

LINGUISTICS COURSE

# Historical Linguistics and Language Families

Why languages change, how linguists prove relationships, and how the world's languages are grouped into families.

10 Chapters

Philip Osztromok

Generated with Claude

# Contents


**Chapter 1** Why Languages Change ..... 2

**Chapter 2** Sound Change ..... 7

**Chapter 3** The Comparative Method ..... 12

**Chapter 4** Reconstruction ..... 17

**Chapter 5** Language Families ..... 22

**Chapter 6** Trees and Waves ..... 27

**Chapter 7** Meaning and Grammar Change ..... 32

**Chapter 8** Evidence from the Past ..... 37

**Chapter 9** Beyond Indo-European ..... 42

**Chapter 10** Putting It to Work ..... 48

# Historical Linguistics and Language Families

Course 1 · Chapter 1 · Why Languages Change

Phonetics and the IPA 8 showed that in a single sentence *going to* can become *gonna* and *handbag* can become [ˈhæmbæg]. Now ask what happens when changes like that are repeated by millions of speakers for a thousand years. The answer is that one language becomes several, and that is the subject of this course.

> **THE IDEA IN A BOX**
>
> **Languages are always changing.** Historical linguistics is "the scientific study of how languages change over time" (Wikipedia). Its job is to describe the changes, explain them, work out which languages are related, and reconstruct the ancestors nobody recorded.

## Change Is Normal

Wikipedia's article on language change defines it as the alteration of a language's features, or of languages in general, over time. It usually happens gradually, with old and new forms coexisting for a while. The same article rejects the word "corruption": innovations cannot be judged good or bad scientifically. Modern English descends from Old English, and the two are different enough that a modern speaker cannot read the older form without study.

## Two Ways of Looking at a Language

| APPROACH | LOOKS AT | QUESTION |
|---|---|---|
| Synchronic | a language at one point in time, usually the present | how does it work now? |
| Diachronic | developments through time | how did it get this way? |

Historical linguistics is mainly diachronic. The article gives a neat reason why the two views need each other: synchronic forms result from earlier diachronic changes. English *sing / sang / sung* looks irregular today, but it reflects a once-regular Indo-European vowel-change system.

The field rests on the **uniformitarian principle**: processes of change seen today were likely at work in the past, unless evidence suggests otherwise. That is why watching English change now is evidence for how it changed before.

## What Changes

| KIND | WHAT CHANGES | EXAMPLE FROM THE SOURCES |
|---|---|---|
| Phonetic / phonological | how phonemes are pronounced; phonemes can merge and reduce in number | loss of rhoticity in Received Pronunciation (Phonetics and the IPA 7) |
| Lexical | new words are borrowed or recombined; old words are lost | (general; later chapters give cases) |
| Spelling | standardisation began after the printing press in the late 15th century | English spelling no longer tracks speech (Phonetics and the IPA 1) |
| Semantic | meaning shifts | pejoration (*villain*), amelioration (*wicked*), broadening (*dog*), narrowing (*hound*) |
| Syntactic | sentence structure; can follow creolisation, relexification, or phonological or morphological change | (later chapters) |

### Meaning change in detail

Wikipedia's article on semantic change uses Bloomfield's 1933 scheme and gives these dated cases:

| TYPE | WORD | SHIFT |
|---|---|---|
| Amelioration | *nice* | "agreeable, delightful" (1769), then "kind, thoughtful" (1830) |
| Amelioration | *terrific* | "terrifying", then "astonishing", then "very good" |
| Pejoration | *awful* | "awe-inspiring" to "very bad" |
| Narrowing | *skyline* | any horizon, then one lined with skyscrapers |
| Metaphor | *broadcast* | casting seeds, then transmitting signals |

| TYPE | WORD | SHIFT |
| --- | --- | --- |
| Metonymy | *jaw* | "cheek", then "mandible" |

Chapter 7 returns to meaning change and to grammar change.

## Why Does It Happen?

The language-change article lists these causes:

- **Economy**: least effort, leading to reductions like *going to* to *gonna*.
- **Expressiveness**: overused language loses intensity.
- **Analogy**: forms are reshaped on the pattern of other forms.
- **Language contact**, **cultural change** and **migration**.
- **Imperfect learning**.
- **Social prestige**: the loss of rhoticity in Received Pronunciation is the page's example.

> **WATCH OUT: CHANGE IS NOT DECAY**
>
> It is tempting to say that *gonna* is lazy speech and *going to* is correct. The source says innovations cannot be called good or bad scientifically, and *nice* has changed its meaning more than once without anyone's permission. The linguist's question is how and why, not whether.

## Is Change Regular?

The Neogrammarians held that sound change is "regular", affecting all relevant words at once. Wikipedia's article says this is controversial, though useful as a heuristic. Chapters 2 and 3 develop the idea, because the whole comparative method depends on sound change being regular enough to track.

## What Historical Linguistics Asks

- How and why do languages change?
- Which languages are related, and how are they grouped into families?

- What did ancestral (proto-) languages look like?
- How did a language's structures develop from earlier forms?
- Where do words come from, and how have their forms and meanings changed (etymology)?
- How do social and cultural factors shape language development?

The main methods are the **comparative method** (comparing related languages to infer shared ancestry and reconstruct proto-languages) and **internal reconstruction** (inferring earlier forms from patterns within one language). Modern historical linguistics dates from the late 18th century, grew out of philology, and began with Indo-European and Uralic languages.

## Conventions Used in This Course

```
// Conventions (my summary; Chapter 4 attributes the asterisk to a source)
*word      a reconstructed form: nobody recorded it
A > B      A became B over time
B < A      B comes from A
[ ]  /  /  phonetic and phonemic notation, as in Phonetics and the IPA
(tree)     a family tree shown as an ASCII tree
```

**NOTE ON SOURCES**

The definitions, the causes, the semantic-change cases and the questions come from Wikipedia's articles on language change, historical linguistics and semantic change (the historical-linguistics page was read only up to its first 100,000 characters). That page does not mention the asterisk convention or William Jones, so the convention box above is my own summary and no history beyond the late 18th century is claimed. Two cases I might have used (*nice* as "silly", Great Vowel Shift) were not on the pages and are left out.

## Try It Yourself

**EXERCISE 1**

Classify each as sound, semantic or syntactic change: *awful* from "awe-inspiring" to "very bad"; RP losing the /r/ in *car*; *broadcast* from seeds to signals.

View solution

**EXERCISE 2**

Explain *sing / sang / sung* synchronically and diachronically. What would each viewpoint say?

View solution

**EXERCISE 3**

Match each change to a cause from the list: *going to* to *gonna*; the loss of the /r/ in RP; the shift of *terrific* to "very good".

View solution

### CHAPTER 1 QUICK REFERENCE

- **Synchronic** = one moment; **diachronic** = through time
- Uniformitarian principle: today's processes were at work in the past
- Kinds of change: phonetic, lexical, spelling, semantic, syntactic
- Causes: economy, expressiveness, analogy, contact, migration, imperfect learning, prestige
- Change is not decay; sound change is regular enough to use as a heuristic

**COMING NEXT**

*Historical Linguistics and Language Families 2* looks at sound change in detail, with Grimm's law as the standard example.

# Historical Linguistics and Language Families

Course 1 · Chapter 2 · Sound Change

Chapter 1 listed the kinds of language change. Sound change is the kind with the strongest claim to being *regular*, which is why it is the foundation of everything else in this course. This chapter names the types, sets out what "regular" means, and then works through the most famous sound change in the field.

> **THE IDEA IN A BOX**
>
> A **sound change** is a change in a language's pronunciation. It may replace one sound with another (phonetic change) or alter the sound system more broadly, for example by merging two sounds (phonological change). Because it is regular, we can use it to prove relationships between languages.

## The Types of Sound Change

Wikipedia's article on sound change lists these types, and every one connects to something from Phonetics and the IPA.

| TYPE | WHAT HAPPENS | EXAMPLE FROM THE SOURCE |
|---|---|---|
| **Assimilation** | one sound becomes more like a neighbour | Latin *kom-* becomes *con-* before t, d or n (*contactus*, *condere*) |
| **Dissimilation** | one sound becomes less like another | Old Spanish *omne* to *hombre* |
| **Lenition** | a consonant weakens, e.g. a stop becomes a fricative | Italian intervocalic *b* to *v*: *caballum* to *cavallo* |
| **Fortition** | a consonant strengthens (the opposite of lenition) | (none given) |
| **Deletion** | a sound is lost; aphaeresis (initial), syncope (medial), apocope (final) | Old French *estat* to *état* (syncope); *opossum* to *possum* (elision) |
| **Haplology** | a syllable next to a similar one is removed | Old English *Englaland* to *England* |
| **Epenthesis** | a sound is inserted | Latin *humilis* to English *humble* |
| **Prothesis** | a sound is added at the start of a word | Latin *status* to Spanish *estado* |

| TYPE | WHAT HAPPENS | EXAMPLE FROM THE SOURCE |
|---|---|---|
| Metathesis | two sounds switch places | Old English *thridda* to Middle English *third* |
| Merger | two sounds become one (a phonological change) | (none given) |

The article also notes nasalisation: French words ending in *-in* now have the nasal vowel [ɛ̃]. It does not define "split" as a separate type, so this course does not either.

