Constructing Language and Comparative Linguistics

Colin Nixon

CONSTRUCTING LANGUAGE AND COMPARATIVE LINGUISTICS

CONSTRUCTING LANGUAGE AND COMPARATIVE LINGUISTICS

Colin Nixon

BIBLIOTEX
Digital Library

Constructing Language and Comparative Linguistics
by Colin Nixon

www.bibliotex.com

Ebook ISBN: 9781984665263

Published by:
Bibliotex
Canada
Website: www.bibliotex.com

Contents

Chapter 1

Constructed Language

A **constructed language** (sometimes called a **conlang**) is a language whose phonology, grammar, and vocabulary, instead of having developed naturally, are consciously devised or invented as a work of fiction. Constructed languages may also be referred to as **artificial languages**, **planned languages** or **invented languages** and in some cases, fictional languages. Planned languages are languages that have been purposefully designed. They are the result of deliberate controlling intervention, thus of a form of language planning.

There are many possible reasons to create a constructed language, such as to ease human communication (see international auxiliary language and code); to give fiction or an associated constructed setting an added layer of realism; for experimentation in the fields of linguistics, cognitive science, and machine learning; for artistic creation; and for language games. Some people make constructed languages simply because they like doing it.

The expression *planned language* is sometimes used to indicate international auxiliary languages and other languages designed for actual use in human communication. Some prefer it to the adjective *artificial*, as this term may be perceived as pejorative. Outside Esperanto culture, the term language planning means the prescriptions given to a natural language to standardize it; in this regard, even a "natural language" may be artificial in some respects, meaning some of its words have been crafted by

conscious decision. Prescriptive grammars, which date to ancient times for classical languages such as Latin and Sanskrit, are rule-based codifications of natural languages, such codifications being a middle ground between naïve natural selection and development of language and its explicit construction. The term *glossopoeia*is also used to mean language construction, particularly construction of artistic languages.

Conlang speakers are rare. For example, the Hungarian census of 2011 found 8,397 speakers of Esperanto, and the census of 2001 found 10 of Romanid, two each of Interlingua and Ido and one each of Idiom Neutral and Mundolinco. The Russian census of 2010 found that there were in Russia about 992 speakers of Esperanto (on place 120), nine of Ido and one of Edo.

Planned, constructed, artificial

The terms "planned", "constructed", and "artificial" are used differently in some traditions. For example, few speakers of Interlingua consider their language artificial, since they assert that it has no invented content: Interlingua's vocabulary is taken from a small set of natural languages, and its grammar is based closely on these source languages, even including some degree of irregularity; its proponents prefer to describe its vocabulary and grammar as standardized rather than artificial or constructed. Similarly, Latino sine flexione (LsF) is a simplification of Latin from which the inflectionshave been removed. As with Interlingua, some prefer to describe its development as "planning" rather than "constructing". Some speakers of Esperanto and Esperantidos also avoid the term

"artificial language" because they deny that there is anything "unnatural" about the use of their language in human communication.

By contrast, some philosophers have argued that all human languages are conventional or artificial. François Rabelais's fictional giant Pantagruel, for instance, said: "It is a misuse of terms to say that we have natural language; languages *are* through arbitrary institutions and the conventions of peoples: voices, as the dialecticians say, don't signify naturally, but capriciously."

Furthermore, fictional or experimental languages can be considered*naturalistic* if they model real world languages. For example, if a naturalistic conlang is derived *a posteriori* from another language (real or constructed), it should imitate natural processes of phonological, lexical, and grammatical change. In contrast with languages such as Interlingua, naturalistic fictional languages are not usually intended for easy learning or communication. Thus, naturalistic fictional languages tend to be more difficult and complex. While Interlingua has simpler grammar, syntax, and orthography than its source languages (though more complex and irregular than Esperanto or its descendants), naturalistic fictional languages typically mimic behaviors of natural languages like irregular verbs and nouns, and complicated phonological processes.

Overview

In terms of purpose, most constructed languages can broadly be divided into:

- *Engineered languages* (*engelangs*/ˈɛnd͡ʒlæŋz/), further subdivided into logical languages (*loglangs*), philosophical languages and experimental languages; devised for the purpose of experimentation in logic, philosophy, or linguistics;
- *Auxiliary languages* (*auxlangs*) devised for international communication (also IALs, for International Auxiliary Language);
- *Artistic languages* (*artlangs*) devised to create aesthetic pleasure or humorous effect, *just for fun*; usually secret languages and mystical languages are classified as artlangs.

The boundaries between these categories are by no means clear. A constructed language could easily fall into more than one of the above categories. A logical language created for aesthetic reasons would also be classifiable as an artistic language, which might be created by someone with philosophical motives intending for said conlang to be used as an auxiliary language. There are no rules, either inherent in the process of language construction or externally imposed, that would limit a constructed language to fitting only one of the above categories.

A constructed language can have native speakers if young children learn it from parents who speak it fluently. According to *Ethnologue*, there are "200–2000 who speak Esperanto as a first language". A member of the Klingon Language Institute, d'Armond Speers, attempted to raise his son as a native (bilingual with English) Klingon speaker.

As soon as a constructed language has a community of fluent speakers, especially if it has numerous native speakers, it begins to evolve and hence loses its constructed status. For example, Modern Hebrew and its pronunciation norms were developed from existing traditions of Hebrew, such as Mishnaic Hebrew and Biblical Hebrew following a general Sephardic pronunciation, rather than engineered from scratch, and has undergone considerable changes since the state of Israel was founded in 1948 (Hetzron 1990:693). However, linguist Ghil'ad Zuckermann argues that Modern Hebrew, which he terms "Israeli", is a Semito-European hybrid based not only on Hebrew but also on Yiddish and other languages spoken by revivalists. Zuckermann therefore endorses the translation of the Hebrew Bible into what he calls "Israeli". Esperanto as a living spoken language has evolved significantly from the prescriptive blueprint published in 1887, so that modern editions of the *Fundamenta Krestomatio*, a 1903 collection of early texts in the language, require many footnotes on the syntactic and lexical differences between early and modern Esperanto.

Proponents of constructed languages often have many reasons for using them. The famous but disputed Sapir–Whorf hypothesisis sometimes cited; this claims that the language one speaks influences the way one thinks. Thus, a "better" language should allow the speaker to think more clearly or intelligently or to encompass more points of view; this was the intention of Suzette Haden Elgin in creating Láadan, a feminist language embodied in her feminist science fiction series *Native Tongue*. Constructed languages have been included in standardized tests such as the SAT, where they were used to test the applicant's ability to infer and apply grammatical

rules. A constructed language could also be used to *restrict* thought, as in George Orwell's Newspeak, or to *simplify* thought, as in Toki Pona. In contrast, linguists such as Steven Pinker argue that ideas exist independently of language. For example, in the book *The Language Instinct*, Pinker states that children spontaneously re-invent slang and even grammar with each generation. These linguists argue that attempts to control the range of human thought through the reform of language would fail, as concepts like "freedom" will reappear in new words if the old words vanish.

Proponents claim a particular language makes it easier to express and understand concepts in one area, and more difficult in others. An example can be taken from the way various programming languages make it easier to write certain kinds of programs and harder to write others.

Another reason cited for using a constructed language is the telescope rule, which claims that it takes less time to first learn a simple constructed language and then a natural language, than to learn only a natural language. Thus, if someone wants to learn English, some suggest learning Basic English first. Constructed languages like Esperanto and Interlingua are in fact often simpler due to the typical lack of irregular verbs and other grammatical quirks. Some studies have found that learning Esperanto helps in learning a non-constructed language later (see propaedeutic value of Esperanto).

Codes for constructed languages include the ISO 639-2 "art" for conlangs; however, some constructed languages have their own ISO 639 language codes (e.g. "eo" and "epo" for Esperanto,

"jbo" for Lojban, "ia" and "ina" for Interlingua, "tlh" for Klingon, and "io" and "ido" for Ido).

One constraint on a constructed language is that if it was constructed to be a natural language for use by fictional foreigners or aliens, as with Dothraki and High Valyrian in the *Game of Thrones* series, which was adapted from the *A Song of Ice and Fire* book series, the language should be easily pronounced by actors, and should fit with and incorporate any fragments of the language already invented by the book's author, and preferably also fit with any personal names of fictional speakers of the language.

A priori and *a posteriori* languages

An *a priori* constructed language is one whose features (including vocabulary, grammar, etc.) are not based on an existing language, and an *a posteriori* language is the opposite. This categorization, however, is not absolute, as many constructed languages may be called*a priori* when considering some linguistic factors, and at the same time *a posteriori* when considering other factors.

A priori language

An *a priori* language (from Latin*a priori*, "from the former") is any constructed language of which all or a number of features are not based on existing languages, but rather invented or elaborated as to work in a different way or to allude different purposes. Some *a priori* languages are designed to be international auxiliary languages that remove what could be considered an unfair learning advantage for native speakers of

a source language that would otherwise exist for *a posteriori* languages. Others, known as philosophical or taxonomic languages, try to categorize their vocabulary, either to express an underlying philosophy or to make it easier to recognize new vocabulary. Finally, many artistic languages, created for either personal use or for use in a fictional medium, employ consciously constructed grammars and vocabularies, and are best understood as *a priori*.

Examples of *a priori* languages

A priori international auxiliary languages

- Balaibalan, attributed to Fazlallah Astarabadi or Muhyi Gulshani (14th century)
- Solresol by François Sudre (1827)
- Ro by Edward Foster (1906)
- Sona by Kenneth Searight (1935)
- Babm by Rikichi Okamoto (1962)
- Kotava by Staren Fetcey (1978)

Experimental languages

- Láadan by Suzette Haden Elgin (1982)
- Ithkuil by John Quijada (2011)

A priori artistic languages

- Quenya and Sindarin by J. R. R. Tolkien for *The Lord of the Rings* (published 1954)
- aUI by W. John Weilgart (1962)
- Klingon by Marc Okrand for the science-fiction franchise *Star Trek* (1985)
- Kēlen by Sylvia Sotomayor (1998)

- Na'vi by Paul Frommer for the movie *Avatar* (2005)
- Dothraki and Valyrian by David Peterson for the television series *Game of Thrones* (2011)
- Kiliki by Madhan Karky for the *Baahubali* films (2015)

Community languages

- Damin (Yangkaal and Lardil people, 19th century or earlier)
- Eskayan (Eskaya people, ca. 1920)
- Medefaidrin (Ibibio, 1930s)

A posteriori language

An ***a posteriori* language** (from Latin*a posteriori*, "from the latter"), according to French linguist Louis Couturat, is any constructed language whose elements are borrowed from or based on existing languages. The term can also be extended to controlled versions of natural languages, and is most commonly used to refer to vocabulary despite other features. Likewise, zonal constructed languages (auxiliary languages for speakers of a particular language family) are a posteriori by definition.

While most auxiliary languages are *a posteriori* due to their intended function as a medium of communication, many artistic languages are fully a posteriori in design—many for the purposes of alternate history. In distinguishing whether the language is *a priori* or *a posteriori*, the prevalence and distribution of respectable traits is often the key.

