indo-European, this has proven difficult (cf. Kroonen 2024a).

2.1.1 Criteria for substrate origin

The concept of linguistic substrates has featured in historical linguistic research from its early stages onwards (e.g. Hervás 1979: 18 [1800]; Pott 1853: 451–452; Ascoli 1882; Kretschmer 1896: 121; see Craddock 1969: 18–47 for an overview of the field). The first steps toward formalizing the methodology for the identification of substrate words can already be seen in Meillet (1908) and Cuny (1910). On the basis of semantic domain, unusual phonotactics and morphology, and irregular sound correspondences, they identify a number of pre-Indo-European words in Latin and Greek. However, clear criteria to study words of substrate origin in a more structured and objective manner were formulated only in the nineties of the previous century. None of the criteria is decisive on its own, but some constitute stronger evidence in favor of substrate origin than others. Roughly in descending order of strength, the proposed criteria are:

- lack of an etymology
- irregular correspondence
- unusual phonotactics
- unusual morphology
- specific semantic categories
- limited geographical distribution

Each of these will be discussed below.

2.1.1.1 Lack of an etymology

A straightforward reason to suspect non-inheritance, and potentially a substrate origin, is the absence of a clear etymology. This criterion has been employed in various Indo-European branches (Meillet 1908: 161; Polomé 1986: 661; 1989: 54; 1990: 276–277). In the case of Sardinian, its ancestor Latin is well attested and much of the Sardinian lexicon is therefore etymologically transparent. If a Sardinian word does not have a clear Latin pre-form, this can suggest a non-inherited origin. Naturally, there is always a possibility that a word was not recorded by chance. Given our extensive knowledge of Latin however, these chances are slim. Conversely, if a plausible Latin etymology is available, it should always be preferred over a hypothetical substrate origin.

2.1.1.2 Irregular correspondence

There are cases in which a certain word is found in several related languages and could thus, based on its linguistic distribution, be inherited. Still, a non-inherited origin may be posited when the attested cognates deviate from the known sound laws and sound correspondences. In such cases, one would assume that the word

was borrowed from a third language, or that it spread through borrowing between the related languages.

This criterion derives from a principle central to the comparative method: the regularity of sound change. It is used already by Meillet (1908) and Cuny (1910: 158–160), and remains important in substrate methodology today (Polomé 1989: 55; 1990: 278; Hamp 1990: 296, 300; Huld 1990: 392–393; Schrijver 1997: 294; Lubotsky 2001: 303). The criterion can be used on different levels: either between two languages that are "vertically" related, such as Sardinian and Latin, or between two languages that share a common ancestor, like Latin and Greek. Irregular correspondence between dialectal varieties of a single language can indicate a non-inherited origin of words too: in this case it is often referred to as "irregular alternation" (e.g. Cuny 1910: 156). With sufficient material, the nature of the observed irregularities can potentially offer clues about the phonology of the substrate source language (cf. § 2.2). The Sardinian evidence is discussed in § 8.

A special case are seemingly corresponding forms in unrelated languages (e.g. Cuny 1910: 161–164), such as Sardinian and Basque (§ 10.2.2). Since these languages do not share a common ancestor, the detection of regular correspondence is not a possibility, nor is the detection of phonological irregularity. When vocabulary items in unrelated languages share a common origin, this is likely the result of lexical borrowing. If the cognates cannot be etymologized in either language, they must have been borrowed from a third source.

There is an important caveat. The comparative method can reveal when two comparanda do not correspond regularly to each other, but it can only prove a common etymological origin when the comparanda do correspond regularly to each other. This is a fundamental limitation that causes any comparison of two or more irregular comparanda to run a risk of chance correspondence. Without the possibility to systematically verify or falsify the validity of a comparison, we must be watchful not to invalidate our investigations of linguistic prehistory by lumping any two vaguely similar words together. The methodological strategies applied in this study to mitigate this risk are discussed in § 2.1.2.2. Another problem for detecting potential substrate words is the fact that, at times, inherited formations exhibit irregular developments too. The phenomena responsible for this, and how they should be taken into consideration, are discussed in § 2.1.2.1.

2.1.1.3 Unusual phonotactics

The criterion of unusual phonotactics is closely related to the previous criterion and applies to phonological sequences that are unusual in the relevant language

or its ancestor, whether reconstructed or not, (Polomé 1989: 55; Hamp 1990: 298; Huld 1990: 393; Salmons 1991: 267; Beekes 1996: 218; Schrijver 1997: 294; Lubotsky 2001: 303–304). This could for instance be a phoneme that does not have a regular source in the historical phonology of the language, such as e.g. intervocalic *s* in Lat. *rosa* 'rose' (Weiss 2009: 151 fn. 12; Wigman 2023: 121–123) and IE **a* and **b* in Lat. *faba* 'bean' < **b*^{h}*ab-* (Kuiper 1995: 65–67; Wigman 2023: 80). Phonemes that do have a regular source, but that occur in a phonological context that did not occur in the ancestor language, are equally suspicious. An example of this is Srd. *ɸ* before *i* (supposing Lat. **c̯i̯-* or **t̯i̯-*) (§ 8.1.1). Here too however, irregular processes can falsely suggest non-inheritance (§ 2.1.2.1). In Sardinian, for instance, *s-* was irregularly replaced by *ɸ-* in *ɸilímba* (Baunei) 'bean pod' < Lat. *siliqua* (DES II: 417). For the interpretation of unusual phonotactics, cf. § 2.2.1. The Sardinian evidence is presented in § 8.

