

The Grammaticalization Cycle

Languages fall onto a spectrum according to

- how much of their grammar is **morphology** (i.e, endings, paradigms, inflections, agreement). These with more are called **inflected** or **synthetic** languages; Latin and Eskimo are examples, the latter an extreme example called **polysynthetic**, where the distinction between word and sentence is weak.

– and

- how much of their grammar is **syntax** (i.e, word order, constructions, particles, prepositions, idioms). These with more are called **isolating** or **analytic** languages; English and Chinese are examples.

Typically the more a language has of one kind of grammar, the less it has of the other. There's an optimal balance between convenience and efficiency, which is always shifting, like any living system. Old English had 4 noun and pronoun suffix cases comprising various suffix paradigms that were applied to just about every noun and pronoun, and used daily in every sentence. After 1000 years, only fossils are left in Modern English:

- the **genitive** 'apostrophe-S' suffix, which has fossilized into a clitic $-Z_2$, identical in form to the noun plural suffix $-Z_1$, but attaching to a noun phrase (*the King of England's wife*) rather than to a noun (**the King's of England wife*), the way a real case would work.
- the **dative/accusative** suffix, now often called '**objective**', which now is a living fossil, occurring only on 5 words, all pronouns, that are special objective forms:
 - *he ~ him*
 - *we ~ us*
 - *they ~ them*
 - *I ~ me*
 - (and sometimes) *who ~ whom*,which has been dead for ages and is a zombie word now.

All Proto-Indo-European inflection was suffixal. PIE and all of its daughter languages (Sanskrit, Greek, Latin, Proto-Germanic, etc) were synthetic and heavily inflected – Sanskrit had 8 suffixal cases and 3 voices, for instance. English, however, has lost almost all its inflections, and is now a textbook analytic language. How did this happen? The real answer is that nobody knows; it's too complicated and there's too little data. There are, however, theories. Here's one, called *Grammaticalization*, that clearly explains some things.

Most affixes occur at the ends of words; this is also where most sound changes happen. When a phonetic change occurs at the end of a word, like final **-m** and **-s** going silent in Vulgar Latin, a lot of the paradigmatic endings that Latin inherited from PIE disappeared, leading to the modern Romance languages. This moves Latin (catastrophically) from being a synthetic (inflected) type language to being an analytic (uninflected) type.

The process is called **grammaticalization**, and refers to the gradual loss of affixal inflections to sound change, followed rapidly by new syntactic constructions with auxiliary words to fill the gap, followed in time by new inflections based on common constructions with auxiliaries.

The example below happened a thousand years before Old English changed into Middle English, but it's similar. Classical Latin had 5 cases for most nouns, distinguished by suffixes, thus:

Singular	Plural	Sg	Pl	Sg	Pl	Case
amīcus	amīcī	amīca	amīcī	lēx	lēgēs	<u>Nominative</u>
amīcī	amīcōrum	amīcae	amīcārum	lēgis	lēgum	<u>Genitive</u>
amīcō	amīcīs	amīcae	amīcīs	lēgi	lēgibus	<u>Dative</u>
amīcum	amīcōs	amīcam	amīcās	lēgem	lēgēs	<u>Accusative</u>
amīcō	amīcīs	amīcā	amīcīs	lēge	lēgibus	<u>Ablative</u>
Masculine		Feminine		Consonant Stems		

(Those long vowels were pronounced *differently* from the short ones in Classical Latin.)

Around 0 CE, some natural and quite common sound changes occurred to the ends of words in common spoken (“Vulgar”) Latin. It stopped distinguishing

- final long vowels from short vowels
- final high vowels (**i** and **u**) from mid vowels (**e** and **o**, respectively)
- final **e** from **ae**,

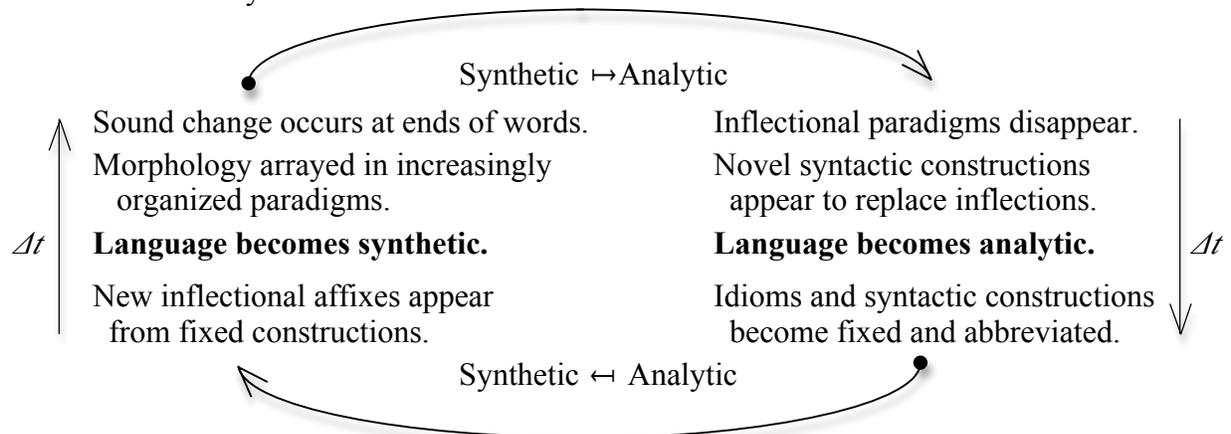
and it stopped pronouncing final **s** and **m**, as well.

To see the resulting paradigms, eliminate all the boldface final **s**’s, **m**’s, and long marks in the paradigm above, change every final **i** and **ae** to **e**, and every final **u** to **o**. The result shows that the case system collapsed because the distinctions necessary to use it collapsed, leaving word order to mark subject/object and only enough suffixes to mark masc/fem and sing/pl. The other case meanings were signalled with loads of prepositions and auxiliaries, leading to the current Romance languages, which come from Vulgar rather than Classical Latin.

After a couple millennia, we now find new morphology in the Romance languages, produced by frozen constructions with new auxiliaries. The future tense in Spanish, Italian, and French is a novel formation different from the Latin future, which got zapped by sound changes like the case system.

In Spanish, for instance, the future is completely regular and consists of the infinitive form of the verb (almost always ending in **-ir**, **-ar**, or **-er**) followed by the present tense form (spelled without the initial silent **h**) of the auxiliary verb **haber**: **-é**, **-ás**, **-á**, **-emos**, **-éis**, **-án**. This came about from the syntactic construction (much like English ‘He is to die tomorrow’, except using *have* instead of *be* as the auxiliary, and adding the auxiliary **after** the infinitive) that replaced the Latin future tense. The reduced auxiliary gradually got glued onto the end by habit, like English *hafta*, *wanna*, *gotta*, *lotta*, *lookit*, etc. Now it’s a regular paradigm; in fact, like all new paradigms, it’s very **very** regular.

But language marches on; in Brazilian Portuguese this standard future has already fallen out of use, and is currently being replaced by a newer form, derived from yet another auxiliary construction.



Δt, in this case, is probably 1 kiloyear or so, as Old English→Modern English shows. The bottom arrow is what has started to happen in French (which is becoming a prefixal language), and in the English eye dialect words *gotta*, *wanna*, *oughta*, *shoulda*, *shouldna*, *coulda*, *hadna*, *usta*, *hafta*, *lookit*, etc.