Examples of *a posteriori* languages

A posteriori artistic languages

- Brithenig by Andrew Smith (1996)
- Atlantean by Marc Okrand for the film *Atlantis: The Lost Empire* (2001)
- Toki Pona by Sonja Lang (2001)
- Wenedyk by Jan van Steenbergen (2002)
- Trigedasleng by David Peterson for the TV series *The 100* (2014)

Controlled auxiliary languages

- Latino sine flexione (Latin, 1911)
- Basic English (English, 1925)
- N'Ko (Manding, 1949)
- Learning English (English, 1959)
- Kitara (SW Ugandan Bantu, 1990)
- Globish (English, 2004)

A posteriori international auxiliary languages

- Volapük (1879)
- Esperanto (1887)
- Ido (1907)
- Interlingue (1922)
- Interlingua (1951)
- Lingua Franca Nova (1965)
- Afrihili (1970)
- Glosa (ca. 1979)
- Sambahsa (2007)

- Lingwa de planeta (2010)
- Idiom Neutral (1880)

Zonal auxiliary languages

- Main article: Zonal constructed language
- Efatese (C. Vanuatu Oceanic, 19th century)
- Romanid (Romance, 1956)
- Folkspraak (Germanic, 1995)
- Interslavic (Slavic, 2011)

History

Ancient linguistic experiments

Grammatical speculation dates from Classical Antiquity, appearing for instance in Plato's *Cratylus* in Hermogenes's contention that words are not inherently linked to what they refer to; that people apply "a piece of their own voice... to the thing". Athenaeus of Naucratis, in Book III of *Deipnosophistae*, tells the story of two figures: Dionysius of Sicily and Alexarchus. Dionysius of Sicily created neologisms like *menandros* "virgin" (from *menei* "waiting" and *andra* "husband"), *menekratēs* "pillar" (from *menei* "it remains in one place" and *kratei* "it is strong"), and *ballantion* "javelin" (from *balletai enantion* "thrown against someone"). Incidentally, the more common Greek words for those three are *parthenos*, *stulos*, and *akon*. Alexarchus of Macedon, the brother of King Cassander of Macedon, was the founder of the city of Ouranopolis. Athenaeus recounts a story told by Heracleides of Lembos that Alexarchus "introduced a peculiar vocabulary, referring to a rooster as a "dawn-crier," a barber as a "mortal-

shaver," a drachma as "worked silver"...and a herald as an *aputēs* [from *ēputa* "loud-voiced"]. "He once wrote something... to the public authorities in Casandreia...As for what this letter says, in my opinion not even the Pythian god could make sense of it."

While the mechanisms of grammar suggested by classical philosophers were designed to explain existing languages (Latin, Greek, and Sanskrit), they were not used to construct new grammars. Roughly contemporary to Plato, in his descriptive grammar of Sanskrit, Pā?ini constructed a set of rules for explaining language, so that the text of his grammar may be considered a mixture of natural and constructed language.

Early constructed languages

A legend recorded in the seventh-century Irish work Auraicept na n-Éces claims that Fénius Farsaid visited Shinar after the confusion of tongues, and he and his scholars studied the various languages for ten years, taking the best features of each to create *in Bérla tóbaide* ("the selected language"), which he named *Goídelc*—the Irish language. This appears to be the first mention of the concept of a constructed language in literature.

The earliest non-natural languages were considered less "constructed" than "super-natural", mystical, or divinely inspired. The Lingua Ignota, recorded in the 12th century by St. Hildegard of Bingen, is an example, and apparently the first entirely artificial language. It is a form of private mystical cant (see also language of angels). An important example from

Middle-Eastern culture is Balaibalan, invented in the 16th century. Kabbalistic grammatical speculation was directed at recovering the original language spoken by Adam and Eve in Paradise, lost in the confusion of tongues. The first Christian project for an ideal language is outlined in Dante Alighieri's *De vulgari eloquentia*, where he searches for the ideal Italian vernacular suited for literature. Ramon Llull's *Ars Magna* was a project of a perfect language with which the infidels could be convinced of the truth of the Christian faith. It was basically an application of combinatorics on a given set of concepts. During the Renaissance, Lullian and Kabbalistic ideas were drawn upon in a magical context, resulting in cryptographic applications.

Perfecting language

Renaissance interest in Ancient Egypt, notably the discovery of the *Hieroglyphica* of Horapollo, and first encounters with the Chinese script directed efforts towards a perfect written language. Johannes Trithemius, in *Steganographia* and *Polygraphia*, attempted to show how all languages can be reduced to one.

In the 17th century, interest in magical languages was continued by the Rosicrucians and Alchemists (like John Dee and his Enochian). Jakob Boehme in 1623 spoke of a "natural language" (*Natursprache*) of the senses.

Musical languages from the Renaissance were tied up with mysticism, magic and alchemy, sometimes also referred to as the language of the birds. The Solresol project of 1817 re-invented the concept in a more pragmatic context.

17th and 18th century: advent of philosophical languages

The 17th century saw the rise of projects for "philosophical" or "a priori" languages, such as:

- Francis Lodwick's *A Common Writing* (1647) and *The Groundwork or Foundation laid (or So Intended) for the Framing of a New Perfect Language and a Universal Common Writing* (1652)
- Sir Thomas Urquhart's *Ekskybalauron* (1651) and *Logopandecteision* (1652)
- George Dalgarno's *Ars signorum*, 1661
- John Wilkins' *Essay towards a Real Character, and a Philosophical Language*, 1668

These early taxonomic conlangs produced systems of hierarchical classification that were intended to result in both spoken and written expression. Leibniz had a similar purpose for his *lingua generalis* of 1678, aiming at a lexicon of characters upon which the user might perform calculations that would yield true propositions automatically, as a side-effect developing binary calculus.

These projects were not only occupied with reducing or modelling grammar, but also with the arrangement of all human knowledge into "characters" or hierarchies, an idea that with the Enlightenment would ultimately lead to the *Encyclopédie*. Many of these 17th–18th centuries conlangs were pasigraphies, or purely written languages with no spoken form or a spoken form that would vary greatly according to the native language of the reader.

Leibniz and the encyclopedists realized that it is impossible to organize human knowledge unequivocally in a tree diagram, and consequently to construct an *a priori* language based on such a classification of concepts. Under the entry *Charactère*, D'Alembert critically reviewed the projects of philosophical languages of the preceding century. After the *Encyclopédie*, projects for *a priori* languages moved more and more to the lunatic fringe. Individual authors, typically unaware of the history of the idea, continued to propose taxonomic philosophical languages until the early 20th century (e.g. Ro), but most recent engineered languages have had more modest goals; some are limited to a specific field, like mathematical formalism or calculus (e.g. Lincos and programming languages), others are designed for eliminating syntactical ambiguity (e.g., Loglan and Lojban) or maximizing conciseness (e.g., Ithkuil).

19th and 20th centuries: auxiliary languages

Already in the *Encyclopédie* attention began to focus on *a posteriori* auxiliary languages. Joachim Faiguet de Villeneuve in the article on *Langue* wrote a short proposition of a "laconic" or regularized grammar of French. During the 19th century, a bewildering variety of such International Auxiliary Languages (IALs) were proposed, so that Louis Couturat and Léopold Leau in *Histoire de la langue universelle* (1903) reviewed 38 projects.

The first of these that made any international impact was Volapük, proposed in 1879 by Johann Martin Schleyer; within a decade, 283 Volapükist clubs were counted all over the globe. However, disagreements between Schleyer and some prominent users of the language led to schism, and by the mid-

1890s it fell into obscurity, making way for Esperanto, proposed in 1887 by L. L. Zamenhof, and its descendants. Interlingua, the most recent auxlang to gain a significant number of speakers, emerged in 1951, when the International Auxiliary Language Association published its Interlingua–English Dictionary and an accompanying grammar. The success of Esperanto did not stop others from trying to construct new auxiliary languages, such as Leslie Jones' Eurolengo, which mixes elements of English and Spanish.

Loglan (1955) and its descendants constitute a pragmatic return to the aims of the *a priori* languages, tempered by the requirement of usability of an auxiliary language. Thus far, these modern *a priori* languages have garnered only small groups of speakers.

Robot Interaction Language (2010) is a spoken language that is optimized for communication between machines and humans. The major goals of ROILA are that it should be easily learnable by the human user, and optimized for efficient recognition by computer speech recognition algorithms.

Artlangs

Language can be artistic to the extent that artists use language as a source of creativity in art, poetry, calligraphy or as a metaphor to address themes as cultural diversity and the vulnerability of the individual in a globalizing world.

Some people prefer however to take pleasure in constructing, crafting a language by a conscious decision for reasons of literary enjoyment or aesthetic reasons without any claim of usefulness. Such artistic languages begin to appear in Early

Modern literature (in Pantagruel, and in Utopian contexts), but they only seem to gain notability as serious projects beginning in the 20th century. *A Princess of Mars* (1912) by Edgar Rice Burroughs was possibly the first fiction of that century to feature a constructed language. J. R. R. Tolkien developed families of related fictional languages and discussed artistic languages publicly, giving a lecture entitled "*A Secret Vice*" in 1931 at a congress. (Orwell's Newspeak is considered a satire of an international auxiliary language rather than an artistic language proper.)

By the beginning of the first decade of the 21st century, it had become common for science-fiction and fantasy works set in other worlds to feature constructed languages, or more commonly, an extremely limited but defined vocabulary which *suggests* the existence of a complete language, or whatever portions of the language are needed for the story, and constructed languages are a regular part of the genre, appearing in *Star Wars*, *Star Trek*, *Lord of the Rings* (Elvish), *Stargate SG-1*, *Atlantis: The Lost Empire*, *Game of Thrones* (Dothraki language and Valyrian languages), *The Expanse*, *Avatar*, *Dune* and the *Myst* series of computer adventure games.

Ownership of constructed languages

The matter of whether or not a constructed language can be owned or protected by intellectual property laws, or if it would even be possible to enforce those laws, is contentious.

In a 2015 lawsuit, CBS and Paramount Pictures challenged a fan film project called Axanar, stating the project infringed

upon their intellectual property, which included the Klingon language, among other creative elements. During the controversy, Marc Okrand, the language's original designer expressed doubt as to whether Paramount's claims of ownership were valid.

David J. Peterson, a linguist who created multiple well-known constructed languages including the Valyrian languages and Dothraki, advocated a similar opinion, saying that "Theoretically, anyone can publish anything using any language I created, and, in my opinion, neither I nor anyone else should be able to do anything about it."

However, Peterson also expressed concern that the respective rights-holders—regardless of whether or not their ownership of the rights is legitimate—would be likely to sue individuals who publish material in said languages, especially if the author might profit from said material.

Furthermore, comprehensive learning material for such constructed languages as High Valyrian and Klingon has been published and made freely accessible on the language-learning platform Duolingo—but those courses are licensed by the respective copyright holders. Because only a few such disputes have occurred thus far, the legal consensus on ownership of languages remains uncertain.

The Tasmanian Aboriginal Center claims ownership of Palawa kani, an attempted composite reconstruction up to a dozen extinct Tasmanian indigenous languages, and has asked Wikipedia to remove its page on the project. However, there is no current legal backing for the claim.

Modern conlang organizations

Various paper zines on constructed languages were published from the 1970s through the 1990s, such as *Glossopoeic Quarterly*, *Taboo Jadoo*, and *The Journal of Planned Languages*. The Conlang Mailing List was founded in 1991, and later split off an AUXLANG mailing list dedicated to international auxiliary languages. In the early to mid-1990s a few conlang-related zines were published as email or websites, such as *Vortpunoj* and *Model Languages*. The Conlang mailing list has developed a community of conlangers with its own customs, such as translation challenges and translation relays, and its own terminology.

Sarah Higley reports from results of her surveys that the demographics of the Conlang list are primarily men from North America and western Europe, with a smaller number from Oceania, Asia, the Middle East, and South America, with an age range from thirteen to over sixty; the number of women participating has increased over time.

More recently founded online communities include the Zompist Bulletin Board (ZBB; since 2001) and the Conlanger Bulletin Board. Discussion on these forums includes presentation of members' conlangs and feedback from other members, discussion of natural languages, whether particular conlang features have natural language precedents, and how interesting features of natural languages can be repurposed for conlangs, posting of interesting short texts as translation challenges, and meta-discussion about the philosophy of conlanging, conlangers' purposes, and whether conlanging is an art or a hobby. Another 2001 survey by Patrick Jarrett

showed an average age of 30.65, with the average time since starting to invent languages 11.83 years. A more recent thread on the ZBB showed that many conlangers spend a relatively small amount of time on any one conlang, moving from one project to another; about a third spend years on developing the same language.

Chapter 2

Comparative Linguistics

Comparative linguistics, or **comparative-historical linguistics** (formerly **comparative philology**) is a branch of historical linguistics that is concerned with comparing languages to establish their historical relatedness.

