Seminar 5
Comparative reconstruction and language families

Contents

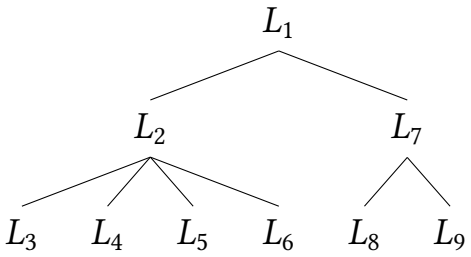
A Linguistic descent and relatedness

§1 A language (more precisely, a grammar) L is said to **descend** from another language (grammar) L' if L is connected to L' by a series of first-language (L1) acquisition events. In this case we also say that L' is an **ancestor** of L. Examples:

- Modern French descends from Old French
- Old French descends from Latin
- (By transitivity:) Modern French descends from Latin, also from Proto-Indo-European, and so on…
- My grammar of Finnish descends from my parents' grammar of Finnish
- Finnish does *not* descend from Latin — it is impossible to show that there is a sequence of L1 acquisition events between the two.

§2 How can we know anything about the ancestors of currently spoken languages? One strategy is to take a handful of languages we suspect to be closely related, compare them, and from this comparison try to infer properties of their ancestor. This is known as **comparative reconstruction** or the **Comparative Method** (henceforth, CM).

§3 What does it mean to say that two languages are "closely related"? To simplify a bit, what we have in mind is the following kind of situation:



Here, the languages L_3 through L_6 are closely related, since they share the same immediate ancestor, L_2. For example L_6 and L_8, on the other hand, are not related at the same time depth (though they are still related through the common ancestor L_1).

We can use comparative reconstruction on languages L_3 through L_6 to try and reconstruct the properties of L_2. Or we can use it on L_8 and L_9 to try and reconstruct the properties of L_7. Or we can use it on each of L_2–L_9 to try and reconstruct L_1.

A bit of terminology: we say that L_3 through L_6 are **sister languages** and the **daughters** of L_2. It is also sometimes said that L_3 through L_6 stand in a **genetic relationship** by virtue of the common ancestor L_2. (Note that this has nothing to do with biological genes! In this usage, the word just descends from Ancient Greek γένεσις, 'origin'.)

§4 It is useful to take a real-life example and to break the CM down into a few steps, following the lines of Campbell (2013).

B Step 1: Compilation of cognate sets

§5 The first step in the CM is to pick a set of languages we suspect are closely related and a set of words which appear to be similar in both meaning and sound structure across these languages. Here's an example from Romance:[1]

Table 1: Four Romance cognate sets.

	Italian	Spanish	Portuguese	French	gloss
1.	*capra* /kapra/	*cabra* /kabɾa/	*cabra* /kabɾa/	*chèvre* /ʃɛvʁ/	'goat'
2.	*caro* /karo/	*caro* /kaɾo/	*caro* /kaɾu/	*cher* /ʃɛʁ/	'dear'
3.	*capo* /kapo/	*cabo* /kabo/	*cabo* /kabu/	*chef* /ʃɛf/	'head, top'
4.	*cane* /kane/	*can* /kan/	*cão* /kãw̃/	*chien* /ʃjɛ̃/	'dog'

[1]Note that these four languages are not *all* the Romance languages — I limit discussion to the four in the interest of simplicity. Similarly, I here abstract away from the tremendous amount of regional variation found in all these languages. The labels here refer to standard varieties, but it should be kept in mind that Italian, itself, branches into a number of dialects each of which is a distinct grammar. Similarly, Peninsular Spanish and Argentinean Spanish are, strictly speaking, different grammars and hence different languages; European and Brazilian Portuguese are different; Metropolitan and Quebec French are different; and so on.

Words such as Italian *capra*, Spanish and Portuguese *cabra* and French *chèvre* are **cognates**: they have a common origin (in this case, a Latin word meaning 'goat'). By extension, each line in the above table is known as a **cognate set**. For comparative reconstruction to work, we first need to collect many of these. Here I have just four, in the interest of simplicity.