2.1.1.4 Unusual morphology

Another indication of borrowing is when lexemes exhibit morphological features that cannot be explained by inheritance (Polomé 1989: 55; Huld 1990: 390; Salmons 1991: 267; Schrijver 1997: 294–295; Lubotsky 2001: 304). This concerns affixes without an etymology, such as Gr. -ι/υνθ- (Pott 1853: 451; Meillet 1908: 162; Cuny 1910: 154–156; Katičić 1976: 41–43). It may also include other types of morphology, like irregular ablaut (cf. Schrijver 1997: 307–312). In the context of this work, a concrete example are Sardinian nouns ending in *-i* that suggest a Latin accusative in *-im*, which is only plausible for a small, closed class of inherited nouns (§ 8.1.5.1). The interpretation of morphological features is discussed in § 2.2.2. The Sardinian evidence is discussed in § 9.

2.1.1.5 Semantic category

It has been observed that lexical material of hypothesized substrate origin often clusters in specific semantic fields. Among these, prominent categories are those referring to the natural world, such as flora, fauna, geography, as well as terminology referring to cultural practices and technologies specific to the area (Meillet 1908: 161; Cuny 1910: 154; Bertoldi 1931: 94; Polomé 1989: 54; 1990: 276; Hamp 1990: 300; Schrijver 1997: 295; Lubotsky 2001: 304–305). This does not mean that any word in such semantic categories must be of substrate origin. As long as a concept existed in the predecessor of a language, the corresponding word could have been inherited. As this criterion much depends on the socio-economic context of the language contact, it is weaker than the purely linguistic criteria discussed above. However, it can inform on the nature of language contact once the material has been identified as being non-inherited (e.g. Polomé 1986:

669)(cf. § 2.3.2). In this study, the lexical material discussed is organized according to semantic category. The categories are as follows: flora (§ 3), fauna (§ 4), geography (§ 5), culture and economy (§ 6), and miscellaneous (§ 7).

2.1.1.6 Limited geographical distribution

A limited geographical distribution is the last criterion introduced here. It has been invoked with regard to the branches of Indo-European. When a word occurs only in Indo-European languages of northern or western Europe, for instance, projecting that word back to Proto-Indo-European is considered dubious (Polomé 1986: 663–665; Huld 1990: 393; Schrijver 1997: 294): even when the comparanda are regular, there is a risk of borrowing (Schrader 1883: 201–202). Within the context of the present study, a word occurring exclusively in Sardinia could quite possibly be a loan from an extinct language once spoken on the island. If, however, a word is found in all of Romance, chances are that we are dealing with a formation that was inherited from (vulgar) Latin. In isolation, this criterion is not conclusive. Inherited material may have been preserved only in a single branch of a language family due to chance. For Sardinian, examples of this are Srd. *dòmo*, *dòmu* 'house' < Lat. *domō* (abl.) 'id.', and Srd. *yánna*, *yènna* etc. 'door' < Lat. *ianua* 'id.' (REW 4575; DES I: 707), which have been replaced in most other Romance varieties by reflexes of Lat. *casa* 'house, hut' and Lat. *porta* 'gate' respectively (REW 1728, 6671). Nevertheless, the geographical distribution of putative substrate words is important for interpreting prehistoric linguistic connections between regions (Polomé 1986: 669; 1990: 282) (§ 11).

2.1.2 Form, meaning, and chance similarities

In isolation, none of the criteria discussed in § 2.1.1 can conclusively demonstrate a word's substrate origin. In practice, it is often the cumulative evidence that decides the issue. The case for a substrate origin becomes stronger if a word answers to two or more of the criteria in § 2.1.1 (Schrijver 1997: 294–296). And even then, rejecting an inherited origin may be impossible. For instance, Srd. *iṡbírru*, *skírru* 'marten' < **squiriolu* (DES I: 646–647) satisfies multiple criteria suggesting a non-inherited origin. It belongs to the semantic category of animal names, is found in the meaning 'marten' only in Sardinia, and corresponds irregularly to Lat. *sciurus* 'squirrel'. Yet, it was likely inherited from Lat. *sciurus* 'squirrel', with a shift in meaning and an irregular metathesis from **skʲur-* to **skʷir-* (DES I: 646-647).

On the other hand, there are instances where accepting an inherited origin is problematic, even if the word question does not satisfy any of the criteria dis-

cussed. Such is the case for Srd. *kála* 'cove, inlet (for ships)', which corresponds to Sp., Cat., OProv., Cors., It., Sic. *cala* 'id.'. It is thus attested in a wider region, in a semantic category different from the ones mentioned, and shows no trace of phonological irregularity. However, it lacks a reliable Latin etymology and has therefore been argued to be of pre-Roman origin (cf. § 5.1.3).