> **WATCH OUT: LENITION IS NOT ONE FIXED LIST**
>
> The source defines lenition simply as a consonant weakening, with a stop becoming a fricative as one example. Do not read that as a rule that all weakening goes in a fixed direction; judge each case from the sounds themselves (Phonetics and the IPA 3 gives you the manner scale).

## Conditioned and Unconditioned Change

An **unconditioned** change applies in all environments. A **conditioned** change applies only in certain contexts, and linguists write it as *A > B / X __ Y*: A becomes B between X and Y. The source's example is the Italian change of intervocalic *b* to *v*, with V__V standing for "between vowels". Sandhi, the sound change at word boundaries, is conditioned change too (Phonetics and the IPA 8 showed several cases).

```
// A > B / X __ Y   ("A becomes B between X and Y")
b > v / V __ V      Italian: caballum > cavallo
```

## The Neogrammarian Idea: Regularity

The Neogrammarians treated sound change as regular and exceptionless, expressed as **sound laws** (the source names Grimm's law and Grassmann's law). Wikipedia lists four heuristics: sound change has no memory, ignores grammar, is exceptionless, and is inevitable. It treats these as guidelines, not inviolable rules.

- Apparent exceptions can come from analogy, dialect borrowing or other causes.

- **Lexical diffusion** challenges exceptionlessness: a change may spread word by word, though it often eventually reaches the whole lexicon. Changes that affect only a few words are called sporadic.
- Each change is limited in space and time.

## Grimm's Law

Grimm's law (also called the First Germanic Consonant Shift) is a set of sound laws describing how Proto-Indo-European (PIE) stop consonants developed in Proto-Germanic during the first millennium BCE. The reconstructed forms are marked with an asterisk; the dedicated discussion of the asterisk comes in Chapter 4.

Wikipedia describes it as a chain shift with three phases, applied in the order 3, 2, 1:

| PHASE | PIE | PROTO-GERMANIC | CHANGES |
|---|---|---|---|
| 1 | voiceless stops | voiceless fricatives | *p > f, *t > þ, *k > h, *kʷ > hw |
| 2 | voiced stops | voiceless stops | *b > p, *d > t, *g > k, *gʷ > kʷ |
| 3 | voiced aspirated stops | voiced stops or fricatives | *bʰ > b, *dʰ > d, *gʰ > g, *gʷʰ > gʷ |

The letter þ ("thorn") is the Old English and Germanic letter for the *th* sound; reading it as [θ] is my gloss, not the page's. The order matters because the sounds move into places that were just vacated: voiced stops become voiceless stops only after the old voiceless stops have already left to become fricatives.

### Examples from the page

| PIE | MEANING | LATIN | ENGLISH |
|---|---|---|---|
| ***pṓds** | foot | *pēs, pedis* | *foot* |
| ***tréyes** | three | *trēs* | *three* |
| ***ḱwón-** | dog | *canis* | *hound* |
| ***déḱm̥t** | ten | *decem* | *ten* |

Latin did not undergo Grimm's law, so it keeps the old sounds; English did, so it shows the new ones. Compare the first consonants: Latin *p* against English *f* (*pedis* and *foot*), Latin *t* against English *th* (*trēs* and *three*), Latin *c* [k] against English *h* (*canis* and *hound*), and Latin *d* against English *t* (*decem* and *ten*).

### The people behind it

- Friedrich von Schlegel noted the Latin *p* / Germanic *f* correspondence in 1806.
- Rasmus Rask extended it to Sanskrit, Greek and the full range of consonants in 1818.
- Jacob Grimm stated the rule in *Deutsche Grammatik* in 1822 and extended it to standard German. He noticed exceptions that his law did not predict.
- Karl Verner, a Danish linguist, explained those exceptions with **Verner's law**, which voiced certain Germanic voiceless fricatives under specific conditions.

Verner's law is a good example of the regularity principle at work. The exceptions to a sound law were not accepted as random; they were assumed to have a cause, and the cause was found.

**NOTE ON SOURCES**

The sound-change types, the conditioned/unconditioned distinction, the regularity heuristics, the three phases of Grimm's law, the four examples and the history come from Wikipedia's articles on sound change and Grimm's law. The page's English *father* example (PIE **ph₂tḗr* to Proto-Germanic **fadēr*) is left out because its *d* needs Verner's law, which this chapter only mentions. The Latin *pater* and *cornu* pairs I might have used are not on the page and are left out. The [θ] reading of þ and the "vacated places" remark are mine.

## Try It Yourself

**EXERCISE 1**

For each pair, name the Grimm's law phase and the sound change: *pedis* / *foot*, *trēs* / *three*, *canis* / *hound*, *decem* / *ten*.

View solution

**EXERCISE 2**

Name the type of sound change in each: Latin *kom-* to *con-*; Old English *thridda* to *third*; Latin *humilis* to *humble*; Old English *Englaland* to *England*; Latin *status* to Spanish *estado*.

View solution

**EXERCISE 3**

Write the Italian change *caballum* to *cavallo* as a conditioned rule. Then explain how Verner's explanation of Grimm's exceptions supports the Neogrammarian idea.

View solution

### CHAPTER 2 QUICK REFERENCE

- Types: assimilation, dissimilation, lenition, fortition, deletion, haplology, epenthesis, prothesis, metathesis, merger
- Conditioned change is written A > B / X __ Y; unconditioned applies everywhere
- Neogrammarians: sound change is regular; exceptions have causes (analogy, borrowing, others); lexical diffusion is the challenge
- Grimm's law: voiceless stops > fricatives, voiced stops > voiceless stops, voiced aspirates > voiced stops or fricatives
- Verner's law explained Grimm's exceptions

**COMING NEXT**

*Historical Linguistics and Language Families 3* turns regular sound correspondences into proof of relationship: the comparative method.

# Historical Linguistics and Language Families

Course 1 · Chapter 3 · The Comparative Method

Chapter 2 showed that sound change is regular. This chapter uses that fact. If sounds change regularly, then two languages that descend from one ancestor will show regular matches between their words, and nothing else produces a pattern like that.

> **THE IDEA IN A BOX**
>
> The **comparative method** compares two or more related languages feature by feature, then infers the properties of their shared ancestor. It looks for systematic sound and meaning correspondences. If they cannot be explained by universals, contact or chance, and are numerous and regular, the languages are assumed to descend from one proto-language (Wikipedia, Comparative method).

## Cognates

**Cognates** are words in different languages that have been "inherited in direct descent from an etymological ancestor" in a shared parent language (Wikipedia, Cognate). Their sound and meaning can change so much that the relationship needs careful historical study to confirm.

| WORD | LANGUAGE |
|---|---|
| *night* | English |
| *Nacht* | German |
| *nox* | Latin |
| *nuit* | French |

All four come from Proto-Indo-European **nókʷts*, according to the source. English *father*, French *père* and Armenian *hayr* also descend from a common ancestor, even though they look quite different.

### Look-alikes that are not cognates

- **False cognates:** Latin *habēre* and German *haben* both mean "to have" but come from different roots. English *much* and Spanish *mucho* are similar but not cognates.

- **Loanwords:** English *beef* was borrowed from Old French *boef*.
- **Chance:** English *day* and Latin *dies* match in meaning and sound, but that is chance.
- **Borrowing that mimics kinship:** English *taboo* resembles Polynesian forms because it was borrowed from Tongan.
- **Heavy borrowing is not kinship:** Persian's heavy Arabic borrowing does not make it related to Arabic.

> **WATCH OUT: SIMILARITY ALONE PROVES NOTHING**
>
> The source says similarity alone has no value. What counts is *regular* correspondence. English *t-* regularly matches Latin *d-*: *ten* / *decem*, *two* / *duo*, *tongue* / *dingua* (the source's own spelling of the old Latin form). That regularity is what Grimm's law (Chapter 2) predicts.

## The Steps

Wikipedia says there is no fixed procedure. Its outline, based on Campbell and Crowley, has five steps:

| STEP | WHAT YOU DO |
|---|---|
| 1 | Assemble potential cognate lists of basic words with similar meanings |
| 2 | Establish **correspondence sets**: the regular sound matches across languages |
| 3 | Find complementary distribution: sets that occur in different contexts may reflect one proto-phoneme, as with conditioned changes (Chapter 2) |
| 4 | Reconstruct proto-phonemes, using typology and the principle of economy (the fewest changes) |
| 5 | Check the system typologically against known constraints, such as symmetry in inventories |

## Worked Example 1: Polynesian

The page compares the words for "one", "two" and "three":

| LANGUAGE | ONE | TWO | THREE |
|---|---|---|---|
| Tongan | taha | ua | tolu |

| LANGUAGE | ONE | TWO | THREE |
|---|---|---|---|
| Samoan | tasi | lua | tolu |
| Māori | tahi | rua | toru |
| Hawaiian | kahi | lua | kolu |

Hawaiian *k* regularly corresponds to *t* elsewhere (*kahi* / *tahi*, *kai* / *tai*). That is a correspondence set: a regular sound match, not a coincidence.