Genetic relatedness implies a common origin or proto-language and comparative linguistics aims to construct language families, to reconstruct proto-languages and specify the changes that have resulted in the documented languages. To maintain a clear distinction between attested and reconstructed forms, comparative linguists prefix an asterisk to any form that is not found in surviving texts. A number of methods for carrying out language classification have been developed, ranging from simple inspection to computerised hypothesis testing. Such methods have gone through a long process of development.

Methods

The fundamental technique of comparative linguistics is to compare phonological systems, morphological systems, syntax and the lexicon of two or more languages using techniques such as the comparative method. In principle, every difference between two related languages should be explicable to a high degree of plausibility; systematic changes, for example in phonological or morphological systems are expected to be highly regular (consistent). In practice, the comparison may be

more restricted, e.g. just to the lexicon. In some methods it may be possible to reconstruct an earlier proto-language. Although the proto-languages reconstructed by the comparative method are hypothetical, a reconstruction may have predictive power. The most notable example of this is Ferdinand de Saussure's proposal that the Indo-Europeanconsonant system contained laryngeals, a type of consonant attested in no Indo-European language known at the time. The hypothesis was vindicated with the discovery of Hittite, which proved to have exactly the consonants Saussure had hypothesized in the environments he had predicted.

Where languages are derived from a very distant ancestor, and are thus more distantly related, the comparative method becomes less practicable. In particular, attempting to relate two reconstructed proto-languages by the comparative method has not generally produced results that have met with wide acceptance.

The method has also not been very good at unambiguously identifying sub-families; thus, different scholars have produced conflicting results, for example in Indo-European. A number of methods based on statistical analysis of vocabulary have been developed to try and overcome this limitation, such as lexicostatistics and mass comparison. The former uses lexical cognates like the comparative method, while the latter uses only lexical similarity. The theoretical basis of such methods is that vocabulary items can be matched without a detailed language reconstruction and that comparing enough vocabulary items will negate individual inaccuracies; thus, they can be used to determine relatedness but not to determine the proto-language.

History

The earliest method of this type was the comparative method, which was developed over many years, culminating in the nineteenth century. This uses a long word list and detailed study. However, it has been criticized for example as subjective, informal, and lacking testability. The comparative method uses information from two or more languages and allows reconstruction of the ancestral language. The method of internal reconstruction uses only a single language, with comparison of word variants, to perform the same function. Internal reconstruction is more resistant to interference but usually has a limited available base of utilizable words and is able to reconstruct only certain changes (those that have left traces as morphophonological variations).

In the twentieth century an alternative method, lexicostatistics, was developed, which is mainly associated with Morris Swadesh but is based on earlier work. This uses a short word list of basic vocabulary in the various languages for comparisons. Swadesh used 100 (earlier 200) items that are assumed to be cognate (on the basis of phonetic similarity) in the languages being compared, though other lists have also been used. Distance measures are derived by examination of language pairs but such methods reduce the information. An outgrowth of lexicostatistics is glottochronology, initially developed in the 1950s, which proposed a mathematical formula for establishing the date when two languages separated, based on percentage of a core vocabulary of culturally independent words. In its simplest form a constant rate of change is assumed, though later versions allow variance but still fail to achieve reliability. Glottochronology has met

with mounting scepticism, and is seldom applied today. Dating estimates can now be generated by computerised methods that have fewer restrictions, calculating rates from the data. However, no mathematical means of producing proto-language split-times on the basis of lexical retention has been proven reliable.

Another controversial method, developed by Joseph Greenberg, is mass comparison. The method, which disavows any ability to date developments, aims simply to show which languages are more and less close to each other. Greenberg suggested that the method is useful for preliminary grouping of languages known to be related as a first step toward more in-depth comparative analysis.

However, since mass comparison eschews the establishment of regular changes, it is flatly rejected by the majority of historical linguists.

Recently, computerised statistical hypothesis testing methods have been developed which are related to both the comparative method and lexicostatistics. Character based methods are similar to the former and distanced based methods are similar to the latter (see Quantitative comparative linguistics). The characters used can be morphological or grammatical as well as lexical. Since the mid-1990s these more sophisticated tree- and network-based phylogenetic methods have been used to investigate the relationships between languages and to determine approximate dates for proto-languages. These are considered by many to show promise but are not wholly accepted by traditionalists. However, they are not intended to replace older methods but to supplement them. Such

statistical methods cannot be used to derive the features of a proto-language, apart from the fact of the existence of shared items of the compared vocabulary. These approaches have been challenged for their methodological problems, since without a reconstruction or at least a detailed list of phonological correspondences there can be no demonstration that two words in different languages are cognate.

Related fields

There are other branches of linguistics that involve comparing languages, which are not, however, part of *comparative linguistics*:

- Linguistic typology compares languages to classify them by their features. Its ultimate aim is to understand the universals that govern language, and the range of types found in the world's languages in respect of any particular feature (word order or vowel system, for example). Typological similarity does not imply a historical relationship. However, typological arguments can be used in comparative linguistics: one reconstruction may be preferred to another as typologically more plausible.
- Contact linguistics examines the linguistic results of contact between the speakers of different languages, particularly as evidenced in loan words. An empirical study of loans is by definition historical in focus and therefore forms part of the subject matter of historical linguistics. One of the goals of etymology is to establish which items in a language's vocabulary result from linguistic contact. This is

also an important issue both for the comparative method and for the lexical comparison methods, since failure to recognize a loan may distort the findings.

- Contrastive linguistics compares languages usually with the aim of assisting language learning by identifying important differences between the learner's native and target languages. Contrastive linguistics deals solely with present-day languages.

Pseudolinguistic comparisons

Comparative linguistics includes the study of the historical relationships of languages using the comparative method to search for regular (i.e. recurring) correspondences between the languages' phonology, grammar and core vocabulary, and through hypothesis testing; some persons with little or no specialization in the field sometimes attempt to establish historical associations between languages by noting similarities between them, in a way that is considered pseudoscientific by specialists (e.g. African/Egyptian comparisons).

The most common method applied in pseudoscientific language comparisons is to search two or more languages for words that seem similar in their sound and meaning. While similarities of this kind often seem convincing to laypersons, linguistic scientists consider this kind of comparison to be unreliable for two primary reasons. First, the method applied is not well-defined: the criterion of similarity is subjective and thus not subject to verification or falsification, which is contrary to the principles of the scientific method. Second, the large size of all

languages' vocabulary and a relatively limited inventory of articulated sounds used by most languages makes it easy to find coincidentally similar words between languages.

There are sometimes political or religious reasons for associating languages in ways that some linguists would dispute. For example, it has been suggested that the Turanian or Ural–Altaic language group, which relates Sami and other languages to the Mongolian language, was used to justify racism towards the Sami in particular. There are also strong, albeit areal not genetic, similarities between the Uralic and Altaiclanguages which provided an innocent basis for this theory. In 1930s Turkey, some promoted the Sun Language Theory, one that showed that Turkic languages were close to the original language. Some believers in Abrahamic religions try to derive their native languages from Classical Hebrew, as Herbert W. Armstrong, a proponent of British Israelism, who said that the word "British" comes from Hebrew *brit* meaning "covenant" and *ish* meaning "man", supposedly proving that the British people are the 'covenant people' of God. AndLithuanian-American archaeologistMarija Gimbutas argued during the mid-1900s that Basque is clearly related to the extinct Pictish and Etruscan languages, in attempt to show that Basque was a remnant of an "Old European culture". In the *Dissertatio de origine gentium Americanarum* (1625), the Dutch lawyer Hugo Grotius "proves" that the American Indians (Mohawks) speak a language (*lingua Maquaasiorum*) derived from Scandinavian languages (Grotius was on Sweden's payroll), supporting Swedish colonial pretensions in America. The Dutch doctor Johannes Goropius Becanus, in his *Origines Antverpiana* (1580) admits *Quis est enim qui non amet patrium sermonem* ("Who does not love his fathers' language?"), whilst asserting

that Hebrew is derived from Dutch. The Frenchman Éloi Johanneau claimed in 1818 (*Mélanges d'origines étymologiques et de questions grammaticales*) that the Celtic language is the oldest, and the mother of all others.

In 1759, Joseph de Guignes theorized (*Mémoire dans lequel on prouve que les Chinois sont une colonie égyptienne*) that the Chinese and Egyptians were related, the former being a colony of the latter. In 1885, Edward Tregear (*The Aryan Maori*) compared the Maori and "Aryan" languages. Jean Prat [fr], in his 1941 *Les langues nitales*, claimed that the Bantu languages of Africa are descended from Latin, coining the French linguistic term *nitale* in doing so. But the Bantu language is also claimed to be related to Ancient Egyptian by Mubabinge Bilolo [fr].

Ancient Egyptian is, according to Cheikh Anta Diop, related to the Wolof language. And, according to Gilbert Ngom, Ancient Egyptian is similar to the Duala language, just as Egyptian is related to Brabantic, following Becanus in his *Hieroglyphica*, still using comparative methods.

The first practitioners of comparative linguistics were not universally acclaimed: upon reading Becanus' book, Scaliger wrote *never did I read greater nonsense*, and Leibniz coined the term *goropism* (from Goropius) to designate a far-sought, ridiculous etymology.

There have also been claims that humans are descended from other, non-primate animals, with use of the voice referred to as the main point of comparison. Jean-Pierre Brisset (*La Grande Nouvelle*, around 1900) believed and asserted that humans descended from the frog, by linguistic means, in that the

croaking of frogs sounds similar to spoken French; he held that the French word *logement*, "dwelling", derived from the word *l'eau*, "water".

Linguistic typology

Linguistic typology (or **language typology**) is a field of linguistics that studies and classifies languages according to their structural features. Its aim is to describe and explain the common properties and the structural diversity of the world's languages. Its subdisciplines include, but are not limited to: qualitative typology, which deals with the issue of comparing languages and within-language variance; quantitative typology, which deals with the distribution of structural patterns in the world's languages; theoretical typology, which explains these distributions; syntactic typology, which deals with word order, word form, word grammar and word choice; and lexical typology, which deals with language vocabulary.

History

Joseph Greenbergis considered the founder of modern linguistic typology, a field that he has revitalized with his publications in the 1960s and 1970s.

Qualitative typology

Qualitative typology develops cross-linguistically viable notions or types that provide a framework for the description and comparison of individual languages. A few examples appear below.

Typological systems

Subject–verb–object positioning

One set of types reflects the basic order of subject, verb, and direct object in sentences:

- Object–subject–verb (OSV)
- Object–verb–subject (OVS)
- Subject–verb–object (SVO)
- Subject–object–verb (SOV)
- Verb–subject–object (VSO)
- Verb–object–subject (VOS)

These labels usually appear abbreviated as "SVO" and so forth, and may be called "typologies" of the languages to which they apply. The most commonly attested word orders are SOV and SVO while the least common orders are those that are object initial with OVS being the least common with only four attested instances.

In the 1980s, linguists began to question the relevance of geographical distribution of different values for various features of linguistic structure. They may have wanted to discover whether a particular grammatical structure found in one language is likewise found in another language in the same geographic location. Some languages split verbs into an auxiliary and an infinitive or participle and put the subject and/or object between them. For instance, German (*Ich* ***habe*** *einen Fuchs im Wald* ***gesehen*** - *"I have a fox in-the woods seen"), Dutch (*Hans* ***vermoedde*** *dat Jan Marie* ***zag leren zwemmen*** - *"Hans suspected that Jan Marie saw to learn to

swim") and Welsh (***Mae****'r gwirio sillafu wedi'i* ***gwblhau*** - *"Is the checking spelling after its to complete"). In this case, linguists base the typology on the non-analytic tenses (i.e. those sentences in which the verb is not split) or on the position of the auxiliary. German is thus SVO in main clauses and Welsh is VSO (and preposition phrases would go after the infinitive).

Many typologists classify both German and Dutch as V2 languages, as the verb invariantly occurs as the second element of a full clause.