§6 A couple of important points:

1. Comparative reconstruction must be conducted on the words themselves, i.e. on the sound forms, not on the written forms. (I have included the written forms in the above table for convenience, but in practice they play no role in reconstruction. For example, we must not be led astray by the fact that the French word for 'dog' has an <n> at the end of its written manifestation — there is in fact no final consonant in the phonemic form itself.)
2. The words in a cognate set must be cognates, not borrowings or other random innovations. For example *can* /kan/ is archaic in Modern Spanish, which nowadays more commonly uses *perro* /pero/ (etymology unknown) for 'dog' instead. However, it would be a mistake to include *perro* in the above table instead of *can*, since it is clearly a different word that at some point in the history of Spanish replaced the native *can*.

C Step 2: Establishment of correspondence sets

§7 The second step in the CM is to establish sound correspondences between the sister languages. From Table 1, we can extract the following **correspondence set**:

(1) Italian k- : Spanish k- : Portuguese k- : French ʃ-

Here I adopt the common practice of leaving out the phoneme slashes and using a trailing dash to indicate 'word-initial'. So the above correspondence set says that where Italian, Spanish and Portuguese have word-initial /k/, French has /ʃ/.

§8 We can also extract the following correspondence sets (∅ means empty, i.e. no sound, -*x*- is word-medial, -*x* is word-final):

(2) Italian -p- : Spanish -b- : Portuguese -b- : French -v-

(3) Italian -a- : Spanish -a- : Portuguese -a- : French -ɛ-

(4) Italian -a : Spanish -a : Portuguese -a : French -∅

And many others, but we'll only focus on these for now.

D A word of warning: false cognates and unsystematic correspondences

§9 For the CM to work, we need many cognate sets (in the above table I only have four, for reasons of simplicity). Sometimes two words in two different languages

happen to have the same meaning and similar sound forms by accident. E.g. English *day* (from OE *dæg*, from Proto-Germanic **dagaz*[2]) and Spanish *día* (from Latin *dies*, from Proto-Indo-European **dyews*). If we only used this one pair of words, we might want to place English among the Romance languages. This would be wrong — English *day* and Spanish *día* are **false cognates**. In fact, it is impossible to find **systematic correspondences** between English and Spanish words in the way we can find between Spanish and Italian, for example:

Table 2: English is not a Romance language.

	Italian	Spanish	English
1.	*dì* /di/[3]	*día* /dia/	*day* /deɪ/
2.	*dente* /dɛnte/	*diente* /djente/	*tooth* /tuθ/
3.	*dito* /dito/	*dedo* /dedo/	*finger* /fɪŋər/
4.	*dare* /dare/	*dar* /dar/	*give* /gɪv/

§10 Having said that, if two languages are more distantly related, they will have cognate words but the correspondences are much more difficult to establish. English and Spanish are a good example, in fact: with much more work, and using much more data, it is possible to trace both *tooth* and *diente* to a reconstructed Proto-Indo-European word **hɜdonts*. This is just to say that both English and Spanish are Indo-European languages (as opposed to, say, Finnish, which is Uralic). However, the kinds of systematic correspondences found within the Romance family (such as Italian and Spanish agreeing for initial /d/ in the above cognate sets) are simply not available when we compare Spanish and English. (The only exception is if we focus on words which were borrowed into English from Latin or Norman French, such as *diabolical*, *diocese*, *deacon*, and so on. But this is a case of **horizontal transfer** between different branches of a genetic tree (more on this later) and therefore not a demonstration of (close) genetic relatedness.)

E Step 3: Reconstruction of proto-phonemes

§11 The next step is to infer back from the sound correspondences to the protolanguage, i.e. try and reconstruct the sounds of the ancestor. Here we must rely on a couple of rules of thumb or principles.

[2]Earlier, we've seen the asterisk (*) used for ungrammaticality. Unfortunately, in historical and reconstructionist linguistics the same symbol is also used to indicate reconstructed forms, i.e. forms which are not attested in artifacts but for which only reconstructive evidence (arrived at through the CM) exists. I will persist in propagating this confusion by continuing to employ the same symbol for the two uses; fortunately there is hardly ever any risk of confusion since the context indicates which reading is intended.