Since no criterion by itself is decisive, the identification of lexical substrate elements depends on careful consideration of form, meaning, etymological status, and geographic distribution. A Romance language like Sardinian has the advantage that its Latin ancestor is exceedingly well attested, allowing us to follow its diachronic development for over two millennia. Because of this, any Romance word lacking a Latin correspondent, is suspect of borrowing from the outset, even if it is otherwise unremarkable in form, meaning, or distribution. Although by no means a conclusive criterion either, it is relatively strong compared to cases in which the linguistic ancestor is not attested. This is the case for many ancient Indo-European subgroups, such as Latin, Greek, Sanskrit, and Hittite, but also for ones that were attested relatively late, such as Baltic and Slavic. Still, even though the study of Sardinian linguistic prehistory is greatly aided by our knowledge of Latin, there are complicating factors that deserve attention. Specifically, inherited material too sometimes exhibits irregularities (§ 2.1.2.1). This raises a methodological question as to when two irregularly corresponding forms should be seen as independently borrowed cognates, or as the result of local irregularizing processes (§ 2.1.2.2).

2.1.2.1 Substrate origin or irregular inheritance?

The exceptionlessness of sound change is the cornerstone of the comparative method (cf. § 2.1.1.2). However, this exceptionlessness is frequently obscured by analogy, folk etymology, onomatopoetic processes, sound-symbolism, and basically any type of "expressivity". This obstacle is all the more evident on the dialectal level. Local varieties of one and the same language can exhibit slightly irregular manifestations of an etymologically identical formation, resulting from local processes or inter-dialectal borrowing (Schuchardt 1885: 11). These irregularizing processes are crucial to take into account, as they may falsely suggest a non-inherited origin for inherited lexical items.

Within the context of this study, it must be noted that many allegedly pre-Roman words in Sardinian have received competing Latin etymologies or etymologies rooted in the historical contact languages of Sardinian. Not infrequently, the choice between a substrate origin and an inherited or recent loan origin

depends on the interpretation of irregular phonological developments or opaque semantic shifts. To an extent, the acceptance of either an inherited or a substrate origin for these cases is thus a matter of intuition. Some scholars are willing to accept a certain degree of formal or semantic deviation to sustain an etymology. Others prefer to take formal and semantic incongruencies at face value, leading them to posit an unattested substrate. Since both approaches entail uncertainties, systematically prioritizing one over the other is counter-productive: we must instead face the reality that both irregular developments of inherited words and borrowings from unattested languages do occur. In order to a achieve a maximally robust methodology, we must acknowledge these uncertainties and carefully weigh the arguments in favor an irregular inherited origin against those in favor of a substrate origin, for each individual word. This is the approach that I adopt in the discussion of the lexical material in § 3 – 7.

2.1.2.1.1 A note on folk etymology

Folk etymology is the process by which the form of a word is altered on the basis of an association with the form of another word. This process is often aided by semantic association. Various linguists have stressed the role of folk etymology in concepts related to animals, plants, and belief systems (Alinei 1984; Beccaria 1995). Paulis (1992), in his work on Sardinian plant names, often invokes folk etymology to account for irregular interdialectal comparanda found across different dialects. The plausibility of a folk-etymological origin depends on the similarity between the irregular form and the putative model, and on the possibility of a semantic connection between the two forms. Needless to say, such evaluation demands a good understanding of the cultural and practical applications and associations of the concept in question (e.g. Paulis 1992: 11–12). As a general rule, I consider the semantic motivation of a supposed folk etymology convincing as long as it does not depend on hypothetical unverifiable cultural contexts; cf. § 2.1.2.2, where a similar attitude towards non-inherited comparanda is adopted.

2.1.2.2 Comparing non-inherited words

Traditionally, a large part of the research into substrate languages has centered on identifying lexical correspondences between the substrate lexicons of different languages or areas. The motivation for this is typically to recover the prehistoric language relationships and contacts between peoples. However, as mentioned in § 2.1.1.2, the comparison of irregularly corresponding cognates, or similar words in unrelated languages, always involves a risk of chance similarity. In order to render the corpus of potential substrate connections as reliable as pos-

sible, it is pertinent to be strict regarding their formal and semantic similarity. A clean corpus is important for a reliable analysis of the linguistic data, on which hypotheses of prehistoric language relationship are necessarily based.

Ideally, a set of non-inherited comparanda are formally and semantically (near-)identical. An example of this is the comparison between Srd. *golóstḭu*, *golóstri* etc. 'holly' and Bq. *gorosti* 'id.', both of which can go back to **golosti* (§ 3.3.7)(Bertoldi 1929: 261 fn. 3). The majority of comparisons, however, require varying degrees of liberty on the formal or semantic side, or both. Statistically, longer corresponding phonological sequences are less likely to be the result of chance similarity. For instance, the shared sequence **al-* in Srd. *aláše* 'butcher's broom' and Prov. *árẹ* 'broom' (§ 3.2.1) stands a higher chance of being due to coincidence than the much longer sequence of sound correspondences between aforementioned Srd. *golóstḭu* and Bq. *gorosti* 'holly'. It is however difficult to draw a line, and great coincidences can of course happen sometimes.