Some languages allow varying degrees of freedom in their constituent order, posing a problem for their classification within the subject–verb–object schema. Languages with bound case markings for nouns, for example, tend to have more flexible word orders than languages where case is defined by position within a sentence or presence of a preposition. To define a basic constituent order type in this case, one generally looks at frequency of different types in declarative affirmative main clauses in pragmatically neutral contexts, preferably with only old referents. Thus, for instance, Russian is widely considered an SVO language, as this is the most frequent constituent order under such conditions—all sorts of variations are possible, though, and occur in texts. In many inflected languages, such as Russian, Latin, and Greek, departures from the default word-orders are permissible but usually imply a shift in focus, an emphasis on the final element, or some special context. In the poetry of these languages, the word order may also shift freely to meet metrical demands. Additionally, freedom of word order may vary within the same language—for example, formal, literary,

or archaizing varieties may have different, stricter, or more lenient constituent-order structures than an informal spoken variety of the same language.

On the other hand, when there is no clear preference under the described conditions, the language is considered to have "flexible constituent order" (a type unto itself).

An additional problem is that in languages without living speech communities, such as Latin, Ancient Greek, and Old Church Slavonic, linguists have only written evidence, perhaps written in a poetic, formalizing, or archaic style that mischaracterizes the actual daily use of the language. The daily spoken language of Sophocles or Cicero might have exhibited a different or much more regular syntax than their written legacy indicates.

OV/VO correlations

A second major way of syntactic categorization is by excluding the subject from consideration. It is a well-documented typological feature that languages with a dominant OV order (object before verb), Japanese for example, tend to have postpositions. In contrast, VO languages (verb before object) like English tend to have prepositions as their main adpositional type. Several OV/VO correlations have been uncovered.

Theoretical issues

Several processing explanations were proposed in the 1980s and 1990s for the above correlations. They suggest that the brain finds it easier to parsesyntactic patterns which are either

right or left branching, but not mixed. The most widely held such explanation is John A. Hawkins' Grammar-Performance Correspondence Hypothesis which argues that language is a non-innate adaptation to innate cognitive mechanisms. Typological tendencies are considered as being based on language users' preference for grammars that are organized efficiently, and on their avoidance of word orderings which cause processing difficulty. Some languages however exhibit regular inefficient patterning. These include the VO languages Chinese, with the adpositional phrase before the verb, and Finnish whichhas postpositions; but there are few other profoundly exceptional languages.

Morphosyntactic alignment

Another common classification distinguishes nominative–accusative alignment patterns and ergative–absolutive ones. In a language with cases, the classification depends on whether the subject (S) of an intransitive verb has the same case as the agent (A) or the patient (P) of a transitive verb. If a language has no cases, but the word order is AVP or PVA, then a classification may reflect whether the subject of an intransitive verb appears on the same side as the agent or the patient of the transitive verb. Bickel (2011) has argued that alignment should be seen as a construction-specific property rather than a language-specific property.

Many languages show mixed accusative and ergative behaviour (for example: ergative morphology marking the verb arguments, on top of an accusative syntax). Other languages (called "active languages") have two types of intransitive verbs—some of them ("active verbs") join the subject in the same case as the agent

of a transitive verb, and the rest ("stative verbs") join the subject in the same case as the patient. Yet other languages behave ergatively only in some contexts (this "split ergativity" is often based on the grammatical person of the arguments or on the tense/aspect of the verb). For example, only some verbs in Georgian behave this way, and, as a rule, only while using the perfective (aorist).

Phonological systems

- Linguistic typology also seeks to identify patterns in the structure and distribution of sound systems among the world's languages. This is accomplished by surveying and analyzing the relative frequencies of different phonological properties. These relative frequencies might, for example, be used to determine why contrastive voicing commonly occurs with plosives, as in English *neat* and *need*, but occurs much more rarely among fricatives, such as the English *niece* and *knees*. According to a worldwide sample of 637 languages, 62% have the voicing contrast in stops but only 35% have this in fricatives. In the vast majority of those cases, the absence of voicing contrast occurs because there is a lack of voiced fricatives and because all languages have some form of plosive, but there are languages with no fricatives. Below is a chart showing the breakdown of voicing properties among languages in the aforementioned sample.

Languages worldwide also vary in the number of sounds they use. These languages can go from very small phonemic

inventories (Rotokas with six consonants and five vowels) to very large inventories (!Xóõ with 128 consonants and 28 vowels). An interesting phonological observation found with this data is that the larger a consonant inventory a language has, the more likely it is to contain a sound from a defined set of complex consonants (clicks, glottalized consonants, doubly articulated labial-velar stops, lateral fricatives and affricates, uvular and pharyngeal consonants, and dental or alveolar non-sibilant fricatives). Of this list, only about 26% of languages in a survey of over 600 with small inventories (less than 19 consonants) contain a member of this set, while 51% of average languages (19-25) contain at least one member and 69% of large consonant inventories (greater than 25 consonants) contain a member of this set. It is then seen that complex consonants are in proportion to the size of the inventory.

Vowels contain a more modest number of phonemes, with the average being 5–6, which 51% of the languages in the survey have. About a third of the languages have larger than average vowel inventories. Most interesting though is the lack of relationship between consonant inventory size and vowel inventory size. Below is a chart showing this lack of predictability between consonant and vowel inventory sizes in relation to each other.

Quantitative typology

Quantitative typology deals with the distribution and co-occurrence of structural patterns in the languages of the world. Major types of non-chance distribution include:

- preferences (for instance, absolute and implicational universals, semantic maps, and hierarchies)
- correlations (for instance, areal patterns, such as with a Sprachbund)

Linguistic universals are patterns that can be seen cross-linguistically. Universals can either be absolute, meaning that every documented language exhibits this characteristic, or statistical, meaning that this characteristic is seen in most languages or is probable in most languages. Universals, both absolute and statistical can be unrestricted, meaning that they apply to most or all languages without any additional conditions. Conversely, both absolute and statistical universals can be restricted or implicational, meaning that a characteristic will be true on the condition of something else (if Y characteristic is true, then X characteristic is true).

Language contact

Language contact occurs when speakers of two or more languages or varieties interact and influence each other. The study of language contact is called**contact linguistics**. When speakers of different languages interact closely, it is typical for their languages to influence each other. Language contact can occur at language borders, between adstratum languages, or as the result of migration, with an intrusive language acting as either a superstratum or a substratum.

Language contact occurs in a variety of phenomena, including language convergence, borrowing and relexification. The common products include pidgins, creoles, code-switching, and mixed languages. Other hybrid languages, such as English, do

not strictly fit into any of these categories. In many other cases, contact between speakers occurs but the lasting effects on the language are less visible; they may, however, include loan words, calques or other types of borrowed material.

Multilingualism has likely been common throughout much of human history, and today most people in the world are multilingual.

Borrowing of vocabulary

The most common way that languages influence each other is the exchange of words. Much is made about the contemporary borrowing of English words into other languages, but this phenomenon is not new, and it is not very large by historical standards. The large-scale importation of words from Latin, French and other languages into English in the 16th and the 17th centuries was more significant.

Some languages have borrowed so much that they have become scarcely recognisable. Armenian borrowed so many words from Iranian languages, for example, that it was at first considered a divergent branch of the Indo-Iranian languages and was not recognised as an independent branch of the Indo-European languages for many decades.

Adoption of other language features

The influence can go deeper, extending to the exchange of even basic characteristics of a language such as morphology and grammar.

Newar, for example, spoken in Nepal, is a Sino-Tibetan language distantly related to Chinese but has had so many centuries of contact with neighbouring Indo-Iranian languages that it has even developed noun inflection, a trait that is typical of the Indo-European family but rare in Sino-Tibetan. Newar has also absorbed grammatical features like verb tenses.

Also, Romanian was influenced by the Slavic languages that were spoken by neighbouring tribes in the centuries after the fall of the Roman Empire in vocabulary but even phonology. English has a few phrases, adapted from French, in which the adjective follows the noun: court-martial, attorney-general, Lake Superior.

It is easy to see how a word can diffuse from one language to another, but it is not as obvious how more basic features can do the same even if the latter phenomenon is common.

Language shift

The result of the contact of two languages can be the replacement of one by the other. This is most common when one language has a higher social position (prestige). This sometimes leads to language endangerment or extinction.

Stratal influence

When language shift occurs, the language that is replaced (known as the substratum) can leave a profound impression on the replacing language (known as the superstratum) when people retain features of the substratum as they learn the new

language and pass these features on to their children, which leads to the development of a new variety. For example, the Latin that came to replace local languages in present-day France during Ancient Rome times was influenced by Gaulish and Germanic. The distinct pronunciation of the Hiberno-English dialect, spoken in Ireland, comes partially from the influence of the substratum of Irish.

Outside the Indo-European family, Coptic, the last stage of ancient Egyptian, is a substratum of Egyptian Arabic.

Creation of new languages: creolization and mixed languages

Language contact can also lead to the development of new languages when people without a common language interact closely. Resulting from this contact a pidgin may develop, which may eventually become a full-fledged creole language through the process of creolization (though some linguists assert that a creole need not emerge from a pidgin). Prime examples of this are Aukan and Saramaccan, spoken in Suriname, which have vocabulary mainly from Portuguese, English and Dutch.

A much rarer but still observed process, according to some linguists, is the formation of mixed languages. Whereas creoles are formed by communities lacking a common language, mixed languages are formed by communities fluent in both languages. They tend to inherit much more of the complexity (grammatical, phonological, etc.) of their parent languages, whereas creoles begin as simple languages and then develop in

complexity more independently. It is sometimes explained as bilingual communities that no longer identify with the cultures of either of the languages they speak, and seek to develop their own language as an expression of their own cultural uniqueness.

Mutual and non-mutual influence

Change as a result of contact is often one-sided. Chinese, for instance, has had a profound effect on the development of Japanese, but Chinese remains relatively free of Japanese influence other than some modern terms that were reborrowed after they were coined in Japan and based on Chinese forms and using Chinese characters. In India, Hindi and other native languages have been influenced by English, and loanwords from English are part of everyday vocabulary.

In some cases, language contact may lead to mutual exchange, but that may be confined to a particular geographic region. For example, in Switzerland, the local French has been influenced by German and vice versa. In Scotland, Scots has been heavily influenced by English, and many Scots terms have been adopted into the regional English dialect.

Linguistic hegemony

A language's influence widens as its speakers grow in power. Chinese, Greek, Latin, Portuguese, French, Spanish, Arabic, Persian, Sanskrit, Russian, German and English have each seen periods of widespread importance and have had varying degrees of influence on the native languages spoken in the areas over which they have held sway.

Especially during and since the 1990s, the internet, along with previous influences such as radio and television, telephone communication and printed materials, has expanded and changed the many ways in which languages can be influenced by each other and by technology.

Dialectal and sub-cultural change

Some forms of language contact affect only a particular segment of a speech community. Consequently, change may be manifested only in particular dialects, jargons, or registers. South African English, for example, has been significantly affected by Afrikaans in terms of lexis and pronunciation, but the other dialects of English have remained almost totally unaffected by Afrikaans other than a few loanwords.

In some cases, a language develops an acrolect that contains elements of a more prestigious language. For example, in England during a large part of the Middle Ages, upper-class speech was dramatically influenced by French to the point that it often resembled a French dialect.

Methods from sociolinguistics, the study of language use in society, are used effectively in the study of language contact in communities. The broader study of contact varieties within a society is calledlinguistic ecology.

Sign languages

Language contact is extremely common in most deaf communities, which are almost always located within a dominant oral language culture. It can also take place between

two or more sign languages, and the expected contact phenomena occur: lexical borrowing, foreign "accent", interference, code switching, pidgins, creoles, and mixed systems. However, between a sign language and an oral language, even if lexical borrowing and code switching also occur, the interface between the oral and signed modes produces unique phenomena: fingerspelling, fingerspelling/sign combination, initialisation, CODA talk, TDD conversation, mouthing and contact signing.

Contrastive analysis

Contrastive analysis is the systematic study of a pair of languages with a view to identifying their structural differences and similarities. Historically it has been used to establish language genealogies.