[3]Like Spanish *can*, this is archaic; Modern Italian uses *giorno* /dʒorno/ (from Latin *diurnum*).

§12 **Majority wins**. The majority wins principle urges us to reconstruct that sound which is most prevalent in the daughter languages. Looking at our first correspondence set in (1), for instance, we are compelled to reconstruct initial /k/ (rather than /ʃ/, or something else entirely) for proto-Romance (i.e. Latin). The alternative hypothesis (proto-/ʃ/) would require three languages to have undergone the change /ʃ/ > /k/ instead of one language undergoing the reverse change; and if we reconstructed something else entirely, we would need to posit four changes in total.

In the case of Latin there is also other, converging evidence that favours this reconstruction, but for lesser studied languages, or for languages with scant written records (often the protolanguage we want to reconstruct isn't attested at all!) such evidence is normally lacking.

§13 **Directionality**. The principle of directionality urges us to reconstruct only such proto-phonemes that do not imply postulating directions of change which are known not to be attested, or attested only very infrequently. The change /k/ > /ʃ/, for example, is documented in a number of languages, while the opposite change /ʃ/ > /k/ does not seem to occur (Campbell, 2013). Thus, the principle of directionality also supports our decision to reconstruct word-initial /k/ for proto-Romance.

§14 Sometimes the two principles give conflicting suggestions. Looking now at our Romance correspondence set (2), the majority wins principle suggests reconstructing -b-. The directionality principle, however, suggests reconstructing -p- in this case. This is because of the prevalence of lenition (recall Seminar 3) as a pathway of change in the world's languages. Briefly put, it is more common for a voiceless stop to become voiced than the other way around, particularly if this stop occurs intervocalically, as is the case in the third cognate set in Table 1. (To see this, note that vowels are intrinsically voiced. Producing a voiceless stop between two vowels thus requires more effort than producing a voiced one: in the first case, one needs to stop the vocal folds from vibrating and then make them vibrate again, whilst in the second case the vocal folds can keep vibrating all along.)

In cases like this we need to be careful. Basically, we have three options:

1. Go with majority wins and reconstruct -b- for proto-Romance. Italian then fortified into -p-, while French lenited even further into the fricative -v-. (Problem: why did Italian fortify, if fortification is uncommon?)
2. Go with directionality and reconstruct -p-. Spanish and Portuguese lenited to -b-, while French underwent -p- > -b- > -v-. (Problem: if voiceless intervocalic stops are "unnatural", how could Latin have them in the first place?)
3. Leave the matter undecided and simply reconstruct a labial stop, without committing ourselves to a specific value on the voicing feature. (Problem: less accurate reconstruction, details missing.)

On balance, it seems that in this case the best option to choose is number 2. However, when reconstructing languages for which less data (and no contemporary grammatical testimony) is available, we may have to settle for number 3.

§15 Which brings us to an important point: historical reconstructions are always approximate and probabilistic. We cannot say we know with 100% certainty that intervocalic stops in Latin were voiceless; rather, we have a degree of confidence in this hypothesis, which may have to be revised if further data comes along. Generally, the deeper the timescales involved the harder and less certain reconstruction becomes: reconstruction of Proto-Indo-European is much more difficult, and much less certain, than reconstruction of Proto-Romance, and reconstruction of "Proto-World" is, generally speaking, impossible.

F Step 4: Resolution of overlapping correspondence sets

§16 Here is more data from Romance:

Table 3: Another four Romance cognate sets.

	Italian	Spanish	Portuguese	French	gloss
1.	*colore* /kolore/	*color* /kolor/	*côr* /kor/	*couleur* /kulœʁ/	'colour'
2.	*correre* /korere/	*correr* /korer/	*correr* /korer/	*courir* /kuʁiʁ/	'to run'
3.	*costare* /kostare/	*costar* /kostar/	*costar* /kostar/	*coûter* /kute/	'to cost'
4.	*cura* /kura/	*cura* /kuɾa/	*cura* /kuɾa/	*cure* /kyʁ/	'cure'

From these data, we extract the following correspondence set:

(5) Italian k- : Spanish k- : Portuguese k- : French k-

This correspondence set now needs to be reconciled with correspondence set (1), which is also about word-initial /k/ but differs from (5) in that French has /ʃ/ instead of /k/. We have three possibilities:

1. The protolanguage had two sounds, /k/ and /ʃ/. French maintained this distinction, whereas Italian, Spanish and Portuguese **merged** the two sounds into /k/.
2. The protolanguage had /k/, and French **innovated** the sound /ʃ/ in some contexts.
3. The protolanguage had /ʃ/, which Italian, Spanish and Portuguese turned into /k/, whilst French innovated /k/ in some contexts but retained /ʃ/ in others.

In this case, both majority wins and directionality support hypothesis 2. In fact, we can construct a convincing case for the change /k/ > /ʃ/ in French. From Table 1, we find that in French, word-initial /ʃ/ is always followed by the vowel /ɛ/. From correspondence set (3), on the other hand, we reconstruct earlier /a/ for later French /ɛ/. The story, then, is as follows:

1. In French, /a/ first changed into /ɛ/.

2. Word-initial /k/ then underwent change into /ʃ/ whenever followed by the new vowel /ɛ/. In symbols: k > ʃ / #_ɛ.[4]
3. When followed by other vowels, initial /k/ remained.

G Step 5: Reconstruction of phonotactics

§17 Working on a large number of cognate sets, it is possible to try and reconstruct the entire phoneme inventory of a protolanguage. Once this is in place, the next step is to try and reconstruct the phonotactics, i.e. the rules whereby sounds are put together to form morphemes. This then allows the reconstruction of entire word forms.

From the above Romance data, for instance, we would reconstruct /kVr/ as a possible (word-initial) syllable in the protolanguage, where V stands for any vowel. On the other hand, there is no support for a syllable like /krk/ in the daughter languages, so we shouldn't reconstruct anything like it.[5]

H Step 6: Sanity checks against language typology

§18 Once the reconstruction of the entire sound system is in place, we must check whether the reconstructed system "makes sense", so to speak. For instance, if our reconstruction leads us to propose a sound system with only nasal vowels (and no oral vowels), we should view this reconstruction with suspicion, as there are no known languages with only nasal vowels. We also shouldn't reconstruct sounds which are impossible to articulate (such as a "glottal trill").

§19 Sometimes symmetry considerations also play a role in checking the sanity of a reconstruction: for example, reconstructionists would typically view with suspicion a proposed system with the voiceless stops /p t k/ but an incomplete voiced series corresponding of just the voiced velar stop /g/ (i.e. if a language has /p t k/, we would assume it to have each of /b d g/ as well). Symmetry considerations of this kind are, however, not absolute, since it is possible to find languages that violate them, and hence they are not absolute typological universals.

§20 When reconstructing phonotactics, we also need to check we are not postulating sound clusters which are deemed impossible or too difficult to pronounce on purely articulatory grounds. For example, a consonant cluster like /pgtkb/ practically has a zero probability of occurring in a human language, so we shouldn't reconstruct anything like it.

[4]Read this as: /k/ changed into /ʃ/ in the context: word-initial, followed by /ɛ/. The symbol # marks word boundary.