The semantic evaluation of lexical comparisons is even more intuitive (cf. § 2.1.2.1). The approach of this study is to accept only those comparisons whose semantic shifts are trivial. As indicated above, this includes only those shifts that do not require hypothetical cultural contexts (cf. Polomé 1990: 282). Thus, the shift 'maple' > 'ash tree' needed to connect Srd. *kòsti*, OProv. *agast* 'maple' to Tusc. *còstolo* 'ash tree' is accepted under this criterion (§ 3.3.9). The change 'poppy' > 'lady' needed for Lang. *ãndèr* etc. 'poppy' and Bq. *andere* 'lady' (§ 3.1.15), on the other hand, is not. Needless to say, the approach described here is but a crude way of drawing a line. Inevitably, it will exclude comparisons that are correct despite their opaque semantics. For instance, the Romance words for 'liver' (Nuor. *fíkatu*, Camp. *fɨɣáu*, Sp. *hígado*, Fr. *foie*, It. *fegato* etc.), which are known to be derived from Lat. *ficus* 'fig', would not be considered convincing for reasons of stringency.[26] However, it is preferable to infer prehistoric language relationships on a smaller corpus of strong comparisons than on a larger corpus of weak ones.

2.1.2.2.1 Toponyms as evidence

Placenames, or toponyms, take up a prominent place in substrate research in general (e.g. already Kretschmer 1896: 400). Especially in the earliest days of the "Mediterranean hypothesis", many of the proposed pre-Roman connections between different areas in the Mediterranean were based in large part on toponyms (§ 1.3.2). Toponyms are among the parts of a language that are most resilient even in the event of a wholesale language-shift (Bertoldi 1931: 95). In situa-

[26] Cf. FEW III: 490–493 and DES I: 518–520 for the specificities of this etymology.

tions where speakers shifted from a language prior to its attestation, the toponymical evidence potentially contains valuable traces of the preexisting linguistic landscape. However, even though toponyms can be preserved over long periods of time and across multiple language shifts, they suffer from a significant drawback: they are usually semantically opaque. As discussed in § 2.1.2.2, scrutiny of the semantics of comparisons is an important step to minimize the risk of chance correspondence. To quote Bertoldi (1933: 260): "abandoning the lexicon […] means letting go of one of our most precious research instruments: the possibility of semantic verification".[27] Since the original meaning of substrate toponyms cannot typically be retrieved, such comparisons are limited to formal similarity.[28]

Toponymic evidence can on the other hand be mobilized as an additional check for verifying the characteristics of a substrate, once it has been identified in the lexicon. Provided that the source language of substrate toponyms is the same as the source language of the substrate lexicon, we should expect features exhibited by the lexicon to reappear in local toponymy (Aikio 2004: 9). In addition, toponyms can support the identification of recurring sequences, i.e. potential morphemes of the source language. These too can then be compared to the lexical evidence. In Sardinia, such an approach has been applied to the toponymy of the Barbagia-region by Wolf (1998a) and integrated with the findings from lexical comparisons by Paulis (2008). Thus, toponyms can aid the reconstruction of the phonology and morphology of a hypothetical substrate source language (§ 8 – 9).

2.1.2.2.2 Conclusion: dealing with chance similarities

As discussed, the main methodological issue in comparing potential substrate words is that chance correspondence can never fully be ruled out. This does not mean that the matter is to be abandoned altogether. In order to isolate compelling evidence, etymological proposals should be evaluated along the following two lines:

---

[27] "Abbandonare il lessico significa però lasciarsi sfuggire uno degli strumenti di ricerca più preziosi: la possibilità del controllo semantico." (Translation mine).

[28] A quick glance at some similar European city names whose etymologies are known, reveals the danger of postulating cognacy without semantic checks: cf. *Lyon* (France) < Lat. *Lugdūnum* << Gaul. **Lugudūnon* 'stronghold of Lug' vs. *León* (Spain) < Lat. *legiō, -ōnem* 'legion'; *Almería* (Spain) << Arab. *al-Mariyya* vs. *Almere* (The Netherlands) < *Aelmere* 'big lake'. For additional examples, see Bottiglioni (1929: 15–16) and Wolf (2011: 607).

- The forms compared must formally be sufficiently similar or identical. This is difficult to quantify, but the longer the corresponding sequence, the better. In doing this, it is crucial that the internal phonological developments of the compared languages be taken into account.
- The meanings of the comparanda should be evident. They should as a minimum not involve any semantic shifts that are only understandable with additional, unverifiable assumptions (about *Benennungsmotive* or cultural practices etc.).

Positing these guidelines is easier than putting them into practice. There is no final answer as to how divergent the forms and meanings of two or more comparanda can be, before their comparison is to be rejected. Each comparison should be weighed carefully on the basis of the relevant data. The semantic guideline eliminates toponyms as reliable evidence.

2.2 Analyzing the data

In the above, we have explored the methodological strategies for identifying and comparing non-inherited material between different languages to detect prehistoric linguistic relationships between regions. It is equally informative to investigate the phonological and morphological features of the non-inherited lexicon. This enables us to exploit the body of non-inherited material found in Sardinian to gain insights into the properties of the substrate language(s) of Sardinia (cf. Devoto 1940: 40–43), as well as the dynamics of its contact with Latin.