Second language acquisition

Contrastive analysis was used extensively in the field of second language acquisition (SLA) in the 1960s and early 1970s, as a method of explaining why some features of a target language were more difficult to acquire than others. According to the behaviourist theories prevailing at the time, language learning was a question of habit formation, and this could be reinforced or impeded by existing habits.

Therefore, the difficulty in mastering certain structures in a second language (L2) depended on the difference between the learners' mother language (L1) and the language they were trying to learn.

History

The theoretical foundations for what became known as the contrastive analysis hypothesis were formulated in Robert Lado's *Linguistics Across Cultures* (1957). In this book, Lado claimed that "those elements which are similar to [the learner's] native language will be simple for him, and those elements that are different will be difficult". While it was not a novel suggestion, Lado was the first to provide a comprehensive theoretical treatment and to suggest a systematic set of technical procedures for the contrastive study of languages. That involved describing the languages (using structuralist linguistics), comparing them and predicting learning difficulties.

During the 1960s, there was a widespread enthusiasm with this technique, manifested in the contrastive descriptions of several European languages, many of which were sponsored by the Center for Applied Linguistics in Washington, DC. It was expected that once the areas of potential difficulty had been mapped out through contrastive analysis, it would be possible to design language courses more efficiently. Contrastive analysis, along with behaviourism and structuralism exerted a profound effect on SLA curriculum design and language teacher education, and provided the theoretical pillars of the audio-lingual method.

Criticism and its response

In its strongest formulation, the contrastive analysis hypothesis claimed that all the errors made in learning the L2 could be attributed to 'interference' by the L1. However, this

claim could not be sustained by empirical evidence that was accumulated in the mid- and late 1970s. It was soon pointed out that many errors predicted by Contrastive Analysis were inexplicably not observed in learners' language. Even more confusingly, some uniform errors were made by learners irrespective of their L1. It thus became clear that contrastive analysis could not predict all learning difficulties, but was certainly useful in the retrospective explanation of errors.

In response to the above criticisms, a moderate version of the Contrastive Analysis Hypothesis (CAH) has developed which paradoxically contradicts Lado's original claim. The new CAH hypothesizes that the more different the L2 is with one's L1, the easier it is for one to learn the target language. The prediction is based on the premise that similarities in languages create confusion for learners.

With the help of technological advancement, contrastive analysis has adopted a more efficient method in obtaining language data, a corpus-based approach, which generates vast amount of juxtapositions of language differences in various fields of linguistics, for example lexis and syntax.

Applications

There are multiple fields in the realms of linguistics to which Contrastive Analysis (CA) is applicable:

- Historical linguistics, a former application of CA, which is subsumed under the name comparative linguistics, a branch in linguistics not to be confused with CA.

- Second language teaching: Despite CA's limitation in the prediction of L2 learners' errors, it provides insights to at least some of the major mistakes that are frequently made by L2 learners irrespective of their L1. Hence, more tailor-made language design can be adopted; examples include awareness raising teaching method and hierarchical learning teaching curriculum.
- Second language learning: Awareness raising is the major contribution of CA in second language learning. This includes CA's abilities to explain observed errors and to outline the differences between two languages; upon language learners' realization of these aspects, they can work to adopt a viable way to learn instead of rote learning, and correct fossilized language errors.
- Sociolinguistics, psycholinguistics, bilingualism, pragmatics and others cultural-related areas: CA is, in itself, a cross-linguistic/cross-cultural study, and its ability to apply both linguistic and non-linguistic features is one of its major merits. This permits a better linguistic-cultural understanding, which is essential for learning a language in its entirety.
- Translation: CA provides better understanding of linguistic difference between two languages and therefore may be applied to the field of translation. Primarily, CA certainly lays a foundation for translation as it is integral that translators and interpreters have a thorough understanding of not only the languages they work between, but of the differences between them as well. Also, it might balance the word-for-word vs. sense-for sense debate

by developing strategies to overcome the linguistic hindrance. Moreover, it may avoid awkward translations such as translationese and Europeanization.

- Language therapy: Distinguishing the difference between language disorder patients from non-standard dialect speakers. This is essential in identifying speech pathology and their corresponding treatment.
- Criminal investigation: CA research offers insight to subtle differences among languages. Language patterns can be used as clues to investigate criminal activities, for example analyzing phishing texts designed to deceive users into giving away confidential information.

Chapter 3

Comparative Method

In linguistics, the **comparative method** is a technique for studying the development of languages by performing a feature-by-feature comparison of two or more languages with common descent from a shared ancestor and then extrapolating backwards to infer the properties of that ancestor. The comparative method may be contrasted with the method of internal reconstruction in which the internal development of a single language is inferred by the analysis of features within that language. Ordinarily, both methods are used together to reconstruct prehistoric phases of languages; to fill in gaps in the historical record of a language; to discover the development of phonological, morphological and other linguistic systems and to confirm or to refute hypothesised relationships between languages.

The comparative method was developed over the 19th century. Key contributions were made by the Danish scholars Rasmus Rask and Karl Verner and the German scholar Jacob Grimm. The first linguist to offer reconstructed forms from a proto-language was August Schleicher, in his *Compendium der vergleichenden Grammatik der indogermanischen Sprachen*, originally published in 1861. Here is Schleicher's explanation of why he offered reconstructed forms:

In the present work an attempt is made to set forth the inferred Indo-European original language side by side with its really existent derived languages. Besides the advantages

offered by such a plan, in setting immediately before the eyes of the student the final results of the investigation in a more concrete form, and thereby rendering easier his insight into the nature of particular Indo-European languages, there is, I think, another of no less importance gained by it, namely that it shows the baselessness of the assumption that the non-Indian Indo-European languages were derived from Old-Indian (Sanskrit).

Definition

Principles

The aim of the comparative method is to highlight and interpret systematic phonological and semantic correspondences between two or more attested languages. If those correspondences cannot be rationally explained as the result of language contact (borrowings, areal influence, etc.), and if they are sufficiently numerous and systematic that they cannot be dismissed as chance similarities, then it must be assumed that they descend from a single proto-language.

A sequence of regular sound changes (along with their underlying sound laws) can then be postulated to explain the correspondences between the attested forms, which eventually allows for the reconstruction of a proto-language by the methodical comparison of "linguistic facts" within a generalized system of correspondences.

Every linguistic fact is part of a whole in which everything is connected to everything else. One detail must not be linked to another detail, but one linguistic system to another.

- *— Antoine Meillet, La méthode comparative en linguistique historique, 1966 [1925], pp. 12–13.*

Relation is deemed certain only if at least a partial reconstruction of the common ancestor is feasible, and regular sound correspondences can be established, with chance similarities ruled out.

Terminology

*Descent*is defined as transmission across the generations: children learn a language from the parents' generation and, after being influenced by their peers, transmit it to the next generation, and so on. For example, a continuous chain of speakers across the centuries links Vulgar Latin to all of its modern descendants.

Two languages are *genetically related* if they descended from the same ancestor language. For example, Italian and French both come from Latin and therefore belong to the same family, the Romance languages. Having a large component of vocabulary from a certain origin is not sufficient to establish relatedness; for example, heavy borrowing from Arabic into Persian has caused more of the vocabulary of Modern Persian to be from Arabic than from the direct ancestor of Persian, Proto-Indo-Iranian, but Persian remains a member of the Indo-Iranian family and is not considered "related" to Arabic.

However, it is possible for languages to have different degrees of relatedness. English, for example, is related to both German and Russian but is more closely related to the former than to the latter. Although all three languages share a common ancestor, Proto-Indo-European, English and German also share

a more recent common ancestor, Proto-Germanic, but Russian does not. Therefore, English and German are considered to belong to a different subgroup, the Germanic languages.

Shared retentions from the parent language are not sufficient evidence of a sub-group. For example, German and Russian both retain from Proto-Indo-European a contrast between the dative case and the accusative case, which English has lost. However, that similarity between German and Russian is not evidence that German is more closely related to Russian than to English but means only that the *innovation* in question, the loss of the accusative/dative distinction, happened more recently in English than the divergence of English from German. The division of related languages into sub-groups is accomplished more certainly by finding *shared linguistic innovations* that differentiate them from the parent language, rather than shared features that are retained from the parent language.

Origin and development

In Antiquity, Romans were aware of the similarities between Greek and Latin, but did not study them systematically. They sometimes explained them mythologically, as the result of Rome being a Greek colony speaking a debased dialect.

Even though grammarians of Antiquity had access to other languages around them (Oscan, Umbrian, Etruscan, Gaulish, Egyptian, Parthian...), they showed little interest in comparing, studying, or just documenting them. Comparison between languages really began after Antiquity.

Early works

In the 9th or 10th century AD, Yehuda Ibn Quraysh compared the phonology and morphology of Hebrew, Aramaic and Arabic but attributed the resemblance to the Biblical story of Babel, with Abraham, Isaac and Joseph retaining Adam's language, with other languages at various removes becoming more altered from the original Hebrew.

In publications of 1647 and 1654, Marcus van Boxhorn first described a rigorous methodology for historical linguistic comparisons and proposed the existence of an Indo-European proto-language, which he called "Scythian", unrelated to Hebrew but ancestral to Germanic, Greek, Romance, Persian, Sanskrit, Slavic, Celtic and Baltic languages. The Scythian theory was further developed by Andreas Jäger (1686) and William Wotton (1713), who made early forays to reconstruct the primitive common language. In 1710 and 1723, Lambert ten Kate first formulated the regularity of sound laws, introducing among others the term root vowel.

Another early systematic attempt to prove the relationship between two languages on the basis of similarity of grammar and lexicon was made by the Hungarian János Sajnovics in 1770, when he attempted to demonstrate the relationship between Sami and Hungarian. That work was later extended to all Finno-Ugric languages in 1799 by his countrymanSamuel Gyarmathi. However, the origin of modern historical linguisticsis often traced back to Sir William Jones, an English philologist living in India, who in 1786 made his famous observation:

The Sanscrit language, whatever be its antiquity, is of a wonderful structure; more perfect than the Greek, more copious than the Latin, and more exquisitely refined than either, yet bearing to both of them a stronger affinity, both in the roots of verbs and the forms of grammar, than could possibly have been produced by accident; so strong indeed, that no philologer could examine them all three, without believing them to have sprung from some common source, which, perhaps, no longer exists. There is a similar reason, though not quite so forcible, for supposing that both the Gothick and the Celtick, though blended with a very different idiom, had the same origin with the Sanscrit; and the old Persian might be added to the same family.

Comparative linguistics

The comparative method developed out of attempts to reconstruct the proto-language mentioned by Jones, which he did not name but subsequent linguists have labelled Proto-Indo-European (PIE). The first professional comparison between the Indo-European languages that were then known was made by the German linguist Franz Bopp in 1816. He did not attempt a reconstruction but demonstrated that Greek, Latin and Sanskrit shared a common structure and a common lexicon. In 1808, Friedrich Schlegel first stated the importance of using the eldest possible form of a language when trying to prove its relationships; in 1818, Rasmus Christian Rask developed the principle of regular sound-changes to explain his observations of similarities between individual words in the Germanic languages and their cognates in Greek and Latin.Jacob Grimm, better known for his *Fairy Tales*, used the comparative method in *Deutsche Grammatik* (published 1819–

1837 in four volumes), which attempted to show the development of the Germanic languages from a common origin, which was the first systematic study of diachronic language change.

Both Rask and Grimm were unable to explain apparent exceptions to the sound laws that they had discovered. Although Hermann Grassmann explained one of the anomalies with the publication of Grassmann's law in 1862, Karl Verner made a methodological breakthrough in 1875, when he identified a pattern now known as Verner's law, the first sound-law based on comparative evidence showing that a phonological change in one phoneme could depend on other factors within the same word (such as neighbouring phonemes and the position of the accent), which is now called *conditioning environments*.