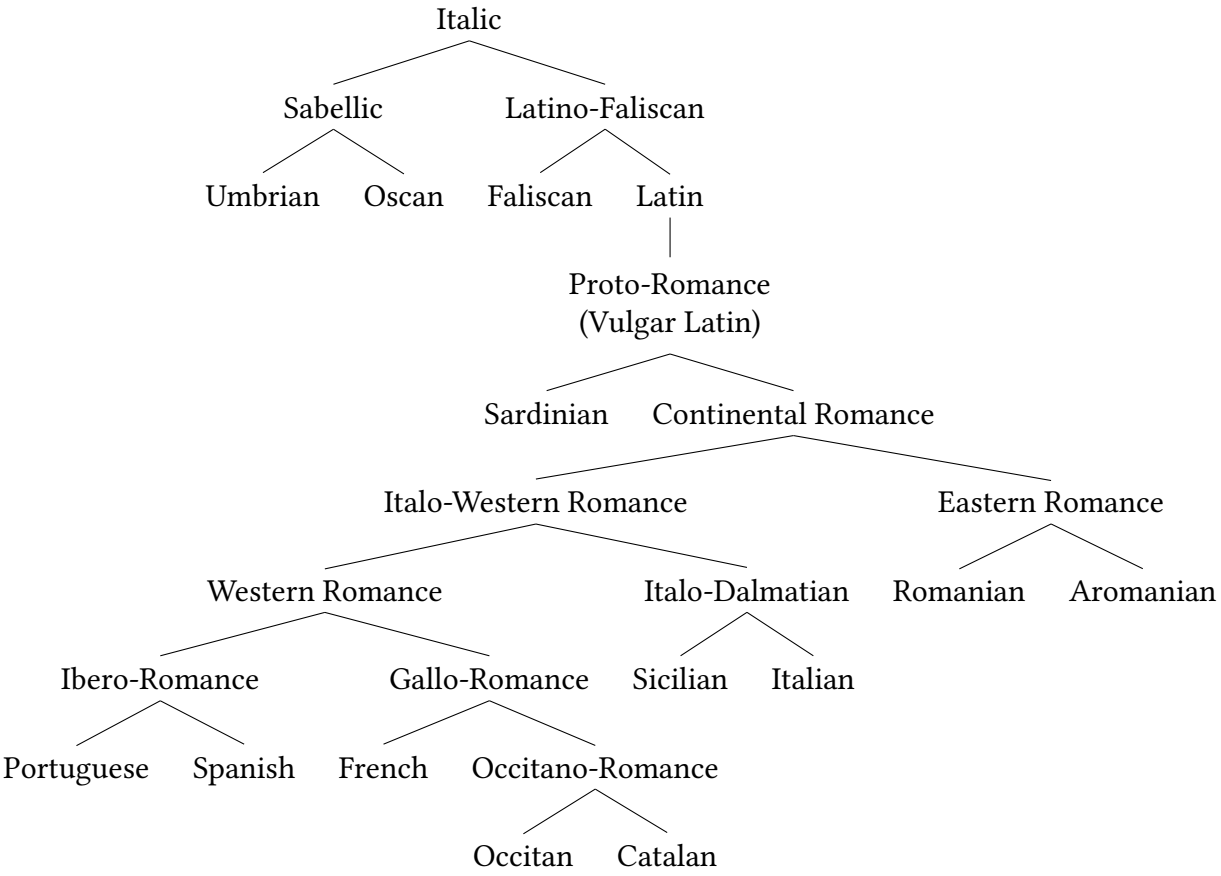
[5]Such syllables are possible in other languages, notably in many languages of the Slavic family.

I Families and subgrouping

§21 Application of the CM gives, indirectly, evidence of language relatedness. We started with the idea that Italian, Spanish, Portuguese and French were somehow related. The fact that the CM allows the reconstruction of a protolanguage for these four languages points to the conclusion that they are indeed related — compare our attempted comparative reconstruction with Spanish and English in Table 2, which failed.

§22 More generally a **language family** is a group of languages that are related. More precisely, two languages belong to the same family if the two languages share a common ancestor.

Language families can be specified on various levels of granularity, e.g. Spanish belongs not just to the Romance family, but also to the family of Ibero-Romance languages (a subfamily of the Romance languages) and to the Italic family (a superfamily of Romance), as illustrated here:[6]



This also means that some families are subsets of other families (e.g. Romance is a subfamily of Italic).

§23 Language families are posited on various kinds of evidence, mostly based on the CM (but later on we will see other methods of establishing relatedness, particularly on longer timescales which are out of the reach of the CM). In a family tree,

[6]It should again be stressed that this tree does not contain all the Romance, or all the Italic, languages — just a selection of them to illustrate the broad outlines!

each branching point is associated with one or more **innovations**. For example, in our above Romantic reconstructions intervocalic lenition (VpV > VbV) separates Spanish, Portuguese and French from Italian; and the first three languages indeed form a subfamily of their own, apart from Italian, known as Western Romance (see above tree).[7]

It should be pointed out that a great deal of disagreement exists among scholars about the precise placement of languages on these trees, or about how to place the branching points corresponding to now extinct protolanguages. The broad outlines, however, are clear — no one would contest the claim that Spanish is a Romance language, for example.