2.2.1 Phonological analysis

The notion that phonological properties of suspected substrate words reflect phonological properties of the donor language has been present from an early age. As early as 1800, Don Lorenzo Hervás (1979: 18) notes:

"Had these things [i.e. the presence of other languages in France and Spain in Roman times] not been known to us through history or tradition, we could infer them by observation and comparison of the Latin dialects, which the French and the Spanish speak presently, with Celtic and Cantabrian [i.e. Basque], which are still spoken among them, with ancient Latin, and with the pure dialects of the latter, which in Italy are spoken by the Romans' descendants. This observation and comparison would make us notice and discover many features of Celtic in French, and many of Cantabrian in Spanish. If by chance the latter and the Celtic language had perished, and we would thus consequently not have been able to compare to them the languages that the French and Spanish actually speak, nevertheless, because of the differences of the idioms and of other things in

those languages, we would conjecture that in ancient times the French and Spanish spoke different languages."[29]

This view was taken up by Ascoli (1882), who argues that certain typical Gallo-Romance sound shifts resulted from a Celtic substrate.[30]

Later, the attention was directed to phonological irregularities and alternations found in specific lexical items. Wagner (1907a: 408–410) already considers ancient Iberian as a source for unexplained lexical similarities between Sardinia and the Iberian Peninsula (cf. § 11.3.1). Meillet (1908) sees a pre-Indo-European language impacting Greek, Latin, Armenian and Iranian. Both scholars express a degree of methodological hesitation. Wagner (1907a: 408) asks: "if the Sardinian word [i.e. *vega*] is old, as we must believe, the etymon **vica* [proposed for Sp. *vega*] is not acceptable; but is it therefore certain that it is Iberian, or do we need to look for another explanation?".[31] Meillet (1908: 164) warns that "it would be unwise to want to specify too much detail in such matters [i.e. which populations the people speaking Indo-European dialects encountered]"[32] — words of caution that are valid to this day. Nevertheless, the idea that irregularly corresponding words could represent parallel borrowings from a substrate language was well-established by the beginning of the 20th century.

What followed was the interpretation of the irregularities and alternations in these putative substrate words. Meillet (1908: 163) and Cuny (1910) account for various irregular correspondences between Greek and Latin words by postulat-

[29] "Si todas estas cosas no nos constaran por la historia ó tradicion, podriamos inferirlas de la observacion y cotejo de los dialectos latinos que al presente hablan los francesses y españoles, con el céltico y cántabro, que aun se hablan entre ellos, con el latin antiquo, y con los dialectos puros de éste, que en Italia hablan los descendientes de los romanos. Esta observacion y cotejo nos harian advertir y descubrir en la lengua francesa muchas cosas de la céltica, y en la española muchas de la cántabra. Si por ventura ésta y la lengua céltica hubieran perecido, y consiguientemente no pudiéramos hacer entónces cotejo con ellas y los lenguages que en la actualidad hablan los franceses y los españoles, no obstante, por razon de la diferencia de los idiotismos y de otras cosas en dichos lenguages, conjeturariamos que antiguamente los franceses y los españoles hablaban diversos idiomas." (Translation mine).

[30] Cf. Kretschmer (1896: 121) for some more early examples.

[31] "Se il vocabolo è anche del sardo antico, come bisogna credere, l'etimo **vica* non è ammissibile per questo; ma è per questo sicuro che è iberico o bisogna cercare un'altra spiegazione?" (Translation mine).

[32] "Il serait imprudent de vouloir trop préciser le détail en pareil matière". (Translation mine).

ing independent borrowings from a third language. They attribute the irregularities to a phonological mismatch between the respective source and target languages. Several alternations in the Greek lexicon have since been attributed to "imperfect" borrowing of non-native phonemes from a substrate language (e.g. Kretschmer 1940: 269; Heubeck 1949: 201–202; 1959: 56–57; Kuiper 1956: 219–220; Schachermeyr 1964: 260–261; Ruijgh 1967: 53 fn. 35 etc.). For examples outside Greek, cf. Martinet (1955: 387) and Kronasser (1966: 41).

The notion that formal irregularities in substrate words could be attributed to phonological properties of the source language, has become fundamental to linguistic substrate research. Attempts to find recurring patterns in irregular alternations are made by Oštir (e.g. 1921; 1930) and later by Furnée (1972; 1979). Both try to interpret irregular alternations between words in various languages as a function of the phonology of substrate languages. Neither scholar's efforts have been met with great enthusiasm. In Oštir's case, this is likely due to the density and idiosyncrasy of the data presentation and of his complicated writing style. Furnée (1972: 92–83) interprets the phonological alternations observed in Greek as reflecting independently borrowed "expressive" and/or "euphonic" variants in the source language, thus pushing the problem from Greek to the unattested source language. This approach is continued and improved by Hamp (1979; 1990), Huld (1990), Salmons (1991), Beekes (1996), Schrijver (1997) and Boutkan (1998). More so than Oštir and Furnée, these scholars identify well-defined irregular correspondences recurring in words found across the (northern) European branches of Indo-European. On the basis of these, they posit clearly formulated ideas on some of the morphophonological properties of a hypothetical pre-Indo-European language that could be the source of these words. Furnée's (1972) extensive material is used by Beekes (2014) to interpret "pre-Greek" alternations as the result of phonological mismatches between source and target language.

This is the methodology that will also be adopted in this study, specifically in § 8. Detailed analysis of the dialectal variation found in the Sardinian lexicon (in § 3 – 7) allows us to identify irregular correspondences that cannot plausibly be explained by means of processes such as folk etymology (cf. § 2.1.2.1). If any of these irregularities recur in multiple words, this can be indicative of features of the phonology of their pre-Roman Sardinian source language(s) (cf. Salmons 1991: 267; Schrijver 1997: 312).[33] Irregular correspondences with single occurrenc-

[33] The phonology of pre-Roman Sardinian toponyms has been studied by Serra (1960) and Wolf (1998a: 77–81), with a focus on syllable structure.

es only cannot be used to make any reliable inferences about the phonology of a substrate language and are therefore left out of consideration.