## Worked Example 2: Romance

| LANGUAGE | BODY | RAW | CHAIN | TO HUNT |
|---|---|---|---|---|
| Italian | corpo | crudo | catena | cacciare |
| Spanish | cuerpo | crudo | cadena | cazar |
| Portuguese | corpo | cru | cadeia | caçar |
| French | corps | cru | chaîne | chasser |

The sets are *k* : *k* and *k* : [ʃ]. French [ʃ] appears only before *a*, so the sets are complementary (step 3) and reflect a single proto-phoneme, **k* (spelled *c* in Latin). That is the logic of conditioned sound change from Chapter 2, run backwards.

## Shared Innovations

Two languages can share a feature for two reasons, and only one of them shows a close relationship.

- **Shared innovations** indicate a closer subgroup. The source's example is Grimm's law in English and German.
- **Shared retentions** do not. German and Russian both keep the dative / accusative contrast; that does not group them together, because it simply shows the loss happened later in English.

## Limits of the Method

- **Time depth:** the page gives no fixed age limit. It says reconstruction can be only partial when languages are scarcely attested, the time depth is too great, or internal evolution obscures sound laws. Relatedness is then "plausible, but uncertain".
- **Contact:** areal diffusion and Sprachbunds can create false classifications (Chapter 6).
- **Irregular change:** random mutations, analogy (Russian *dʲevʲatʲ* "nine", influenced by "ten") and gradual lexical diffusion can obscure regularity.
- **Lost features:** Latin declension was lost in Romance, so it cannot be fully reconstructed from comparison.
- **Certainty:** reconstructions of unknown proto-languages are "only likely, not certain".

**NOTE ON SOURCES**

Everything here comes from Wikipedia's articles on the comparative method (read to its first 100,000 characters) and on cognates. The Cognate page gives no entry for *bad*, *theos/deus* or Latin *pater*, so they are left out. The page names no fixed age limit for the method, so none is given here. The "run backwards" remark is mine.

## Try It Yourself

**EXERCISE 1**

For each pair, say whether the source treats it as a regular correspondence, chance, borrowing or false cognate: *ten* / *decem*; *day* / *dies*; *taboo* / Polynesian; *much* / *mucho*.

View solution

**EXERCISE 2**

In the Polynesian table, find the correspondence among Samoan, Māori and Hawaiian for the consonant in *lua*, *rua*, *lua*, and in *tolu*, *toru*, *kolu*. Write the set for each position.

View solution

**EXERCISE 3**

Using the Romance table, explain why the sets *k* : *k* and *k* : [ʃ] are treated as one proto-phoneme, and name the step of the method this shows.

View solution

## CHAPTER 3 QUICK REFERENCE

- **Cognates** are inherited from a common ancestor; loanwords, chance matches and false cognates are not
- Similarity alone proves nothing; **regular correspondence** does
- Steps: cognate lists, correspondence sets, complementary distribution, reconstruct, check typologically
- **Shared innovations** show subgroups; shared retentions do not
- Limits: great time depth, contact, irregular change, lost features

**COMING NEXT**

*Historical Linguistics and Language Families 4* takes the correspondence sets and builds a proto-language from them: reconstruction.

# Historical Linguistics and Language Families

Course 1 · Chapter 4 · Reconstruction

Chapter 3 found the regular correspondences between related languages. Reconstruction takes the next step: it asks what single ancestor sound would, after regular change, give each of the sounds seen today. The result is a form in a language that nobody has recorded. This chapter explains how that is done and how far to trust it.

> **THE IDEA IN A BOX**
>
> **Linguistic reconstruction** establishes features of an "unattested ancestor language" of one or more given languages. The result of comparative reconstruction is a **proto-language**, the common ancestor of a language family (Wikipedia, Linguistic reconstruction).

## The Asterisk

Reconstructed forms are prefixed with an asterisk (*) to distinguish them from attested forms. "Asterisks are used by linguists as a conventional mark of reconstructed words" (Wikipedia, Proto-Indo-European language). So **tréyes* is not a word anyone wrote down; it is a conclusion. Chapter 1 flagged this convention as my own summary; this chapter attributes it to its sources.

> **WATCH OUT: AN ASTERISK FORM IS A HYPOTHESIS**
>
> Never quote a starred form as if it were attested. When two sources give slightly different starred forms for the same word, that is normal: the Grimm's law page in Chapter 2 gave the PIE word for "dog" as **ḱwón-*, while the Proto-Indo-European page gives **ḱwn̥tós* ("hound"). I have not worked out why the two differ, so this course reports both as given.

## Two Kinds of Reconstruction

| KIND | EVIDENCE |
| --- | --- |
| Internal reconstruction | irregularities within a single language, used to infer an earlier stage of that language |

| KIND | EVIDENCE |
| --- | --- |
| Comparative reconstruction | the comparative method (Chapter 3) applied to two or more related languages of one family |

A **reflex** is a known derivative of an earlier form, attested or reconstructed. An attested word from which a proto-language root is reconstructed is a reflex. A reflex predictable from the reconstructed history is "regular", and reflexes of the same source are cognates.

## Rules of Thumb

| PRINCIPLE | WHAT IT SAYS |
| --- | --- |
| Simplicity | posit the fewest phonemes that fit the data, and prefer the sound values requiring the fewest changes |
| Majority | a pattern shared by most cognates is likely inherited from the parent language |
| Most natural development | four common tendencies: word-final vowels are dropped, voiceless sounds between vowels voice, stops become fricatives, word-final consonants devoice |
| Subgrouping | needs shared innovations and shared retentions; do not assume linguistic groupings match cultural or ethnic ones |

### A worked example from the source

Spanish *cantar* and French *chanter* suggest that the stop [k] is older than the fricative [ʃ], because stops tend to become fricatives. This is the "most natural development" principle in action, and it is the same direction as the Romance [k] : [ʃ] set in Chapter 3.

```
// Reconstructing the older sound
Spanish   cantar   [k]
French    chanter  [ʃ]
Principle: stops tend to become fricatives
Conclusion: [k] is older
```

## Proto-Indo-European

Proto-Indo-European (PIE) is the reconstructed common ancestor of the Indo-European language family. No direct record of it has been found; its features were derived by comparing documented descendant languages. Examples from the page:

| PIE FORM | MEANING |
|---|---|
| *wódr̥ | water |
| *ḱwn̥tós | hound |
| *tréyes | three |
| *méh₂tēr | mother |
| *ph₂tḗr | father |
| *bʰréh₂tēr | brother |
| *swésōr | sister |
| *dʰugh₂tḗr | daughter |
| *gʷṓus | cow |

The sound correspondence is the one from Chapter 2: Italian *piede*, *padre* and *pesce* against English *foot*, *father* and *fish* show a consistent initial p / f pairing. Grimm's law (1822) and Verner's law (1876) describe such regular changes, and the Neogrammarians held that sound laws apply without exception.

### When and where?

- **Date:** PIE is hypothesised to have been spoken as a single language from about 4500 BCE to 2500 BCE. Other estimates place the period more than a thousand years later.
- **Kurgan hypothesis** (first proposed by Marija Gimbutas in 1956): the most popular view. It places the language with the Yamnaya culture on the Pontic steppe around 3400 BCE.
- **Anatolian hypothesis:** PIE spread from Anatolia with agriculture, beginning around 7500-6000 BCE.
- The page calls these two the most widely accepted and the most debated against each other. Colin Renfrew, the Anatolian hypothesis's original proponent, has accepted migrations from the Pontic steppe after 2015 ancient DNA studies. The Armenian hypothesis, the Paleolithic continuity paradigm and the indigenous Aryans theory are not regarded as credible.

I show both main views because the page treats the question as open. Where scholars disagree, this course does not choose.

### A reconstruction that was later confirmed

The laryngeal theory was proposed by Ferdinand de Saussure in 1879, based only on internal reconstruction. It explains irregularities in PIE phonology as effects of hypothetical sounds. It gained general acceptance after Jerzy Kuryłowicz found consonantal reflexes of these sounds in Hittite. That is why many PIE forms above contain a subscript $h_2$.

### How reliable is it?

The page says PIE phonology "has been reconstructed in some detail", but that the most widely accepted reconstruction is "not uncontroversial". Some points remain debated, including the existence of /a/ as a separate phoneme and the accent system, which some linguists reconstruct as a tone system instead. It also says the sound laws are reliable enough to support the Neogrammarian hypothesis.

**NOTE ON SOURCES**

The definitions, principles, worked example, PIE forms, dates, homeland views and the laryngeal theory come from Wikipedia's articles on linguistic reconstruction and Proto-Indo-European language (the PIE page was read only up to its first 100,000 characters; its sections on morphology and numerals were not reached, so there are no PIE numerals or verbs here). The reconstruction article does not discuss laryngeals, reliability or the debate over whether a proto-language is a real language, so none of those is attributed to it. The remark that $h_2$ is the laryngeal in these forms is my connection between the two pages.