§24 Since each branching point in a tree corresponds to some innovation(s), distances between languages in a tree correspond to differences between languages — the more separation in the tree between two languages, the more different they ought to be (unless later changes undo earlier ones, but this is highly unlikely). E.g. we would expect Sardinian to be closer to Latin than, say, Catalan. This is in fact the case, comparing for example the indicative paradigm (present tense) of the Latin verb *cantare* 'to sing':

Table 4: Comparison of Latin (reconstructed), Nuorese Sardinian and Central Catalan: present indicative of *cantare*.

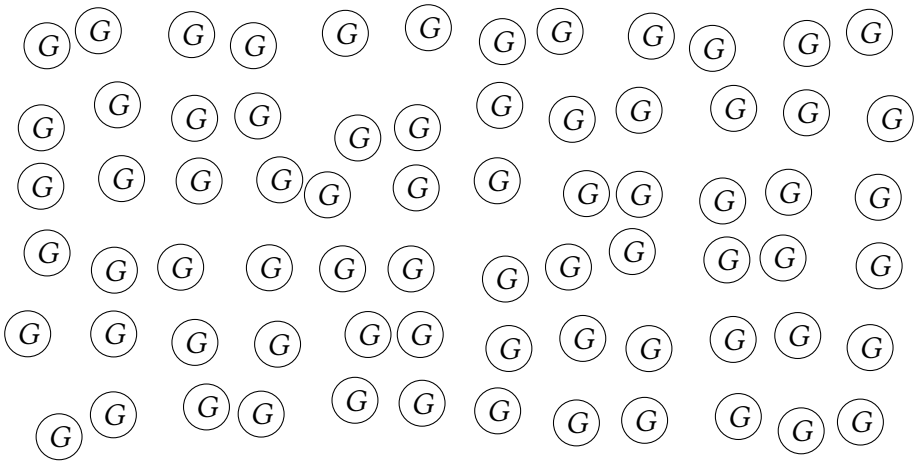
	Latin	Sardinian	Catalan
1sg	/ˈkanto:/	/ˈkanto/	/ˈkantu/
2sg	/ˈkanta:s/	/ˈkantaza/	/ˈkantəs/
3sg	/ˈkantat/	/ˈkantata/	/ˈkantə/
1pl	/kanˈta:mus/	/kanˈtamuzu/	/kənˈtɛm/
2pl	/kanˈta:tis/	/kanˈtateze/	/kənˈtɛw/
3pl	/ˈkantant/	/ˈkantana/	/ˈkantən/

J A note on grammars

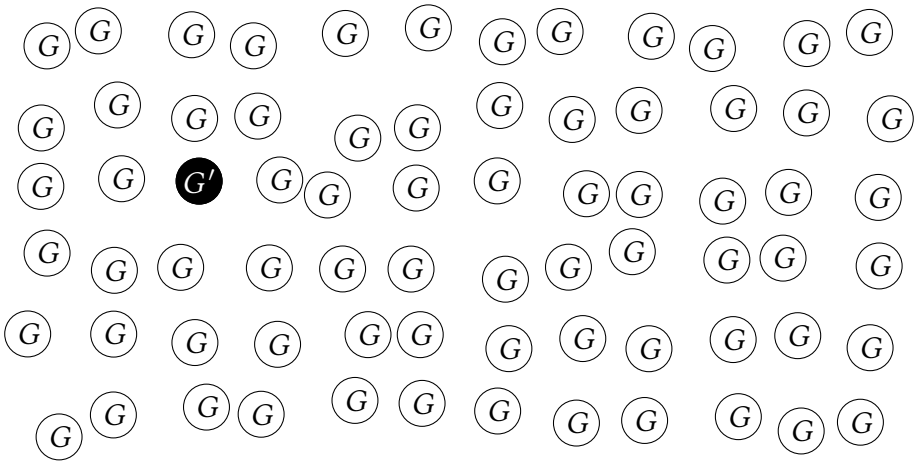
§25 Recall that previously we have insisted that the proper object of study of linguistics, both synchronic and historical, is the grammar, i.e. the I-language. Yet here we have been speaking loosely about languages: "Spanish descends from Latin", "Spanish and Portuguese are sister languages", and so on. Does this represent a new ontological confusion?