2.2.2 Morphological analysis

The methodology for identifying and interpreting morphological features of substrate words has gone hand in hand with that of phonological features (§ 2.2.1). Pott (1853: 451–452) considers the Greek elements -νθ- and -σσος/-σσα, also found in toponyms, to "maybe [be] a remainder of from the language of a people that preceded the Hellenes".[34, 35] For Sardinian, Forsyth Major (1893: 154) and Guarnerio (1904b: 259) analyze the element *činči-*, *tsintsi-*, found in various animal names, as a diminutive prefix of Iberian origin. Kretschmer's (1896: 402–405) discussion of Greek forms containing -νθ- allows us to distill three criteria for establishing a substrate origin for a segment:

1. The sequence is not an isolated phenomenon but occurs in various words of non-inherited origin.
2. The "morpheme-hood" of the sequence is secured through alternations in identifiable cognates.
3. The sequence cannot be of inherited origin because of formal or semantic reasons.

Kretschmer (1896: 404) additionally notes that a pre-Greek origin of -νθ- is suggested by the fact that it is not productive and only occurs in words of obscure origin. However, this criterion only works in one direction. While its occurrence on etymologically obscure words may indeed point to a substrate origin for -νθ-, a word containing -νθ- cannot mechanically be considered a substrate word, as secondary productivity of the suffix cannot easily be ruled out. This applies to all hypothetical substrate morphemes.[36] Over the years, numerous substrate morphemes have been posited on the basis of both toponyms and lexical nouns, especially in the context of the Mediterranean hypothesis (see § 1.3.2). The foundations for the identification of pre-Roman Sardinian morphemes were laid by Terracini (1927), who identifies various pre-Roman suffixes found in Sardinian toponyms (*-V́ɻ́*, *-án*, *-́in*, *-énnVr*, *-́Vr* etc.; cf. § 9.2).

A risk concerning the study of substrate morphology is to give a sequence morphemic status when there is in fact no compelling reason to do so. An example of

[34] "… vielleicht Ueberrest aus der Sprache eines den Hellenen vorausgegangenen Geschlechts, …" (Translation mine).

[35] Cf. Katičić (1976: 39–55) and Kroonen (2024b) for an overview.

[36] Cf. the complex case of athematic velar suffixes in Greek (Chantraine 1933: 376–383).

this is Bertoldi's (1928: 231) suggestion of a "Mediterranean" derivational suffix *-*st*(*r*)-, represented in Srd. *golóstḭu* etc. 'holly' (§ 3.3.7) and *giḍḍòstre* etc. 'tree heather' (§ 3.3.20)(cf. also Bertoldi 1929: 261 fn. 3; 1930). Although the first and third criteria listed above are fulfilled, with two non-inherited words containing an etymologically unexplained sequence, neither of these words have corresponding comparanda without *-*st*(*r*)-.[37] Even if such an alternation can be found elsewhere (e.g. Prov. *árę* 'broom' besides Lang. *orousto* 'broom branch'; § 3.2.1), the Sardinian data do not provide evidence that *-*st*(*r*)- was an independent morpheme, rather than a part of the lexical stem, in the Sardinian substrate. This example illustrates how one's evaluation of the second criterion depends on the cognates that one accepts, which in the case of non-inherited lexicon suffers from the additional uncertainties described in § 2.1.2.2. Thus, FEW's (II: 490) and Hubschmid's (1953: 30) comparison between Srd. *kaðúmbu*(*lu*) 'great mullein' < **katúmbu*- and Campan. *kwátenə*, *kwātanə* 'id.' is only possible if both Srd. *-úmbu* and Campan. *-enə* are analyzed as suffixes to a root **k*(*w*)*at*-. However, identifying these morphemes is only possible in the first place if we accept the cognacy between these two words.[38] This approach evidently runs the risk of becoming circular, and for that reason I exclusively consider potential substrate morphemes if their alternations are attested within Sardinian, in indubitably cognate forms (in § 9). This will inevitably exclude many of the comparisons proposed in previous scholarship (cf. § 1.3), but in turn greatly increases the reliability of the accepted cases.

Once a substrate morpheme has been identified, the logical next step is to determine its function. This is usually challenging, however. If we exclusively accept a sequence as a morpheme when it alternates in forms recognizable as cognates due to their (near-)identical meaning (e.g. *-(*i*)*k*- in *kòstike* besides *kòsti* 'maple'; § 3.3.9, § 9.2.2.2), we inevitably lose any derivational morphemes whose very purpose it was to alter the original meaning of a word. What remains are those morphemes that are semantically bland, including but not limited to case markers, gender/noun class markers, number markers etc. It does not come as a surprise that so many alleged substrate affixes have been interpreted as having a plural or "collective" function: cf. § 9.2.1, § 9.2.2.1, § 9.2.6.2.1 for the Sardinian cases alone. It is implausible that all of these proposals are correct.

---

[37] The comparison between Srd. *giḍḍòstre* and Bq. *gillar* etc. by Hubschmid (1953: 29) is uncertain (cf. § 3.3.20).

[38] For many more such comparisons to Srd. *kaðúmbu* 'mullein', cf. Hubschmid (1953: 29–33).