Neo-grammarian approach

Similar discoveries made by the *Junggrammatiker* (usually translated as "Neogrammarians") at the University of Leipzig in the late 19th century led them to conclude that all sound changes were ultimately regular, resulting in the famous statement by Karl Brugmann and Hermann Osthoff in 1878 that "sound laws have no exceptions". That idea is fundamental to the modern comparative method since it necessarily assumes regular correspondences between sounds in related languages and thus regular sound changes from the proto-language. The *Neogrammarian hypothesis* led to the application of the comparative method to reconstruct Proto-Indo-European since Indo-European was then by far the most well-studied language family. Linguists working with other

families soon followed suit, and the comparative method quickly became the established method for uncovering linguistic relationships.

Application

There is no fixed set of steps to be followed in the application of the comparative method, but some steps are suggested by Lyle Campbell and Terry Crowley, who are both authors of introductory texts in historical linguistics. This abbreviated summary is based on their concepts of how to proceed.

Step 1, assemble potential cognate lists

This step involves making lists of words that are likely cognates among the languages being compared. If there is a regularly-recurring match between the phonetic structure of basic words with similar meanings, a genetic kinship can probably then be established. For example, linguists looking at the Polynesian family might come up with a list similar to the following (their actual list would be much longer):

Borrowings or false cognates can skew or obscure the correct data. For example, English *taboo* ([tæbu]) is like the six Polynesian forms because of borrowing from Tongan into English, not because of a genetic similarity. That problem can usually be overcome by using basic vocabulary, such as kinship terms, numbers, body parts and pronouns. Nonetheless, even basic vocabulary can be sometimes borrowed. Finnish, for example, borrowed the word for "mother", *äiti*, from Proto-Germanic *aiþį̄ (compare to Gothic*aiþei*). English borrowed the pronouns "they", "them",

and "their(s)" from Norse. Thai and various other East Asian languages borrowed their numbers from Chinese. An extreme case is represented by Pirahã, a Muran language of South America, which has been controversially claimed to have borrowed all of its pronouns from Nheengatu.

Step 2, establish correspondence sets

The next step involves determining the regular sound-correspondences exhibited by the lists of potential cognates. For example, in the Polynesian data above, it is apparent that words that contain *t* in most of the languages listed have cognates in Hawaiian with *k* in the same position. That is visible in multiple cognate sets: the words glossed as 'one', 'three', 'man' and 'taboo' all show the relationship. The situation is called a "regular correspondence" between *k* in Hawaiian and *t* in the other Polynesian languages. Similarly, a regular correspondence can be seen between Hawaiian and Rapanui *h*, Tongan and Samoan *f*, Maori *ɸ*, and Rarotongan *ʔ*.

Mere phonetic similarity, as between English*day* and Latin*dies* (both with the same meaning), has no probative value. English initial *d-* does not *regularly* match Latin *d-* since a large set of English and Latin non-borrowed cognates cannot be assembled such that English *d* repeatedly and consistently corresponds to Latin *d* at the beginning of a word, and whatever sporadic matches can be observed are due either to chance (as in the above example) or to borrowing (for example, Latin *diabolus* and English *devil*, both ultimately of Greek origin). However, English and Latin exhibit a regular correspondence of *t-* : *d-* (in which "A : B" means "A corresponds to B"), as in the following examples:

Step 3, discover which sets are in complementary distribution

During the late 18th to late 19th century, two major developments improved the method's effectiveness.

First, it was found that many sound changes are conditioned by a specific *context*. For example, in both Greek and Sanskrit, an aspiratedstop evolved into an unaspirated one, but only if a second aspirate occurred later in the same word; this is Grassmann's law, first described for Sanskrit by Sanskrit grammarianPā□ini and promulgated by Hermann Grassmann in 1863.

Second, it was found that sometimes sound changes occurred in contexts that were later lost. For instance, in Sanskrit velars (*k*-like sounds) were replaced by palatals (*ch*-like sounds) whenever the following vowel was **i* or **e*. Subsequent to this change, all instances of **e* were replaced by *a*. The situation could be reconstructed only because the original distribution of *e* and *a* could be recovered from the evidence of other Indo-European languages. For instance, the Latin suffix *que*, "and", preserves the original **e* vowel that caused the consonant shift in Sanskrit:

Verner's Law, discovered by Karl Verner c. 1875, provides a similar case: the voicing of consonants in Germanic languages underwent a change that was determined by the position of the old Indo-European accent. Following the change, the accent shifted to initial position. Verner solved the puzzle by comparing the Germanic voicing pattern with Greek and Sanskrit accent patterns.

This stage of the comparative method, therefore, involves examining the correspondence sets discovered in step 2 and seeing which of them apply only in certain contexts. If two (or more) sets apply in complementary distribution, they can be assumed to reflect a single original phoneme: "some sound changes, particularly conditioned sound changes, can result in a proto-sound being associated with more than one correspondence set".

Since French *ʃ* occurs only before *a* where the other languages also have *a*, and French *k* occurs elsewhere, the difference is caused by different environments (being before *a* conditions the change), and the sets are complementary. They can, therefore, be assumed to reflect a single proto-phoneme (in this case **k*, spelled |c| in Latin). The original Latin words are *corpus*, *crudus*, *catena* and *captiare*, all with an initial *k*. If more evidence along those lines were given, one might conclude that an alteration of the original *k* took place because of a different environment.

A more complex case involves consonant clusters in Proto-Algonquian. The Algonquianist Leonard Bloomfield used the reflexes of the clusters in four of the daughter languages to reconstruct the following correspondence sets:

Although all five correspondence sets overlap with one another in various places, they are not in complementary distribution and so Bloomfield recognised that a different cluster must be reconstructed for each set. His reconstructions were, respectively, **hk*, **xk*, **čk* (=[tʃk]), **šk* (=[ʃk]), and *çk* (in which '*x*' and '*ç*' are arbitrary symbols, rather than attempts to guess the phonetic value of the proto-phonemes).

Step 4, reconstruct proto-phonemes

Typology assists in deciding what reconstruction best fits the data. For example, the voicing of voiceless stops between vowels is common, but the devoicing of voiced stops in that environment is rare. If a correspondence *-t-* : *-d-* between vowels is found in two languages, the proto-phoneme is more likely to be **-t-*, with a development to the voiced form in the second language. The opposite reconstruction would represent a rare type.

However, unusual sound changes occur. The Proto-Indo-European word for *two*, for example, is reconstructed as **dwō*, which is reflected in Classical Armenian as *erku*. Several other cognates demonstrate a regular change **dw-→erk-* in Armenian. Similarly, in Bearlake, a dialect of the Athabaskan language of Slavey, there has been a sound change of Proto-Athabaskan **ts→* Bearlake k^w. It is very unlikely that **dw-* changed directly into *erk-* and **ts* into k^w, but they probably instead went through several intermediate steps before they arrived at the later forms. It is not phonetic similarity that matters for the comparative method but rather regular sound correspondences.

By the principle of economy, the reconstruction of a proto-phoneme should require as few sound changes as possible to arrive at the modern reflexes in the daughter languages. For example, Algonquian languages exhibit the following correspondence set:

An earlier voiceless aspirated row was removed on grounds of insufficient evidence. Since the mid-20th century, a number of

linguists have argued that this phonology is implausible and that it is extremely unlikely for a language to have a voiced aspirated (breathy voice) series without a corresponding voiceless aspirated series.

Thomas Gamkrelidze and Vyacheslav Ivanov provided a potential solution and argued that the series that are traditionally reconstructed as plain voiced should be reconstructed as glottalized: either implosive(ɓ, ɗ, ɠ) or ejective(p', t', k'). The plain voiceless and voiced aspirated series would thus be replaced by just voiceless and voiced, with aspiration being a non-distinctive quality of both. That example of the application of linguistic typology to linguistic reconstruction has become known as the glottalic theory. It has a large number of proponents but is not generally accepted.

The reconstruction of proto-sounds logically precedes the reconstruction of grammatical morphemes (word-forming affixes and inflectional endings), patterns of declension and conjugation and so on. The full reconstruction of an unrecorded protolanguage is an open-ended task.

Complications

The history of historical linguistics

The limitations of the comparative method were recognized by the very linguists who developed it, but it is still seen as a valuable tool. In the case of Indo-European, the method seemed at least a partial validation of the centuries-old search for an Ursprache, the original language. The others were

presumed to be ordered in a family tree, which was the tree model of the neogrammarians.

The archaeologists followed suit and attempted to find archaeological evidence of a culture or cultures that could be presumed to have spoken a proto-language, such as Vere Gordon Childe's *The Aryans: a study of Indo-European origins*, 1926. Childe was a philologist turned archaeologist. Those views culminated in the *Siedlungsarchaologie*, or "settlement-archaeology", of Gustaf Kossinna, becoming known as "Kossinna's Law". Kossinna asserted that cultures represent ethnic groups, including their languages, but his law was rejected after World War II. The fall of Kossinna's Law removed the temporal and spatial framework previously applied to many proto-languages. Fox concludes:

The Comparative Method *as such* is not, in fact, historical; it provides evidence of linguistic relationships to which we may give a historical interpretation.... [Our increased knowledge about the historical processes involved] has probably made historical linguists less prone to equate the idealizations required by the method with historical reality.... Provided we keep [the interpretation of the results and the method itself] apart, the Comparative Method can continue to be used in the reconstruction of earlier stages of languages.

Proto-languages can be verified in many historical instances, such as Latin. Although no longer a law, settlement-archaeology is known to be essentially valid for some cultures that straddle history and prehistory, such as the Celtic Iron Age (mainly Celtic) and Mycenaean civilization (mainly Greek).

None of those models can be or have been completely rejected, but none is sufficient alone.

The Neogrammarian principle

The foundation of the comparative method, and of comparative linguistics in general, is the Neogrammarians' fundamental assumption that "sound laws have no exceptions". When it was initially proposed, critics of the Neogrammarians proposed an alternate position that summarised by the maxim "each word has its own history". Several types of change actually alter words in irregular ways. Unless identified, they may hide or distort laws and cause false perceptions of relationship.

Borrowing

All languages borrow words from other languages in various contexts. They are likely to have followed the laws of the languages from which they were borrowed, rather than the laws of the borrowing language. Therefore, studying borrowed words will probably mislead the investigator since they reflect the customs of the donor language, which is the source of the word.

Areal diffusion

Borrowing on a larger scale occurs in areal diffusion, when features are adopted by contiguous languages over a geographical area. The borrowing may be phonological, morphological or lexical. A false proto-language over the area may be reconstructed for them or may be taken to be a third language serving as a source of diffused features.

Several areal features and other influences may converge to form a Sprachbund, a wider region sharing features that appear to be related but are diffusional. For instance, the Mainland Southeast Asia linguistic area, before it was recognised, suggested several false classifications of such languages as Chinese, Thai and Vietnamese.

Random mutations

Sporadic changes, such as irregular inflections, compounding and abbreviation, do not follow any laws. For example, the Spanish words *palabra* ('word'), *peligro* ('danger') and *milagro* ('miracle') would have been *parabla*, *periglo*, *miraglo* by regular sound changes from the Latin *parabŏla*, *perīcŭlum* and *mīrācŭlum*, but the *r* and *l* changed places by sporadic metathesis.

Analogy

Analogy is the sporadic change of a feature to be like another feature in the same or a different language. It may affect a single word or be generalized to an entire class of features, such as a verb paradigm. An example is the Russian word for *nine*. The word, by regular sound changes from Proto-Slavic, should have been /nʲevʲatʲ/, but it is in fact /dʲevʲatʲ/. It is believed that the initial *nʲ-* changed to *dʲ-* under influence of the word for "ten" in Russian, /dʲesʲatʲ/.

Gradual application

Those who study contemporary language changes, such as William Labov, acknowledge that even a systematic sound

change is applied at first inconsistently, with the percentage of its occurrence in a person's speech dependent on various social factors. The sound change seems to gradually spread in a process known as lexical diffusion. While it does not invalidate the Neogrammarians' axiom that "sound laws have no exceptions", the gradual application of the very sound laws shows that they do not always apply to all lexical items at the same time. Hock notes, "While it probably is true in the long run every word has its own history, it is not justified to conclude as some linguists have, that therefore the Neogrammarian position on the nature of linguistic change is falsified".