## Try It Yourself

**EXERCISE 1**

Use the "most natural development" principle to decide which is older in each pair: Spanish *cantar* [k] / French *chanter* [ʃ]; and a voiceless stop between vowels versus a voiced one in a related language.

View solution

**EXERCISE 2**

Why is Saussure's 1879 laryngeal theory called internal reconstruction, and why did Hittite matter to it?

View solution

**EXERCISE 3**

Pair each starred form with its meaning, and state in one sentence why it is written with an asterisk: **ph₂tḗr*, **wódr̥*, **gʷṓus*. Then give the two main homeland views and one date for each.

View solution

## CHAPTER 4 QUICK REFERENCE

- Asterisk = reconstructed, never attested; a hypothesis, not a fact
- Internal reconstruction (one language) versus comparative reconstruction (a family)
- Principles: simplicity, majority, most natural development, subgrouping
- PIE: about 4500-2500 BCE; Kurgan (steppe, c. 3400 BCE) versus Anatolian (7500-6000 BCE)
- Laryngeal theory: Saussure 1879, confirmed through Hittite

**COMING NEXT**

*Historical Linguistics and Language Families 5* zooms out from one proto-language to the families of the world, with the Indo-European family in detail.

# Historical Linguistics and Language Families

Course 1 · Chapter 5 · Language Families

Chapters 3 and 4 showed how to prove that languages are related and how to reconstruct their ancestor. This chapter gives the result a name and a map: a **language family** is a set of languages with a common ancestor. Then it looks at the best-studied one, Indo-European, which contains English.

**THE IDEA IN A BOX**

A **language family** is "a group of languages related through descent from a common ancestor", the **proto-language** (Wikipedia, Language family). The proto-language is seldom known directly and is usually reconstructed with the comparative method. Its descendants are **daughter languages**.

## Key Terms

| TERM | MEANING | EXAMPLE FROM THE SOURCE |
|---|---|---|
| Proto-language | the common ancestor of a family | Proto-Indo-European |
| Daughter languages | languages descended from a proto-language; "genetically related" | the Indo-European languages |
| Branch / subfamily | a smaller group within a family that shares a more recent common ancestor | Germanic, within Indo-European |
| Language isolate | a language with no proven relatives, so a family of one | Basque; about 129 isolates today |
| Dialect continuum | closely related varieties with no clear geographic boundaries between them, so individual languages cannot be clearly identified or counted | Arabic, where speakers at the extremes have no mutual intelligibility |

**WATCH OUT: "GENETIC" MEANS DESCENT, NOT BIOLOGY**

The source calls daughter languages "genetically related", but the word refers only to descent from a common ancestor language. It says nothing about the speakers' biology, and Chapter 4's subgrouping principle warned against assuming that linguistic groupings match cultural or ethnic ones.

## How Many Families?

The answer depends on who is counting, and the source gives three figures:

| SOURCE | COUNT |
|---|---|
| Ethnologue | 7,151 living languages across 142 families |
| Lyle Campbell (2019) | 406 independent families, including isolates |
| Glottolog | 423 families, including 184 isolates |

The gap between 142 and 406 or 423 is not an error: the counts follow different methods, and the source does not say how they differ, so this course reports all three. Ethnologue's figure of 142 is the smallest, which suggests it groups more languages together.

### The largest families

| ETHNOLOGUE 27 (2024) | LANGUAGES | GLOTTOLOG 5.3 (2026) | LANGUAGES |
|---|---|---|---|
| Niger-Congo | 1,552 | Atlantic-Congo | 1,408 |
| Austronesian | 1,256 | Austronesian | 1,276 |
| Trans-New Guinea | 481 | Indo-European | 586 |
| Sino-Tibetan | 458 | Sino-Tibetan | 519 |
| Indo-European | 454 | Afroasiatic | 381 |

By speakers, Indo-European, Sino-Tibetan, Afro-Asiatic, Niger-Congo and Austronesian together account for almost 83.3% of the world's population. Notice that Indo-European is fifth by number of languages in one count and third in the other, yet first by speakers: a family can be large in people and modest in languages.

## Indo-European

Wikipedia's article on the Indo-European languages gives these figures: over 3.4 billion people, 42% of the global population, speak an Indo-European language as a first language. About 446 living Indo-European languages exist according to Ethnologue, 313 of them in Indo-Iranian. (The

family-level table above gives Ethnologue's later count as 454; the two pages are from different editions, and this course reports both as given.)

### The branches

| BRANCH | STATUS | REPRESENTATIVE LANGUAGES | NATIVE SPEAKERS |
|---|---|---|---|
| Indo-Iranian | living | Hindustani (Hindi, Urdu), Bengali, Punjabi, Marathi, Persian | hundreds of millions |
| Italic | living | the Romance languages, including Spanish, Portuguese, French | hundreds of millions |
| Germanic | living | English, German, Dutch | hundreds of millions |
| Balto-Slavic | living | Russian, Polish, Lithuanian, Latvian | hundreds of millions |
| Hellenic | living | Greek | fewer than 15 million |
| Albanoid | living | Albanian | fewer than 15 million |
| Armenian | living | Armenian | fewer than 15 million |
| Celtic | living | Welsh, Irish, Scottish Gaelic, Breton | fewer than 15 million |
| Anatolian | extinct | Hittite, Luwian | none |
| Tocharian | extinct | (not named in the summary I read) | none |

```
Proto-Indo-European
├── Anatolian        (extinct: Hittite, Luwian)
├── Tocharian        (extinct)
├── Indo-Iranian     (Hindi, Urdu, Bengali, Persian ...)
├── Hellenic         (Greek)
├── Albanoid         (Albanian)
├── Armenian         (Armenian)
├── Italic           (Spanish, Portuguese, French ...)
├── Celtic           (Welsh, Irish, Scottish Gaelic, Breton)
├── Germanic         (English, German, Dutch)
└── Balto-Slavic     (Russian, Polish, Lithuanian, Latvian)
```

The tree lists the ten branches as the source table gives them, not in order of splitting: the page's text I read does not give a branching order, and Chapter 6 explains why a clean tree can mislead.

### The discovery

Sir William Jones lectured to the Asiatic Society of Bengal in 1786 on similarities among Latin, Greek and Sanskrit. He conjectured an earlier ancestor language, calling it "a common source, which, perhaps, no longer exists." That sentence is the idea behind Chapters 3 and 4 in a single line.

Look at where the languages you have met in this course sit: English and German are Germanic; Spanish, Portuguese and French are Italic (Romance) and appeared in Chapter 3's cognate tables; Hittite is Anatolian and appeared in Chapter 4's laryngeal story. Hungarian is not on this tree at all; it belongs to a different family, which Chapter 9 introduces.

**NOTE ON SOURCES**

The definitions, the family counts, the largest-families table and the Indo-European figures and branch table come from Wikipedia's articles on language family and the Indo-European languages. Both pages were read only up to their first 100,000 characters, so the Tocharian languages, branch subdivisions and any branching order were not seen. The ASCII tree, the "large in people, modest in languages" remark, the suggestion that Ethnologue groups more languages together, and the placing of Hungarian outside Indo-European are mine; the Hungarian point is a promise for Chapter 9, which will cite a source.

## Try It Yourself

**EXERCISE 1**

Define proto-language, daughter language, branch and isolate in one line each, and give the source's example for each.

View solution

**EXERCISE 2**

Place each language in its Indo-European branch: English, Persian, Greek, Hittite, Welsh, Russian, Albanian, French. Which are extinct?

View solution

**EXERCISE 3**

The three family counts (142, 406, 423) differ widely. Give two reasons the source or this chapter suggests, and say what you would need to find out before trusting one.

View solution

## CHAPTER 5 QUICK REFERENCE

- Family = descent from a common proto-language; branch = more recent common ancestor; isolate = family of one (Basque)
- Counts differ: Ethnologue 142 families; Campbell 406; Glottolog 423
- Largest by speakers: Indo-European, Sino-Tibetan, Afro-Asiatic, Niger-Congo, Austronesian (about 83.3% of people)
- Indo-European: 10 branches (8 living, 2 extinct), over 3.4 billion first-language speakers (42%)
- Jones 1786: "a common source, which, perhaps, no longer exists"

**COMING NEXT**

*Historical Linguistics and Language Families 6* looks at the limits of the tree: the wave model, language contact and areas where unrelated languages share features.

# Historical Linguistics and Language Families

Course 1 · Chapter 6 · Trees and Waves

Chapter 5 drew Indo-European as a tree. A tree is a good picture of one thing: a language splits into daughters and they drift apart. But languages also meet. Words are borrowed, features spread between neighbours, and unrelated languages can end up looking alike. This chapter looks at what the tree leaves out.

> **THE IDEA IN A BOX**
>
> The **tree model** assumes languages change only through social splitting and divergence. The **wave model** pictures an innovation that starts in one area and spreads speaker to speaker and dialect to dialect, like waves on water. A full account of language history needs both.