The two pre-Roman morphemes identified in the Sardinian lexicon that have received most attention with regard to the question of their original function, are **-ar/-Vr* and **ϑi-* (and variants). The former is found in lexical items as well as toponyms. On the basis of early attested toponyms like *Gennor* and *Mandara*, whose endings were over time replaced with the inherited plural suffix *-s*, resulting in *Gennos* and *Mandas* (§ 9.2.6.2), Terracini (1927: 139) has convincingly analyzed *-r* as a plural marker. The prefix **ϑi-* is found in lexical items and has been interpreted as a diminutive marker (Forsyth Major 1893: 154; Guarnerio 1904a: 57–58; 1904a: 260; Wagner 1932: 223–224; 1997: 263) as well as a morphosyntactic element (Pittau 1995: 197; Swanenvleugel 2024)(see § 9.1.2). It goes to show that, unless there is a fortuitous attestation of relevant forms (as was the case for **-Vr-*), establishing the original function of a non-productive morpheme can remain challenging despite more than a century of debate.

2.3 Interpreting results

Once the evidence has been examined in the greatest possible detail, a logical next step is to offer a linguistic scenario that is consistent with the findings. If sufficient evidence for a substrate language has been identified, possible follow-up questions are:

1. Stratification: Was there a single substrate language or were there multiple?
2. Sociolinguistics: What was the nature of contact between the substrate language and the attested language leading up to the eventual language shift?
3. Chronology: For how long did the substrate and superstrate languages coexist?
4. Origin: What was the linguistic affiliation of the substrate language(s)?
5. Interdisciplinarity: Is there any extralinguistic (e.g. historical, archeological or genetic) evidence corroborating the linguistically inferred scenario?

It is unlikely that all of these questions can be satisfactorily answered for every single linguistic substrate. Even in a region like southern Europe, where relatively many non-Indo-European languages have survived to the brink of history (e.g. Tartessian, Iberian, Etruscan, Rhaetian, Minoan, Lemnian, Eteocypriot) and where the existence of other, non-attested languages is apparent from classical sources, it is still difficult to reconstruct the specifics of linguistic affiliation and the degree of contact between these languages.

Situations of language contact are inherently complex. A survey of modern-day Sardinia may serve as an illustration. In most of the island Sardinian is spoken, but with considerable variation between the various dialects. In addition, there are areas with Catalan, Ligurian and Corsican dialects. To add to this complexity, Italian has imposed itself as a superstrate to all these speech varieties. If someone in the future, when the shift to standard Italian has been completed,[39] were to attempt to reconstruct the island's former linguistic diversity, chances are low that they would succeed to fully capture it. There is *a priori* no reason to think that Sardinia in pre-Roman and Roman times was linguistically less diverse. It is thus important to acknowledge the fact that many aspects of the original linguistic landscape are irrecoverable. Any hypothesis claiming the opposite is inescapably wrong. This does not, however, mean that all attempts to elucidate linguistic prehistory are necessarily futile. Even extracting the tiniest bit of information on the past linguistic configuration of a certain region, is potentially of great value to our understanding of general (pre-)history.

2.3.1 Unity of the substrate

A good starting point for the investigation of any prehistoric linguistic situation, is to assume that it was to some extent linguistically diverse, unless there are strong reasons to assume homogeneity. When analyzing the corpus of non-inherited material in a given language, we should therefore ask ourselves the question whether the assumption of a single extinct language can account for all the evidence.[40] One way to demonstrate such linguistic homogeneity, is to analyze the dialectal distribution of the features of the linguistic substrate: lexemes, phonology, morphemes etc. If a set of features is consistently found in a particular area, this may point to the existence of a distinct language limited to that area. Conversely, if many different features are found across all of the speech varieties in the researched region, it seems more likely that a single language was spoken throughout that region. Naturally, the latter scenario does not preclude that other languages were present, in smaller pockets, like in present-day Sardinia.

With regard to the linguistic situation on Sardinia at the time of the Roman conquest, at least two languages were spoken on the island. One was Phoenician, the language of Punic settlers from Carthage, who left ample inscriptional evidence

[39] This is, and will hopefully remain, a purely hypothetical scenario.

[40] For a case in point, cf. Beekes' (2010: xli–xlii; 2014: 45) assumption of the unity of pre-Greek, and the critique by Verhasselt (2011: 279), Meissner (2013: 6–7), De Decker (2015: 5), and Meester (2024).

(Bondì 1987d; Adams 2003: 209). In addition, one or more non-colonial languages native to Sardinia were spoken, but these left no written record. In order to be able to stratify the pre-Roman lexicon, it is necessary to analyze and interpret the distributions of non-inherited lexemes, and of their phonological (§ 8) and morphological features (§ 9).

2.3.2 Nature of language contact

Thomason and Kaufman (1988: 121) argue that it is impossible to predict the outcome of linguistic interference in case of a language shift. The opposite, interpreting the nature of language contact on the basis of linguistic interference, is not much easier. As discussed in § 1.3.1.3, it is difficult to prove that a certain feature of Sardinian reflects influence from a substrate language. However, on the basis of the lexical data, discussed in this study, we may attempt to employ linguistic paleontology to make inferences about the nature of economic and cultural contacts between speakers of Latin, Phoenician, and the other pre-Roman language(s) spoken on Sardinia.