Non-inherited features

The comparative method cannot recover aspects of a language that were not inherited in its daughter idioms. For instance, the Latin declension pattern was lost in Romance languages, resulting in an impossibility to fully reconstruct such a feature via systematic comparison.

The presumption of a well-defined node

The tree model features nodes that are presumed to be distinct proto-languages existing independently in distinct regions during distinct historical times. The reconstruction of unattested proto-languages lends itself to that illusion since they cannot be verified, and the linguist is free to select whatever definite times and places seems best. Right from the outset of Indo-European studies, however, Thomas Young said:

It is not, however, very easy to say what the definition should be that should constitute a separate language, but it seems most natural to call those languages distinct, of which the one cannot be understood by common persons in the habit of speaking the other.... Still, however, it may remain doubtfull whether the Danes and the Swedes could not, in general, understand each other tolerably well... nor is it possible to say if the twenty ways of pronouncing the sounds, belonging to the Chinese characters, ought or ought not to be considered as so many languages or dialects.... But,... the languages so nearly allied must stand next to each other in a systematic order...

The assumption of uniformity in a proto-language, implicit in the comparative method, is problematic. Even small language communities are always have differences in dialect, whether they are based on area, gender, class or other factors. The Pirahã language of Brazilis spoken by only several hundred people but has at least two different dialects, one spoken by men and one by women. Campbell points out:

It is not so much that the comparative method 'assumes' no variation; rather, it is just that there is nothing built into the comparative method which would allow it to address variation directly.... This assumption of uniformity is a reasonable idealization; it does no more damage to the understanding of the language than, say, modern reference grammars do which concentrate on a language's general structure, typically leaving out consideration of regional or social variation.

Different dialects, as they evolve into separate languages, remain in contact with and influence one another. Even after they are considereddistinct, languages near one another

continue to influence one another and often share grammatical, phonological, and lexical innovations. A change in one language of a family may spread to neighboring languages, and multiple waves of change are communicated like waves across language and dialect boundaries, each with its own randomly delimited range. If a language is divided into an inventory of features, each with its own time and range (isoglosses), they do not all coincide.

History and prehistory may not offer a time and place for a distinct coincidence, as may be the case for Proto-Italic, for which the proto-language is only a concept. However, Hock observes:

The discovery in the late nineteenth century that isoglosses can cut across well-established linguistic boundaries at first created considerable attention and controversy. And it became fashionable to oppose a wave theory to a tree theory.... Today, however, it is quite evident that the phenomena referred to by these two terms are complementary aspects of linguistic change....

Subjectivity of the reconstruction

The reconstruction of unknown proto-languages is inherently subjective. In the Proto-Algonquian example above, the choice of **m* as the parent phoneme is only *likely*, not *certain*. It is conceivable that a Proto-Algonquian language with **b* in those positions split into two branches, one that preserved **b* and one that changed it to **m* instead, and while the first branch developed only into Arapaho, the second spread out more widely and developed into all the other Algonquian tribes. It is

also possible that the nearest common ancestor of the Algonquian languages used some other sound instead, such as **p*, which eventually mutated to **b* in one branch and to **m* in the other.

Examples of strikingly complicated and even circular developments are indeed known to have occurred (such as Proto-Indo-European **t*> Pre-Proto-Germanic **þ*>Proto-Germanic**ð*> Proto-West-Germanic **d*>Old High German*t* in *fater*> Modern German *Vater*), but in the absence of any evidence or other reason to postulate a more complicated development, the preference of a simpler explanation is justified by the principle of parsimony, also known as Occam's razor. Since reconstruction involves many such choices, some linguists prefer to view the reconstructed features as abstract representations of sound correspondences, rather than as objects with a historical time and place.

The existence of proto-languages and the validity of the comparative method is verifiable if the reconstruction can be matched to a known language, which may be known only as a shadow in the loanwords of another language. For example, Finnic languages such as Finnish have borrowed many words from an early stage of Germanic, and the shape of the loans matches the forms that have been reconstructed for Proto-Germanic. Finnish *kuningas* 'king' and *kaunis* 'beautiful' match the Germanic reconstructions **kuningaz* and **skauniz* (> German *König* 'king', *schön* 'beautiful').

Additional models

The wave model was developed in the 1870s as an alternative to the tree model to represent the historical patterns of

language diversification. Both the tree-based and the wave-based representations are compatible with the comparative method.

By contrast, some approaches are incompatible with the comparative method, including glottochronology and mass lexical comparison, both of which are considered by most historical linguists to be flawed and unreliable.

Chapter 4

Aspects of Comparative Linguistics

Glottochronology

Glottochronology (from Attic Greekγλ□ττα*tongue, language* and χρόνος*time*) is the part of lexicostatistics which involves comparative linguistics and deals with the chronological relationship between languages.

The idea was developed by Morris Swadesh in the 1950s in his article on Salish internal relationships. He developed the idea under two assumptions: there indeed exists a relatively stable *basic vocabulary* (referred to as *Swadesh lists*) in all languages of the world; and, any replacements happen in a way analogous to radioactive decay in a constant percentage per time elapsed. Using mathematics and statistics, Swadesh developed an equation that could determine when languages separated and give an approximate time of when the separation occurred. His methods aid linguistic anthropologists by giving them a definitive way to determine a separation date between two languages. The formula he created finds an approximate number of centuries since two languages separated from a singular common ancestor. His methods also provided information on when ancient languages may have existed.

Despite multiple studies and literature containing the information of glottochronology, it is not widely used today and

is surrounded with controversy. Glottochronology tracks language separation from thousands, if not millions of years ago but many linguists are skeptical of the concept because it is more of a 'probability' rather than a 'certainty.' On the other hand, some linguists may say that glottochronology is gaining traction because of its relatedness to archaeological dates. Glottochronology is not as accurate as archaeological data but it can provide a solid estimate.

Over time many different extensions of the Swadesh method evolved; however, Swadesh's original method is so well known that 'glottochronology' is usually associated with him.

Methodology

Word list

The original method of glottochronology presumed that the core vocabulary of a language is replaced at a constant (or constant average) rate across all languages and cultures and so can be used to measure the passage of time. The process makes use of a list of lexical terms and morphemes which are similar to multiple languages.

Lists were compiled by Morris Swadesh and assumed to be resistant against borrowing (originally designed in 1952 as a list of 200 items, but the refined 100-word list in Swadesh (1955) is much more common among modern day linguists). The core vocabulary was designed to encompass concepts common to every human language such as personal pronouns, body parts, heavenly bodies and living beings, verbs of basic actions, numerals, basic adjectives, kin terms, and natural

occurrences and events. Through a basic word list, one eliminates concepts that are specific to a particular culture or time period. It has been found through differentiating word lists that the ideal is really impossible and that the meaning set may need to be tailored to the languages being compared. Word lists are not homogenous throughout studies and they are often changed and designed to suit both languages being studied. Linguists find that it is difficult to find a word list where all words used are culturally unbiased. Many alternative word lists have been compiled by other linguists and often use fewer meaning slots. The percentage of cognates (words with a common origin) in the word lists is then measured. The larger the percentage of cognates, the more recently the two languages being compared are presumed to have separated.

Results

Glottochronology was found to work in the case of Indo-European, accounting for 87% of the variance. It is also postulated to work for Afro-Asiatic (Fleming 1973), Chinese (Munro 1978) and Amerind (Stark 1973; Baumhoff and Olmsted 1963). For Amerind, correlations have been obtained with radiocarbon dating and blood groups as well as archaeology.

The approach of Gray and Atkinson, as they state, has nothing to do with "glottochronology".

Discussion

The concept of language change is old, and its history is reviewed in Hymes (1973) and Wells (1973). In some sense, glottochronology is a reconstruction of history and can often be

closely related to archaeology. Many linguistic studies find the success of glottochronology to be found alongside archaeological data. Glottochronology itself dates back to the mid-20th century. An introduction to the subject is given in Embleton (1986) and in McMahon and McMahon (2005).

Glottochronology has been controversial ever since, partly because of issues of accuracy but also because of the question of whether its basis is sound (for example, Bergsland 1958; Bergsland and Vogt 1962; Fodor 1961; Chrétien 1962; Guy 1980). The concerns have been addressed by Dobson et al. (1972), Dyen (1973) and Kruskal, Dyen and Black (1973). The assumption of a single-word replacement rate can distort the divergence-time estimate when borrowed words are included (Thomason and Kaufman 1988).

An overview of recent arguments can be obtained from the papers of a conference held at the McDonald Institute in 2000. The presentations vary from "Why linguists don't do dates" to the one by Starostin discussed above. Since its original inception, glottochronology has been rejected by many linguists, mostly Indo-Europeanists of the school of the traditional comparative method. Criticisms have been answered in particular around three points of discussion:

- Criticism levelled against the higher stability of lexemes in Swadesh lists alone (Haarmann 1990) misses the point because a certain amount of losses only enables the computations (Sankoff 1970). The non-homogeneity of word lists often leads to lack of understanding between linguists. Linguists also have difficulties finding a completely unbiased list of

basic cultural words. it can take a long time for linguists to find a viable word list which can take several test lists to find a usable list.

- Traditional glottochronology presumes that language changes at a stable rate.
- Thus, in Bergsland & Vogt (1962), the authors make an impressive demonstration, on the basis of actual language data verifiable by extralinguistic sources, that the "rate of change" for Icelandic constituted around 4% per millennium, but for closely connected Riksmal (Literary Norwegian), it would amount to as much as 20% (Swadesh's proposed "constant rate" was supposed to be around 14% per millennium).
- That and several other similar examples effectively proved that Swadesh's formula would not work on all available material, which is a serious accusation since evidence that can be used to "calibrate" the meaning of *L* (language history recorded during prolonged periods of time) is not overwhelmingly large in the first place.
- It is highly likely that the chance of replacement is different for every word or feature ("each word has its own history", among hundreds of other sources:).
- That global assumption has been modified and downgraded to single words, even in single languages, in many newer attempts (see below).
- There is a lack of understanding of Swadesh's mathematical/statistical methods. Some linguists reject the methods in full because the statistics lead to 'probabilities' when linguists trust 'certainties' more.

- A serious argument is that language change arises from socio-historical events that are, of course, unforeseeable and, therefore, uncomputable.
- New methods developed by Gray & Atkinson are claimed to avoid those issues but are still seen as controversial, primarily since they often produce results that are incompatible with known data and because of additional methodological issues.

Modifications

Somewhere in between the original concept of Swadesh and the rejection of glottochronology in its entirety lies the idea that glottochronology as a formal method of linguistic analysis becomes valid with the help of several important modifications.

Thus, inhomogeneities in the replacement rate were dealt with by Van der Merwe (1966) by splitting the word list into classes each with their own rate, while Dyen, James and Cole (1967) allowed each meaning to have its own rate. Simultaneous estimation of divergence time and replacement rate was studied by Kruskal, Dyen and Black.

Brainard (1970) allowed for chance cognation, and drift effects were introduced by Gleason (1959). Sankoff (1973) suggested introducing a borrowing parameter and allowed synonyms.

A combination of the various improvements is given in Sankoff's "Fully Parameterised Lexicostatistics".

In 1972, Sankoff in a biological context developed a model of genetic divergence of populations. Embleton (1981) derives a

simplified version of that in a linguistic context. She carries out a number of simulations using this which are shown to give good results.

Improvements in statistical methodology related to a completely different branch of science, phylogenetics; the study of changes in DNA over time sparked a recent renewed interest.

The new methods are more robust than the earlier ones because they calibrate points on the tree with known historical events and smooth the rates of change across them. As such, they no longer require the assumption of a constant rate of change (Gray & Atkinson 2003).