[7] Strictly speaking the lenition is not restricted to just intervocalic sequences: compare Italian /kapra/, Spanish /kabra/ and French /ʃɛvʁ/. Here the change is fortis > lenis / V_R, where V is any vowel and R is a rhotic. French lenites more than Spanish in the sense that the fricative /v/ is higher up in the sonority hierarchy than the voiced stop /b/ (see Seminar 3; though it should be pointed out that in this kind of context the Spanish sound is usually produced as the bilabial fricative [ß]). It is probable that the pathway French took was in fact something like p > b > ß > v.

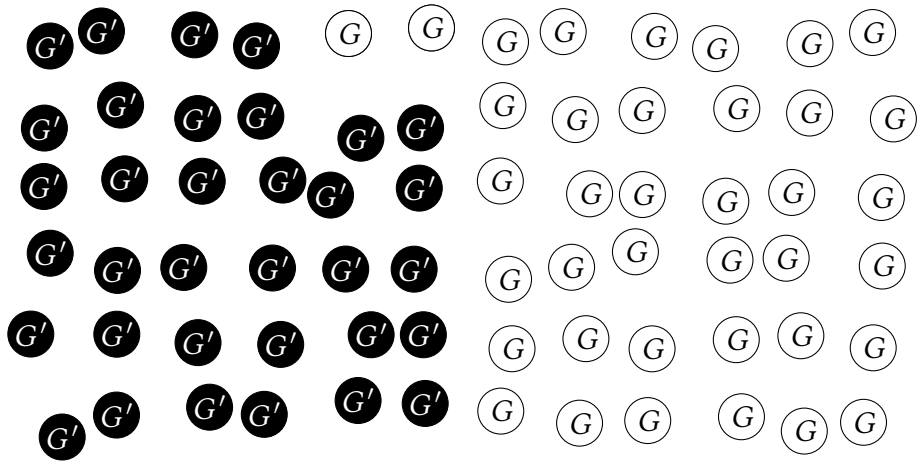
§26 The key to understanding the relation between grammars and languages is to see languages as **populations** of grammars (I-languages). Consider the following picture, in which every circle represents one speaker, every speaker sharing the same grammar *G*; for the purposes of illustration, let's suppose that this is the grammar of Italo-Western Romance (cf. above tree):



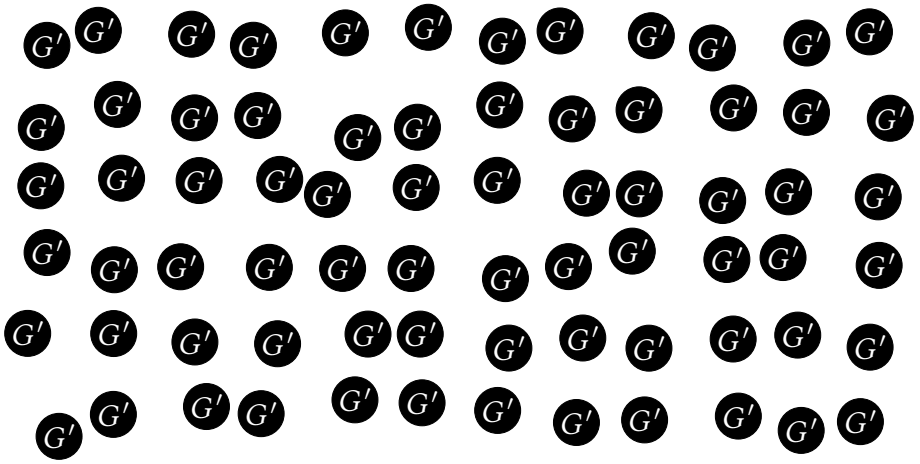
The first prerequisite for family branching is for there to be an **innovation** event somewhere in the population. Suppose one speaker innovates phonological lenition of intervocalic stops (e.g. VpV > VbV); let's call this grammar G':