2.3.3 Dynamics of language shift

The question of the nature of the language contact between Latin and the preexisting languages in Sardinia is intrinsically linked to the romanization process (§ 1.2). Multiple scenarios can be considered. Did native Sardinians shift to Phoenician prior to shifting to Latin? If not, did speakers of Phoenician and other pre-Roman language(s) shift to Latin at the same time, or did one language outlive the other(s)? The patterns of distribution of non-inherited linguistic elements can be employed to shed more light on the various scenarios that can be proposed for the relative chronology and direction of the romanization process.

2.3.4 Classifying a non-attested language

A question that has proven particularly popular in the study of linguistic substrates is to which languages they are related. The Sardinian substrate is no exception. It has been proposed to be related to Basque, Etruscan, and Berber, as well as to other unattested hypothetical languages constituting substrates elsewhere in the Mediterranean (cf. § 1.3.2, § 1.3.2.1). Like in the case of living languages, the classification of substrate elements as belonging to a certain language family should ideally be based on morphological as well as lexical evidence. The difference with living languages is however that the available material for substrate languages is usually much scarcer, especially in the morphological domain, and intrinsically transmitted indirectly through the language that replaced it.

In order to confidently posit a genetic link between a presumed substrate language and another language, the hypothesis must be able to account for a significant portion of the identified substrate elements. In Sardinia, this is the case for Punic, to which a number of non-inherited words can be ascribed (§ 10.1) (Paulis 1990).[41] A hypothetical genetic link between two unattested substrate languages is harder to evaluate, due to the greater amount of uncertainties in these cases. It is difficult to distinguish "native" substrate words from *Wanderwörter* (e.g. into 1st century BCE Sardinian Latin), or from words borrowed into the substrate language (e.g. a loan from Greek into 4th century BCE Sardinian Punic). We must thus rely on the number of shared forms between two putative substrate languages when evaluating the likelihood of their genetic relatedness. An important caveat is that it cannot be expected that the geographical extent of extinct languages coincides exactly with that of modern languages. In order to reliably interpret the potentially putatively shared between two substrate languages, close attention should be paid to the geographical distribution of these features on both sides of the comparison. This will be discussed in § 11.

2.3.5 Extralinguistic arguments

Once we have formulated hypotheses about the pre-Roman linguistic configuration, and subsequent romanization of Sardinia on the basis of linguistic evidence, a possible next step is to establish whether these can be corroborated by other lines of evidence. Such evidence may come from the fields of archeology, genetics, or from historiographical sources. For overviews of Sardinian archeology, see Lilliu (1963; 2002), Webster (1996), and Dyson and Rowland (2007). In recent times, Sardinia has been the subject of numerous genetic studies. Both ancient and modern Sardinian populations display a remarkably low degree of steppe-derived ancestry, whose prominence in the rest of Europe has been connected to the spread of the Indo-European languages (e.g. Chiang et al. 2018; Fernandes et al. 2020; Calò et al. 2021). These non-linguistic lines of evidence are best considered by experts in the respective fields, and are therefore left outside the scope of this study.

2.4 Conclusion

When it comes to methodology of substrate research, an important realization is that the evidence should be treated with great caution, due to the many uncertainties surrounding extinct, unattested, indirectly transmitted languages. We

[41] Cf. § 10.3.3, where it is argued that Basque–Sardinian comparisons do not live up to this expectation, thus making Basque an unlikely relative to the Sardinian substrate.

must accept that it will never be possible to recover all of the intricacies of lost prehistoric languages. To avoid ending up with grandiose theories (cf. § 1.3.2), only the strongest possible evidence should be used when formulating hypotheses about the pre-Roman languages of Sardinia. Although this means that less of the non-inherited material is exploited to reach conclusions, the conclusions that are reached should be relatively more reliable. In practice, I argue for deploying a maximal degree of strictness in the various avenues of substrate research discussed. Some points that are observed throughout this study are the following:

- Etymological obscurity does not necessarily imply a substrate origin. The latter should only be posited in the presence of compelling positive evidence, e.g. non-native phonemes, irregular alternations, irregular cognates. Isolated words of unknown origin generally do not provide reliable evidence. Cf. § 2.1.1.
- On the other hand, an inherited etymology should not involve too many unmotivated irregular sound changes aor non-trivial semantic shifts. Cf. § 2.1.2.2.
- Comparisons between non-inherited comparanda can only be plausibly made when they display sufficient similarity and when semantic differences are trivial. Cf. § 2.1.2.2.
- The previous point generally rules out toponyms as evidence for substrate language relatedness, as their semantic opacity means that the risk of formal chance correspondence cannot be mitigated by semantic comparison. Cf. § 2.1.2.2.1.
- Phonological alternations are only potentially indicative of phonological features of the substrate language if they occur sufficiently frequently throughout the non-inherited lexicon. Cf. § 2.2.1.
- Morphological features can only be posited for a substrate language if there is sufficient positive evidence, i.e. an alternation between non-inherited endings. “Morpheme-hood” should only be posited based on language-internal evidence, as positing suffix alternation on the basis of forms in different languages renders both the comparison of the forms and the identification of the suffix circular. Cf. § 2.2.2.

It goes without saying that all of these points are subject to a researcher’s personal preferences and intuitions. Still, they provide guidelines, followed throughout the rest of this study, that allow us to distinguish plausibilities from uncertainties, and to formulate hypotheses based on the former.