## The Tree and Its Problem

Wikipedia's article on the wave model says the tree model "struggles to explain some shared features, especially in the Germanic languages". A tree can show that two branches share a recent ancestor; it cannot easily show that a change spread across the boundary between branches after they had separated.

| MODEL | PICTURE | GOOD FOR |
|---|---|---|
| Tree | a line that forks and never rejoins | descent, subgroups (Chapters 3 to 5) |
| Wave | an innovation spreading outward, with waves overlapping in intersecting patterns | dialect continua, areal phenomena, features shared across branches |

### How the wave model works

- An innovation starts in one area and spreads speaker to speaker and dialect to dialect. Innovations can overlap in intersecting patterns, which the tree cannot easily represent.
- Dialectology uses linguistic atlases and maps to show **isoglosses** (the boundaries of a feature) and dialect boundaries, including fuzzy ones such as the Croissant and the Rhenish fan (the page's own examples).

- Schmidt's continuum metaphor: a smooth slope becomes a stepped line, with each step a dialect. Some steps fade, while others, like Standard German, can spread and replace local dialects.
- Advocacy for the model is credited to Johannes Schmidt and Hugo Schuchardt.

> **WATCH OUT: IT IS NOT TREE VERSUS WAVE**
>
> The wave model has "contributed greatly to improve, but not supersede" the tree model. Some scholars argue it should replace the tree for genealogy, and the page says **linkages** (families descended from a former continuum) must be analysed with waves. It was little accepted in the 20th century except for dialect continua and areal phenomena, and has regained popularity because of the tree's shortcomings. Use the tree for descent and the waves for spread.

## Language Contact

Wikipedia's article on language contact lists what can happen when speakers of different languages meet:

| OUTCOME | WHAT IT IS | EXAMPLE FROM THE SOURCE |
|---|---|---|
| **Borrowing** | "The most common way that languages influence each other is the exchange of words"; influence can reach morphology and grammar | Newar absorbed noun inflection and verb tenses from neighbouring Indo-Iranian languages |
| **Code-switching** | multilingual speakers use multiple languages in a single conversation | (none given) |
| **Pidgin, creole** | a pidgin forms among people without a common language and may become a creole; some linguists deny a creole must pass through a pidgin stage | Ndyuka (Aukan) and Saramaccan, in Suriname, draw their vocabulary mainly from Portuguese, English and Dutch |
| **Mixed language** | forms among communities fluent in both source languages; keeps much of the parents' grammatical and phonological complexity | (none given) |
| **Language shift and death** | one language replaces another, most often the one with higher prestige; can lead to endangerment or extinction | (none given) |

### Loanword cases

- **Latin and French into English**: the page describes the large-scale import of words from these languages in the 16th and 17th centuries as more significant than modern English borrowing into other languages.
- **Iranian into Armenian**: Armenian borrowed so many Iranian words that it was at first considered a divergent Indo-Iranian branch. This is a warning for Chapter 3: heavy borrowing can fake relationship, so correspondence sets must come from basic vocabulary. Armenian is, in the Indo-European table of Chapter 5, a branch of its own.
- **English into Hindi and Scots**, **Afrikaans into South African English**, and **Chinese into Japanese**, where some Japanese coinages based on Chinese forms were later reborrowed into Chinese.

During much of the Middle Ages, upper-class English speech was "dramatically influenced" by Norman, the page says, to the point that it often resembled a dialect. It also gives English phrases adapted from French where the adjective follows the noun: *court-martial*, *attorney-general*, *Lake Superior*.

## Linguistic Areas (Sprachbunds)

A **sprachbund** (linguistic area) is a group of languages that share "areal features" from geographic proximity and language contact. The languages may be genetically unrelated or only distantly related. A sprachbund is defined by features shared for reasons other than common descent. Without the history of a region, it can be hard to tell whether shared features indicate a family or a sprachbund. The page says these shared features "might give a false appearance of relatedness."

| AREA | LANGUAGES | SHARED FEATURES |
|---|---|---|
| Balkans | Albanian, Romanian, southern South Slavic languages, Greek, Balkan Turkish, Romani (Turkish is Turkic; the rest are Indo-European) | avoiding the infinitive; forming the future tense in similar ways |
| South Asia | Dravidian and Indo-Aryan, with Munda and Sino-Tibetan also cited (Emeneau, 1956) | retroflex consonants, echo words, SOV word order, discourse markers, the quotative |

| AREA | LANGUAGES | SHARED FEATURES |
|---|---|---|
| Mainland Southeast Asia | Sinitic, Hmong-Mien, Tai, Vietnamese, with Khmer, Cham, Lao | tone systems, similar vowel systems, isolating structure, classifiers, OV order, topic-comment structure; the modern consensus places these languages in unrelated families |
| Standard Average European | modern Indo-European languages of Europe (Whorf, 1939); Romance and West Germanic at the core, North Germanic and Balto-Slavic more peripheral | (the page says Proto-Indo-European lacked most of these features, so inheritance is ruled out) |

The page also lists Mesoamerica as a linguistic area of the Americas, but gives no specific features. It attributes these similarities to diffusion during sustained contact and to convergence caused by language contact.

**NOTE ON SOURCES**

The wave-model, language-contact and linguistic-area material comes from Wikipedia's articles on the wave model, language contact and linguistic area. The wave-model page does not give an 1872 date for Schmidt or mention Schleicher, and the contact page does not mention Norse or give a borrowing hierarchy, so none of those appears here. I did not read these pages beyond the summaries returned. The Armenian-versus-Chapter-3 remark and the closing "tree for descent, waves for spread" are mine.

## Try It Yourself

**EXERCISE 1**

In one sentence each, say what the tree model explains and what it does not, and what the wave model adds. Use the Germanic problem from the source.

View solution

**EXERCISE 2**

Greek, Romanian and Albanian avoid the infinitive, but so does Balkan Turkish, which is not Indo-European. Explain why the shared feature does not prove they are related, using the Balkans entry.

View solution

**EXERCISE 3**

Classify each as borrowing, pidgin or creole, mixed language, or language shift: Armenian taking Iranian words; Ndyuka and Saramaccan; Newar taking verb tenses from neighbours; English phrases such as *attorney-general*.

View solution

## CHAPTER 6 QUICK REFERENCE

- Tree = descent by splitting; wave = spread of an innovation; use both
- Isoglosses map the boundary of a feature; linkages need the wave model
- Contact outcomes: borrowing, code-switching, pidgin and creole, mixed language, shift and death
- Heavy borrowing can fake relatedness (Armenian and Iranian); only basic correspondence sets prove kinship
- A sprachbund shares features through contact, not descent (Balkans, South Asia, Southeast Asia)

**COMING NEXT**

*Historical Linguistics and Language Families 7* turns from sounds and families to meaning and grammar: semantic change, analogy and grammaticalisation.

# Historical Linguistics and Language Families

Course 1 · Chapter 7 · Meaning and Grammar Change

Chapters 2 to 4 followed sounds. Words and grammar change too, and the same logic applies: the changes are not random, they fall into recognisable types, and a pattern seen today is evidence for how patterns formed before. This chapter covers three of them: change of meaning, change by analogy, and the way words become grammar.

> **THE IDEA IN A BOX**
>
> **Meaning changes** by recurring routes (broadening, narrowing, metaphor and others). **Analogy** remodels irregular forms on the pattern of regular ones. **Grammaticalisation** turns content words into grammar, along a path called a cline.

## Semantic Change

Chapter 1 gave a first table. Wikipedia's article on semantic change organises the types using Bloomfield's 1933 scheme and adds more:

| TYPE | WHAT HAPPENS | EXAMPLE FROM THE SOURCE |
|---|---|---|
| Widening (broadening) | a word's sense expands | brand names used generically, such as *Kleenex* |
| Narrowing | from a general to a specific sense | *skyline*: any horizon, then one lined with skyscrapers |
| Amelioration | a shift to a more positive sense | *terrific*: "terrifying" to "astonishing" to "very good"; *nice*: "agreeable, delightful" (1769), then "kind, thoughtful" (1830) |
| Pejoration | a shift to a more negative sense | *awful*: "awe-inspiring" to "very bad"; *egregious*: remarkably good to "remarkably bad or flagrant" |
| Metaphor | change by similarity | *broadcast*: casting seeds to transmitting signals |
| Metonymy | change by nearness | *jaw*: "cheek" to "mandible" |
| Synecdoche | change by part-whole relation | capital cities standing for governments |

| TYPE | WHAT HAPPENS | EXAMPLE FROM THE SOURCE |
|---|---|---|
| Hyperbole (strengthening) | a word becomes stronger | *kill*: "torment" to "slaughter" |
| Meiosis (weakening) | a word becomes weaker | *astound*: "strike with thunder" to "surprise strongly" |

> **WATCH OUT: DATES AND DIRECTIONS COME FROM THE SOURCE, NOT FROM YOU**
>
> *Nice* has several dated senses on the source page (pre-1300, late 14th century, about 1400, the 1500s, 1769, 1830). When you describe a change, cite the dated sense you mean. The source also gives *gay* a 1637 date for negative connotations, but this course does not use it, because the page's other dates for that word (13th, 14th century, 1897, 1935, 1951) need reading in context.