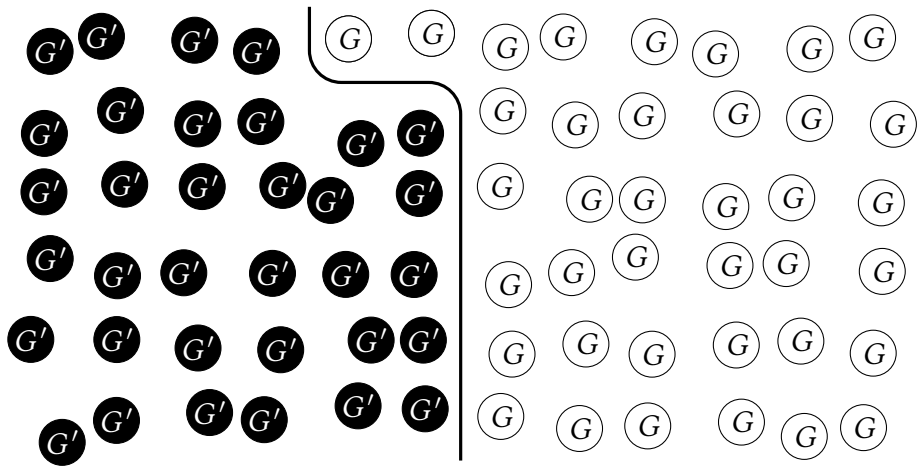
If the innovation propagates to other members of the population, we can say that a **change has occurred**:



Now, if the change propagates to *all* members of the population, what we have at hand is simply a change in Italo-Western Romance, without any branching:



However, if propagation stops at some point, e.g. due to a geographical or political **boundary**, the original population **diversifies** into two subpopulations, one of them with grammar G (Italo-Western Romance) and the other with G' (Western Romance), embodying intervocalic lenition of stops:



This population diversification then corresponds, in broad, abstract terms, to the familial branching of Italo-Western Romance into Western Romance and Italo-Dalmatian. In this case, Western Romance embodies the innovatory feature (lenition) whilst Italo-Dalmatian carries the proto-feature (no lenition). Subsequent developments within Italo-Dalmatian, of course, serve to diversify that language further, e.g. in the split into Sicilian and Mainland Italian dialects.

§27 Thus, even though it is convenient to speak of languages diversifying and branching and standing in descent and sibling relationships, we should always bear in mind that what is at issue is individual grammars innovating and these innovations spreading across populations of speakers.[8]

[8]The above population diagrams also abstract away from a "small" detail: that speakers die and new ones are born (note that language family branching such as the passage from Proto-Romance to the currently spoken Romance languages takes a long time, far longer than the lifetime of any single speaker). Here I simplify matters by assuming that every pair of speakers produces, on average, two

K Review

§28 After this seminar, you should be able to explain what the following terms mean:

linguistic descent	cognate set	directionality
ancestor	correspondence set	language family
Comparative Method	false cognate	innovation
sister language	horizontal transfer	propagation
daughter language	proto-phoneme	branching
genetic relationship	majority wins	

L Further reading

§29 A very good, beginner-friendly hands-on treatment of the CM, complete with exercises (though lacking somewhat in the epistemological side of reconstruction), appears in Campbell (2013, ch. 5). I have relied on Campbell's exposition here heavily. Another good contemporary textbook treatment is to be found in Ringe and Eska (2013, ch. 10). More in-depth theoretical and critical discussion can be found in Lass (1997, chs. 3–5).

References

Campbell, L. (2013). *Historical linguistics: an introduction* (3rd ed.). Edinburgh: Edinburgh University Press.

Lass, R. (1997). *Historical linguistics and language change*. Cambridge: Cambridge University Press.

Ringe, D. A. and J. F. Eska (2013). *Historical linguistics: toward a twenty-first century reintegration*. Cambridge: Cambridge University Press.

offspring, thereby keeping population size constant. Other matters, such as language contact and borrowing, also introduce complications which it is not possible to discuss here, but to which we will return toward the end of this course.