## Analogy

Wikipedia defines analogical change as a process in which word-forms "perceived as breaking rules or ignoring general patterns" are altered "to more typical forms that follow them."

- **Morphological levelling.** The verb *help* once had the past tense *holp* and past participle *holpen*. These were replaced by *helped*, through the analogy of other past-tense forms with *-ed* (*jumped*, *carried*, *defeated*).
- **Analogy that breaks rules.** Analogy does not always regularise. American English *dove* (past of *dive*) was formed on the pattern of *drove* (from *drive*) and *strove* (from *strive*).
- **New words by analogy.** *Software* (from *hardware*), then *firmware* and *vapourware* after it, and the humorous *underwhelm* (from *overwhelm*).

The general form of an analogy is "A is to B as C is to D." The page does not list *brethren*, *kine* or the extension of plural *-s* in its linguistics section, so they are not used here.

```
// Analogy: A is to B as C is to D
jump : jumped  ::  help : helped    (replacing holp)
drive : drove  ::  dive : dove      (American English)
```

## Grammaticalisation

**Grammaticalisation** is the process by which words shift from expressing objects or actions to serving grammatical functions. Content words such as nouns and verbs can become auxiliaries, case markers, inflections or connectives. Hopper and Traugott's **cline** runs:

```
content word  >  grammatical word  >  clitic  >  inflectional affix
```

Each step is more grammatical and less lexical than the one before it.

### Examples from the source

- **will**: Old English *willan* ("to want", "to wish") became an auxiliary for intention, and later for future time. In speech it reduces to *'ll*.
- **-ly and -like**: both suffixes derive from a Proto-Germanic word meaning "body" or "corpse".
- **going to to gonna**: an example of phonetic erosion. (You met this in Phonetics and the IPA 8 and in Chapter 1.)
- **Romance future**: Latin *cantare habeo* ("I have to sing") took on future meaning and became the future tense in most Romance languages, as in French *chanterai*.

### The mechanisms

| MECHANISM | MEANING |
|---|---|
| Semantic bleaching | loss of lexical content, leaving only grammatical content |
| Decategorialisation | loss of features tied to the original word class, such as number marking on English *that* |
| Phonetic erosion | loss of phonetic substance, though not required for grammaticalisation |
| Obligatorification | increasing obligatoriness of a structure, such as grammatical number |

### A debate: does it only go one way?

The **unidirectionality** hypothesis holds that change generally moves from less to more grammatical; Heine calls grammaticalisation "a unidirectional process". Critics argue that counterexamples exist, such as degrammaticalisation in North Saami. Lyle Campbell says advocates often minimise or redefine these cases, and Richard Janda cites over 70 works critical of the hypothesis. The source leaves the question open, and so does this course.

> NOTE ON SOURCES

Semantic change, analogy and grammaticalisation come from Wikipedia's articles on semantic change, analogy (its linguistics section; the page was read only up to its first 100,000 characters) and grammaticalization. The Grammaticalization page does not discuss *do*, the Analogy page's linguistics section lacks *brethren*, *kine* and plural *-s*, and the Semantic change page lacks *dog* and *girl*; none is used. The outline promised Hungarian examples in this chapter, but I found no source for them, so there are none. The code diagram, which is my layout of the source's examples, and the dating caution are mine.

## Try It Yourself

**EXERCISE 1**

Name the type of semantic change in each: *skyline*; *jaw*; *awful*; *kill*; *astound*; *terrific*.

View solution

**EXERCISE 2**

Show the analogy behind *helped* replacing *holp*, and explain why *dove* is an analogy that does not regularise.

View solution

**EXERCISE 3**

Place *willan* to *will* to *'ll* on the cline, and name two mechanisms at work. Then give one argument for and one against unidirectionality.

View solution

### CHAPTER 7 QUICK REFERENCE

- Semantic change: widening, narrowing, amelioration, pejoration, metaphor, metonymy, synecdoche, hyperbole, meiosis
- Analogy: odd forms pulled into line (*holp* to *helped*); it can also create irregularity (*dove*)

- Cline: content word > grammatical word > clitic > inflectional affix
- Mechanisms: semantic bleaching, decategorialisation, phonetic erosion, obligatorification
- Unidirectionality is a hypothesis, not a law; critics cite degrammaticalisation

**COMING NEXT**

*Historical Linguistics and Language Families 8* asks what evidence survives from the past: texts, place names, loanwords and dialects.

# Historical Linguistics and Language Families

Course 1 · Chapter 8 · Evidence from the Past

The comparative method works from living languages, but historians of language also have direct evidence: things people wrote, names they gave places, and words other languages borrowed from them. This chapter looks at those kinds of evidence, then follows one language, English, across its documented history to see them working together.

> **THE IDEA IN A BOX**
>
> The evidence for earlier languages comes from **texts** (studied by philology), **place names** (toponymy), and **loanwords** that preserve forms from another language. Each has limits, so the best conclusions come from agreement between several.

## Written Texts and Philology

Wikipedia defines philology as the study of language in oral and written historical sources. It overlaps with textual criticism, literary criticism, history and linguistics, and it also covers the authenticity, original form and meaning of texts. In older, especially British, usage it is conflated with comparative and historical linguistics, and in British English it remains largely synonymous with historical linguistics. Historical and comparative linguistics often rely on philological materials.

| IDEA | WHAT THE SOURCE SAYS |
|---|---|
| Textual criticism | tries to reconstruct an author's original text from variant manuscript copies; critical editions list the variants in a critical apparatus |
| New philology | some scholars reject textual criticism because it introduces editorial interpretation, and favour a strict diplomatic transcription of the manuscript |
| Decipherment | allowed scholars to read older records: Old Persian and Mycenaean Greek, and Hittite and Sumerian from cuneiform |

> **WATCH OUT: A TEXT IS NOT A RECORDING**
>
> The source lists the limits. Some scripts remain undeciphered, such as Linear A, whose language is unknown. Where texts carry political or religious weight, reaching objective conclusions is difficult.

> Textual criticism and interpretation have no clear boundary, and editorial reconstruction can distort the manuscript record. A written form is also not a pronunciation (Phonetics and the IPA 1).

## Place Names

Toponymy studies place names: their origins, meanings, usage and types. Many names have a **generic** (the kind of place) and a **specific** (what distinguishes it). In *Oldham* and *Newham* the generic is *ham* ("homestead") and the specifics are *old* and *new*. Some names are "simplex", a generic only: *Bath*, *Chester*, *Poole*.

- **Names can preserve languages that left no texts.** *Hellespont* was explained by Greek poets through Helle, but the name is probably from an older language, such as Pelasgian, that those poets did not know. George R. Stewart suggests it originally meant something like "narrow Pontus".
- **Names show settlement.** Toponyms "illustrate ethnic settlement patterns" and can help identify "discrete periods of immigration" (McDavid, 1958; Kaups's study of Finnish place names in Minnesota, 1966).
- **Names change** through language change, politics and border adjustments.

F. M. Powicke (1954) said place-name study "uses, enriches and tests the discoveries of archaeology and history and the rules of the philologists". The page presents this as a scholarly view, not a settled conclusion. The page does not discuss Celtic river names, Norse *-by* endings or Old English *-ton*, so they are not used here.

## Loanwords

Chapter 6 showed that loanwords reveal contact. Here is the evidence side: a loanword can preserve a form, or a word that later disappeared in the source language. The next section shows English loanwords from two sources.

## Worked Case: English

Wikipedia's History of English gives these periods, with approximate dates from the page:

| PERIOD | DATES ON THE PAGE | NOTES |
|---|---|---|
| Old English | from the Anglo-Saxon settlement, c. 450 | evolved into Middle English "some time after" 1066 |
| Middle English | from the Norman Conquest in 1066 "until the end of the 15th century" | |
| Early Modern English | "around 1500" | includes the language of Shakespeare |
| Modern English | often dated from the Great Vowel Shift | "in place by the late 17th century" (the page marks this as needing a citation) |

### The settlers

Germanic settlers from northwest Germany, southern Denmark and the Netherlands arrived in Britain in the mid-5th to 7th centuries. Their Ingvaeonic dialects displaced Celtic languages and possibly British Latin in much of England. The page cites the Anglo-Saxon Chronicle's account of Vortigern inviting the Angles around 449, but notes it was not a contemporaneous source.

### Old Norse

Viking raids began in the late 8th century. The Great Heathen Army invaded in 865, bringing much of northern and eastern England under Scandinavian control (the Danelaw). English borrowed about 2,000 Old Norse words, several hundred of which survive: *anger*, *bag*, *law*, *sky* and *they*. Norse contact is believed to have contributed to grammatical simplification, such as the loss of grammatical gender and most case endings.

### French and Latin

After 1066, Anglo-Norman replaced English for a time as the language of the upper classes. About 10,000 French and Norman loanwords entered Middle English, especially in government, church, law, the military, fashion and food. The page's pie chart gives about 29% of modern English words from Anglo-Norman or French and about 29% from Latin.

### Spelling and the Great Vowel Shift

- English gained the *th* spelling for /θ/ and /ð/ in place of the Old English letters þ and ð. Printing had a standardising effect; most early presses came from continental Europe, so native letters such as þ and ð died out.

- The Great Vowel Shift is dated c. 1400-1700 in the timeline, "mainly during the 15th century" in the text, and c. 1500 in the phonology section. It "affected the qualities of most long vowels". It occurred after the spelling system was fixed, which explains pairs like *mat* and *mate*.
- The first English dictionary, *A Table Alphabeticall*, was published in 1604, and Samuel Johnson's *Dictionary of the English Language* followed in 1755.

### Samples of the texts

```
// Old English: Beowulf (8th to early 11th century), opening
Hwæt! Wē Gār-Dena ...

// Middle English: Chaucer, The Canterbury Tales (end of 14th century,
// London dialect), opening
Whan that Aprill with his shoures soote ...
```

Read side by side with Chapter 7's examples: *willan* is Old English, and *'ll* is modern. Each text is evidence for the stage of the language it was written in, within the limits of the philology section.

**NOTE ON SOURCES**

The philology, place-name and English-history material comes from Wikipedia's articles on philology, toponymy and the History of English. Each page was read only up to its first 100,000 characters (the History of English page runs to 158,455, so later sections were not reached). The pages give no examples of Celtic river names, Norse *-by* or Old English *-ton*, and the History of English page does not discuss dialects, so the "dialects" item from the outline appears only through Chapter 6. The loanword paragraph's linking sentence and the closing remark about Chapter 7 are mine.

## Try It Yourself

**EXERCISE 1**

List the limits of written evidence from the philology source, and say which one stops a text telling you how a word was pronounced.

View solution

**EXERCISE 2**

Using the English timeline, give the approximate date and one consequence for the language of each: the Anglo-Saxon settlement, the Great Heathen Army, the Norman Conquest.

View solution

**EXERCISE 3**

Why does the Great Vowel Shift explain pairs like *mat* and *mate*? Use the dates and the point about spelling from the source.

View solution

## CHAPTER 8 QUICK REFERENCE

- Philology studies language in historical sources; textual criticism rebuilds original texts from variant manuscripts
- Limits: undeciphered scripts, politically charged texts, editorial distortion, writing is not pronunciation
- Place names: generic plus specific; they can preserve languages with no texts and show settlement patterns
- English: Old English c. 450, Middle English from 1066, Early Modern about 1500; about 2,000 Norse and 10,000 French loanwords
- The Great Vowel Shift came after spelling was fixed, so spelling no longer matches sound

**COMING NEXT**

*Historical Linguistics and Language Families 9* looks beyond Indo-European: the Uralic family (Hungarian), language isolates, planned languages, and creoles.

# Historical Linguistics and Language Families

Course 1 · Chapter 9 · Beyond Indo-European

Most of this course has used Indo-European, because it is the best-studied family and contains English. This chapter looks at what lies outside it, and at languages that fit the family tree badly or not at all. It also keeps a promise from Chapter 5: Hungarian is not on the Indo-European tree, and here is where it belongs.

> **THE IDEA IN A BOX**
>
> The family model works when a language descends from a proto-language by regular change. Some languages belong to a different family (Hungarian, Uralic), some have no proven relatives (isolates), some arise from contact (creoles), and some were designed (Esperanto).

## The Uralic Family and Hungarian

Wikipedia's article on the Uralic languages describes a language family native to northern Eurasia, with an estimated 21.3 million native speakers, concentrated in Central, Northern and Eastern Europe and Northern Asia. The page says the family is "not part of what is now known as the Indo-European family."

| LANGUAGE | SHARE OF SPEAKERS IN THE PAGE'S TABLE |
|---|---|
| Hungarian | 61.33% |
| Finnish | 26.97% |
| Estonian | 5.68% |

Erzya, Moksha, Mari, Udmurt and Komi each have over 100,000 speakers, all in European Russia.

### The branches

- The family has nine undisputed groups: Sámi, Finnic, Mordvinic, Mari, Permic, Hungarian, Mansi, Khanty and Samoyedic.
- Finno-Ugric is a disputed subdivision that excludes Samoyedic, and it is sometimes used as a synonym for Uralic.

- The page presents Donner's 1879 model, a traditional tree with Ugric and Finno-Permic branches, but notes that higher-order groupings are increasingly doubted.
- Hungarian belongs to the Ugric branch with Mansi and Khanty (Ob-Ugric); some scholars instead group Hungarian with Mansi.

### Cognates: the evidence at work

The page's table of Proto-Uralic cognates shows the comparative method (Chapter 3) at work on Hungarian:

| PROTO-URALIC | MEANING | FINNISH | ESTONIAN | HUNGARIAN |
|---|---|---|---|---|
| ***kala** | fish | kala | kala | hal |
| ***wete** | water | vesi | vesi | víz |
| ***käte** | hand, arm | käsi | käsi | kéz |

Look for the correspondence sets: Finnish *k-* against Hungarian *h-* (*kala*, *hal*) but Finnish *k-* against Hungarian *k-* (*käsi*, *kéz*); Finnish *-s-* against Hungarian *-z* (*vesi*, *víz*; *käsi*, *kéz*). The first pair differs by the vowel that follows, which is the kind of complementary distribution that Chapter 3 described, though I have not checked that this is the source's analysis. The page notes that cognates are not translations, since meanings can shift and loanwords can replace them. Numerals are less settled: Janhunen's proposal reconstructs Proto-Uralic **kektä* "2" and **wixti* "5", and Hungarian *hét* "7" is a replacement form.

### How was the relationship found?

- Finnish-Hungarian affinity was first proposed in the late 17th century by Fogel, Stiernhielm and Skytte.
- In 1717, Rudbeck proposed about 100 Finnish-Hungarian etymologies, of which about 40 are still considered valid.
- Sajnovics (1770) argued for a relationship based on grammatical features, and Budenz later attempted a Proto-Finno-Ugric reconstruction.
- Shared traits include agglutination, extensive case systems (Finnish has 15) and vowel harmony in many languages.

**WATCH OUT: RELATEDNESS DOES NOT MEAN YOU CAN UNDERSTAND IT**

> Mall Hellam's sample sentences in Estonian, Finnish and Hungarian were claimed to be mutually intelligible, but Geoffrey Pullum reports that neither Finns nor Hungarians could understand the other's version. The languages are related at a time depth of thousands of years; Honkola et al. (2013) estimate the family's root at 5,300 years before present, Finno-Ugric at 3,900 and Ugric at 3,300.

### Where did Proto-Uralic come from?

Several homelands have been proposed, including the Volga region, western Siberia and central Siberia. Grünthal et al. (2022) place it east of the Urals, near the Minusinsk Basin, and reject a Volga/Kama origin. Zeng et al. (2025) link Uralic speakers to ancient populations in Yakutia around 4,500 years ago. As in Chapter 4, the scholars disagree, so this course gives the views and does not choose.

## Language Isolates

Chapter 5 defined an isolate as a language with no proven relatives, a family of one. Basque is the standard example, and the source estimates about 129 isolates today. "No proven relatives" is a statement about evidence, not a claim that the language came from nowhere: the comparative method needs enough regular correspondences, and for an isolate they have not been found.

## Creoles

Wikipedia's article on creole languages defines a creole as a stable contact language that forms when a language simplifies and mixes into a new form, often a pidgin, which then expands into a full language with native speakers in a fairly brief period. A creole has a consistent grammar and a large stable vocabulary, and children acquire it as a first language.

| | PIDGIN | CREOLE |
|---|---|---|
| Learned by | adults, usually as a second language | children, as a native language (nativization) |
| Grammar and vocabulary | (not described in the summary I read) | consistent grammar, large stable vocabulary |

- A creole's lexicon is largely supplied by the parent languages, especially those of the most dominant group. Phonetic and semantic shifts are common, and the grammar often has new features that differ substantially from the parents'.
- Creoles do not fit the 19th-century neogrammarian tree model, which assumes regular sound changes. Phylogenetic classification usually relies on inherited core vocabulary and grammar; in creoles the core lexicon often has mixed origins and the grammar is largely original.
- The page says the question of which language is the parent of a creole often has no definitive answer and can lead to long-running controversies. The European dialect origin hypothesis treats French creoles as phylogenetically based on French; the monogenetic theory traces Atlantic creoles to a single Portuguese-based pidgin; others argue creoles are structurally no different from other languages and that "creole" is a sociohistoric rather than linguistic concept.
- Examples on the page include Haitian Creole ("the most creole of creoles" in McWhorter's view) and Sranan, which has existed for centuries in a diglossic relationship with Dutch.

## Planned Languages: Esperanto

Esperanto is a constructed (planned) language, which Wikipedia calls "the world's most widely spoken constructed auxiliary language". L. L. Zamenhof created it and published the first grammar, *Unua Libro*, in 1887 under the pseudonym "Doktoro Esperanto". Its vocabulary, syntax and semantics come mainly from Indo-European languages: about 80% of its vocabulary derives from Romance languages, with the rest from Germanic, Greek and Slavic languages. The infobox lists it as an international auxiliary language and an a posteriori language.

- **Native speakers:** the infobox gives "c. 1,000 (2022)", but the body text says an estimated 2,000. The page's own figures disagree.
- **Second-language speakers:** the infobox gives 30,000 to 2 million. Individual estimates are about 2 million (Culbert), above 2 million (Wandel) and 30,000-180,000 (Vendelbo); the body text also cites two estimates of around 100,000 Esperantists.

The page does not say whether Esperanto sits outside every language family. My reasoning: it was not inherited by descent from a proto-language, so the family model does not apply to its origin, and it borrows from several families. That is a conclusion drawn from the facts above, not the source's statement. Compare the course's earlier material: an Esperanto word such as the ones in Esperanto Word-Building 1 comes from Romance, Germanic or Slavic roots chosen in 1887, not by regular sound change.

**NOTE ON SOURCES**

The Uralic, creole and Esperanto material comes from Wikipedia's articles on the Uralic languages, creole languages and Esperanto (each page was read only up to its first 100,000 characters). The creole page does not mention Tok Pisin or Jamaican Patois, and it uses "parent languages" and "superstrate" rather than "lexifier", so those are left out. The sound-correspondence sentence on *kala/hal*, the claim that Esperanto sits outside the family model, and the isolates remark are mine.

## Try It Yourself

**EXERCISE 1**

Using the cognate table, state the correspondence between Finnish and Hungarian for the consonant at the end of *vesi* / *víz* and *käsi* / *kéz*. Is it a regular correspondence set?

View solution

**EXERCISE 2**

Explain why a creole does not fit the tree model, using the source's two reasons. Then say what the monogenetic theory claims.

View solution

**EXERCISE 3**

Give three facts from the source about Esperanto's origin and vocabulary. Why is "which family does it belong to?" a hard question? Point out which part of your answer is inference.

View solution

**CHAPTER 9 QUICK REFERENCE**

- Uralic: about 21.3 million native speakers; Hungarian, Finnish, Estonian are the largest; not Indo-European
- Hungarian is Ugric, with Mansi and Khanty; cognates such as *kala/hal*, *vesi/víz*, *käsi/kéz*

- Related does not mean intelligible; the family is thousands of years old
- Isolate: no proven relatives (Basque); creole: a native-speaker language that grew from a pidgin
- Esperanto (1887, Zamenhof): planned, about 80% Romance vocabulary; no inheritance from a proto-language

**COMING NEXT**

*Historical Linguistics and Language Families 10* is the last chapter: using the comparative method on a small data set, the limits of how far back language can be reconstructed, and where to go next.

# Historical Linguistics and Language Families

Course 1 · Chapter 10 · Putting It to Work

You now have every part of the toolkit: why languages change (Chapter 1), sound change (2), the comparative method (3), reconstruction (4), families (5), contact and areas (6), meaning and grammar (7), written evidence (8) and languages outside Indo-European (9). This last chapter puts the pieces into one routine, practises it on data from earlier chapters, and is honest about how far the method can see.

> **THE IDEA IN A BOX**
>
> A relationship between languages is shown by **regular correspondences in basic vocabulary**, not by look-alike words (Chapter 3). A reconstruction is a **hypothesis** with a known degree of doubt (Chapter 4). Always ask whether a shared feature could come from contact (Chapter 6) before you call it inheritance.

## The Course in One Table

| CHAPTER | CORE IDEA |
|---|---|
| 1 Why Languages Change | change is normal; synchronic versus diachronic; kinds and causes of change |
| 2 Sound Change | types of sound change; conditioned versus unconditioned; Grimm's law; regularity |
| 3 The Comparative Method | cognates and correspondence sets prove relationship; shared innovations show subgroups |
| 4 Reconstruction | the asterisk; internal and comparative reconstruction; Proto-Indo-European |
| 5 Language Families | families, branches, isolates; counts; the Indo-European branches |
| 6 Trees and Waves | the wave model; contact outcomes; sprachbunds |
| 7 Meaning and Grammar Change | semantic change; analogy; grammaticalisation |
| 8 Evidence from the Past | philology; place names; loanwords; the history of English |
| 9 Beyond Indo-European | Uralic and Hungarian; isolates; creoles; Esperanto |

## A Routine for a Small Data Set

When you meet a set of words from several languages and have to decide what they show, work through these questions in order. The routine is my summary of the course and builds on the five steps in Chapter 3's source, not on a standard from one.

```
// Routine
1. Use basic vocabulary (numbers, body parts, water, fire ...).
2. Line the words up and look for REGULAR sound matches, not just look-alikes.
3. Rule out chance, loanwords and contact (Chapter 3, Chapter 6).
4. Write the correspondence sets; check conditioned changes (A > B / X __ Y).
5. Reconstruct the older sound with the principles of Chapter 4.
6. Mark every reconstruction with * and say how confident you are.
7. Ask what would count as evidence against your conclusion.
```

### Practice data: Finnish, Estonian, Hungarian

From Chapter 9, with the Proto-Uralic forms Wikipedia gives:

| PROTO-URALIC | MEANING | FINNISH | ESTONIAN | HUNGARIAN |
|---|---|---|---|---|
| *kala | fish | kala | kala | hal |
| *wete | water | vesi | vesi | víz |
| *käte | hand, arm | käsi | käsi | kéz |

Apply the routine: the words are basic vocabulary (step 1). The Finnish and Estonian words are identical, so they add little evidence about the Finnish-Hungarian gap, and the differences between Finnish and Hungarian are the interesting part (steps 2 and 4): *-s-* against *-z*, and initial *k-* against *h-* in *kala* but *k-* in *käsi*. Three words are a very small sample, so the honest conclusion is that the table is consistent with a relationship, not a proof. The real argument rests on far more cognates, plus the grammatical traits the source lists (agglutination, many cases, vowel harmony).

## The Limits of Reconstruction

- **Time depth.** The Comparative method page gives no fixed age limit. It says reconstruction can be only partial when languages are scarcely attested, the time depth is too great, or

internal evolution obscures sound laws; relatedness is then "plausible, but uncertain".

- **Reconstructions are likely, not certain.** The same page says reconstructions of unknown proto-languages are "only likely, not certain".
- **Contact can look like kinship.** Sprachbunds "might give a false appearance of relatedness" (Chapter 6).
- **Experts disagree.** The Indo-European homeland (Kurgan versus Anatolian), the Uralic homeland and the shape of the Uralic tree are all open questions in the sources.
- **Texts are not recordings.** Writing preserves spelling, not pronunciation (Chapter 8).

> **WATCH OUT: A CONFIDENT CHART IS NOT A SETTLED FACT**
>
> Family counts range from 142 (Ethnologue) to 406 (Campbell) and 423 (Glottolog), and every tree in this course is a model. When you read a family tree or a starred form, ask who proposed it, from what evidence, and where the disagreement lies.

## Where to Go Next

| DIRECTION | WHY |
|---|---|
| Sociolinguistics: Language, Dialect and Society | who changes language, and why; the social causes of change (prestige, contact) that Chapter 1 listed. This course is planned on this site and is not yet written. |
| Writing Systems of the World | the history of the scripts behind the texts of Chapter 8. Planned, not yet written. |
| Old English | reading the stage of English that Chapter 8 introduced. Planned, not yet written. |
| A real language you study | take a language from this site and trace three words back with the routine above |
| Primary sources | the reference works behind the Wikipedia articles cited in this course, for the full evidence |

> **NOTE ON SOURCES**
>
> All facts and data in this chapter come from earlier chapters and the Wikipedia articles cited there (comparative method, Uralic languages, linguistic area, language family). No new source was fetched

for this chapter. The routine, the practice commentary, the limits list and the "where to go next" table are my own summary.

## Try It Yourself

**EXERCISE 1**

Run the routine on the Finnish-Hungarian data: write down the correspondence you can see for the word-final consonant in *vesi* / *víz* and *käsi* / *kéz*, say why three words is not enough, and name one further kind of evidence the source lists.

View solution

**EXERCISE 2**

Sort these resemblances as inheritance, chance, borrowing or contact: *ten* / *decem*; *day* / *dies*; *taboo* / Polynesian; Armenian and Iranian words; the Balkan avoidance of the infinitive.

View solution

**EXERCISE 3**

Choose two languages you know or study and write a one-paragraph plan: which basic words you would compare, what sources you would use, and what would make you doubt the result.

View solution

### CHAPTER 10 QUICK REFERENCE

- Relationship = regular correspondences in basic vocabulary, not look-alikes
- Rule out chance, loanwords and contact first; then write correspondence sets and reconstruct
- A starred form is a hypothesis; reconstructions are "only likely, not certain"
- Trees are models; counts and homelands are disputed in the sources

- Always state your evidence and what would disprove your conclusion

**COURSE COMPLETE**

**Historical Linguistics and Language Families** is finished. You can explain why languages change, describe sound change and Grimm's law, apply the comparative method, read a reconstruction critically, place a language in a family, and tell inheritance from contact. Sociolinguistics, Writing Systems of the World and Old English are the planned next steps.