Starostin's method

Another attempt to introduce such modifications was performed by the Russian linguist Sergei Starostin, who had proposed the following:

- Systematic loanwords, borrowed from one language into another, are a disruptive factor and must be eliminated from the calculations; the one thing that really matters is the "native" replacement of items by items from the same language. The failure to notice that factor was a major reason in Swadesh's original estimation of the replacement rate at under 14 words from the 100-wordlist per millennium, but the real rate is much slower (around 5 or 6). Introducing that correction effectively cancels out the "Bergsland & Vogt" argument since a thorough analysis of the Riksmal data shows that its basic wordlist includes

about 15 to 16 borrowings from other Germanic languages (mostly Danish), and the exclusion of those elements from the calculations brings the rate down to the expected rate of 5 to 6 "native" replacements per millennium.

- The rate of change is not really constant but depends on the time period during which the word has existed in the language (the chance of lexeme X being replaced by lexeme Y increases in direct proportion to the time elapsed, the so-called "aging of words" is empirically understood as gradual "erosion" of the word's primary meaning under the weight of acquired secondary ones).
- Individual items on the 100 word-list have different stability rates (for instance, the word "I" generally has a much lower chance of being replaced than the word "yellow").

Time-depth estimation

The McDonald Institute hosted a conference on the issue of time-depth estimation in 2000. The published papers give an idea of the views on glottochronology at that time.

They vary from "Why linguists don't do dates" to the one by Starostin discussed above. Note that in the referenced Gray and Atkinson paper, they hold that their methods cannot be called "glottochronology" by confining this term to its original method.

Historical linguistics

Historical linguistics, also termed **diachronic linguistics**, is the scientific study of language change over time. Principal concerns of historical linguistics include:

to describe and account for observed changes in particular languages

to reconstruct the pre-history of languages and to determine their relatedness, grouping them into language families (comparative linguistics)

to develop general theories about how and why language changes

to describe the history of speech communities

to study the history of words, i.e. etymology

Historical linguistics is founded on the Uniformitarian Principle, which is defined by linguist Donald Ringe as:

Unless we can demonstrate significant changes in the conditions of language acquisition and use between some time in the unobservable past and the present,

We must assume that the same types and distributions of structures, variation, changes, etc. existed at that time in the past as in the present.

History and development

Western modern historical linguistics dates from the late-18th century. It grew out of the earlier discipline of philology, the study of ancient texts and documents dating back to antiquity.

At first, historical linguistics served as the cornerstone of comparative linguistics, primarily as a tool for linguistic reconstruction. Scholars were concerned chiefly with establishing language families and reconstructing unrecorded proto-languages, using the comparative method and internal reconstruction. The focus was initially on the well-known Indo-European languages, many of which had long written histories; scholars also studied the Uralic languages, another Eurasian language-family for which less early written material exists. Since then, there has been significant comparative linguistic work expanding outside of European languages as well, such as on the Austronesian languages and on various families of Native American languages, among many others. Comparative linguistics became only a part of a more broadly-conceived discipline of historical linguistics. For the Indo-European languages, comparative study is now a highly specialized field. Most research is being carried out on the subsequent development of these languages, in particular, the development of the modern standard varieties.

Some scholars have undertaken studies attempting to establish super-families, linking, for example, Indo-European, Uralic, and other families into Nostratic. These attempts have not met with wide acceptance. The information necessary to establish relatedness becomes less available as the time depth increases. The time-depth of linguistic methods is limited due to chance

word resemblances and variations between language groups, but a limit of around 10,000 years is often assumed. The dating of the various proto-languages is also difficult; several methods are available for dating, but only approximate results can be obtained.

Diachronic and synchronic analysis

In linguistics, a **synchronic analysis** is one that views linguistic phenomena only at a given time, usually the present, but a synchronic analysis of a historical language form is also possible. It may be distinguished from diachronic, which regards a phenomenon in terms of developments through time. Diachronic analysis is the main concern of historical linguistics; however, most other branches of linguistics are concerned with some form of synchronic analysis. The study of language change offers a valuable insight into the state of linguistic representation, and because all synchronic forms are the result of historically-evolving diachronic changes, the ability to explain linguistic constructions necessitates a focus on diachronic processes.

Initially, all of modern linguistics was historical in orientation. Even the study of modern dialects involved looking at their origins. Ferdinand de Saussure's distinction between synchronic and diachronic linguistics is fundamental to the present day organization of the discipline. Primacy is accorded to synchronic linguistics, and **diachronic linguistics** is defined as the study of successive synchronic stages. Saussure's clear demarcation, however, has had both defenders and critics.

In practice, a purely-synchronic linguistics is not possible for any period before the invention of the gramophone, as written records always lag behind speech in reflecting linguistic developments. Written records are difficult to date accurately before the development of the modern title page. Often, dating must rely on contextual historical evidence such as inscriptions, or modern technology, such as carbon dating, can be used to ascertain dates of varying accuracy.

Also, the work of sociolinguists on linguistic variation has shown synchronic states are not uniform: the speech habits of older and younger speakers differ in ways that point to language change. Synchronic variation is linguistic change in progress.

Synchronic and diachronic approaches can reach quite different conclusions. For example, a Germanic strong verb like English *sing – sang – sung* is irregular when it is viewed synchronically: the native speaker's brain processes them as learned forms, but the derived forms of regular verbs are processed quite differently, by the application of productive rules (for example, adding *-ed* to the basic form of a verb as in *walk – walked*). That is an insight of psycholinguistics, which is relevant also for language didactics, both of which are synchronic disciplines. However, a diachronic analysis shows that the strong verb is the remnant of a fully regular system of internal vowel changes, in this case the Indo-European ablaut; historical linguistics seldom uses the category "irregular verb".

The principal tools of research in diachronic linguistics are the comparative method and the method of internal reconstruction. Less-standard techniques, such as mass lexical comparison,

are used by some linguists to overcome the limitations of the comparative method, but most linguists regard them as unreliable.

The findings of historical linguistics are often used as a basis for hypotheses about the groupings and movements of peoples, particularly in the prehistoric period. In practice, however, it is often unclear how to integrate the linguistic evidence with the archaeological or genetic evidence. For example, there are numerous theories concerning the homeland and early movements of the Proto-Indo-Europeans, each with its own interpretation of the archaeological record.

Sub-fields of study

Comparative linguistics

Comparative linguistics (originally comparative philology) is a branch of historical linguistics that is concerned with comparing languages in order to establish their historical relatedness. Languages may be related by convergence through borrowing or by genetic descent, thus languages can change and are also able to cross-relate.

Genetic relatedness implies a common origin or proto-language. Comparative linguistics has the goal of constructing language families, reconstructing proto-languages, and specifying the changes that have resulted in the documented languages. To maintain a clear distinction between attested language and reconstructed forms, comparative linguists prefix an asterisk to any form that is not found in surviving texts.

Etymology

Etymology is the study of the history of words: when they entered a language, from what source, and how their form and meaning have changed over time. A word may enter a language as a loanword (as a word from one language adopted by speakers of another language), through derivational morphology by combining pre-existing elements in the language, by a hybrid of these two processes called phono-semantic matching, or in several other minor ways.

In languages with a long and detailed history, etymology makes use of philology, the study of how words change from culture to culture over time. Etymologists also apply the methods of comparative linguistics to reconstruct information about languages that are too old for any direct information (such as writing) to be known. By analyzing related languages with a technique known as the comparative method, linguists can make inferences, about their shared parent language and its vocabulary. In that way, word roots that can be traced all the way back to the origin of, for instance, the Indo-Europeanlanguage familyhave been found. Although originating in the philological tradition, much current etymological research is done in language families for which little or no early documentation is available, such as Uralic and Austronesian.

Dialectology

Dialectology is the scientific study of linguistic dialect, the varieties of a language that are characteristic of particular groups, based primarily on geographic distribution and their

associated features. This is in contrast to variations based on social factors, which are studied in sociolinguistics, or variations based on time, which are studied in historical linguistics. Dialectology treats such topics as divergence of two local dialects from a common ancestor and synchronic variation.

Dialectologists are concerned with grammatical features that correspond to regional areas. Thus, they are usually dealing with populations living in specific locales for generations without moving, but also with immigrant groups bringing their languages to new settlements.

Phonology

Phonology is a sub-field of linguistics which studies the sound system of a specific language or set of languages. Whereas phonetics is about the physical production and perception of the sounds of speech, phonology describes the way sounds function within a given language or across languages.

An important part of phonology is studying which sounds are distinctive units within a language. For example, the "p" in "pin" is aspirated, but the "p" in "spin" is not. In English these two sounds are used in complementary distribution and are not used to differentiate words so they are considered allophones of the same phoneme.

In some other languages like Thai and Quechua, the same difference of aspiration or non-aspiration differentiates words and so the two sounds (or phones) are therefore considered two distinct phonemes.

In addition to the minimal meaningful sounds (the phonemes), phonology studies how sounds alternate, such as the /p/ in English, and topics such as syllable structure, stress, accent, and intonation.

The principles of phonological theory have also been applied to the analysis of sign languages, but the phonological units do not consist of sounds. The principles of phonological analysis can be applied independently of modality because they are designed to serve as general analytical tools, not language-specific ones.

Morphology

Morphology is the study of the formal means of expression in a language; in the context of historical linguistics, how the formal means of expression change over time; for instance, languages with complex inflectional systems tend to be subject to a simplification process. This field studies the internal structure of words as a formal means of expression.

Words as units in the lexicon are the subject matter of lexicology. While words are generally accepted as being (with clitics) the smallest units of syntax, it is clear that, in most (if not all) languages, words can be related to other words by rules.

The rules understood by the speaker reflect specific patterns (or regularities) in the way words are formed from smaller units and how those smaller units interact in speech. In this way, morphology is the branch of linguistics that studies patterns of word-formation within and across languages, and attempts to formulate rules that model the knowledge of the

speakers of those languages, in the context of historical linguistics, how the means of expression change over time. See grammaticalisation.

Syntax

Syntax is the study of the principles and rules for constructing sentences in natural languages. The term *syntax*is used to refer directly to the rules and principles that govern the sentence structure of any individual language, as in "the syntax of Modern Irish". Modern researchers in syntax attempt to describe languages in terms of such rules. Many professionals in this discipline attempt to find general rules that apply to all natural languages in the context of historical linguistics, how characteristics of sentence structure in related languages changed over time. See grammaticalisation.

Rates of change and varieties of adaptation

Studies in historical linguistics often use the terms "conservative" or "innovative" to characterize the extent of change occurring in a particular language or dialect as compared with related varieties. In particular, a *conservative* variety changes relatively less than an *innovative* variety. The variations in plasticity are often related to the socio-economic situation of the language speakers. An example of an innovative dialect would be American English because of the vast number of speakers and the open interaction its speakers have with other language groups; the changes can be seen in

the terms developed for business and marketing, among other fields such as technology.

The converse of an innovative language is a conservative language, which is generally defined by its static nature and imperviousness to outside influences. Most but not all conservative languages are spoken in secluded areas that lack any other primary language speaking population.

Neither descriptive terms carries any value judgment in linguistic studies or determines any form of worthiness a language has, compared to any other language.

A particularly-conservative variety that preserves features that have long since vanished elsewhere is sometimes said to be "archaic". There are few examples of archaic language in modern society, but some have survived in set phrases or in nursery rhymes.

Evolutionary context

In terms of evolutionary theory, historical linguistics (as opposed to research into the origin of language) studies Lamarckianacquired characteristics of languages.

Intercontinental Dictionary Series

The Intercontinental Dictionary Series is a large database of topical vocabulary lists in various world languages. The general editor of the database is Bernard Comrie of the Max Planck Institute for Evolutionary Anthropology, Leipzig. Mary Ritchie Key of the University of California, Irvine is the

founding editor. The database has an especially large selection of indigenous South American languages and Northeast Caucasian languages.

The Intercontinental Dictionary Series' advanced browsing function allows users to make custom tables which compare languages in side-by-side columns.

Below are the languages that are currently included in the Intercontinental Dictionary Series. The languages are grouped by language families, some of which are still hypothetical.

It is part of the Cross-Linguistic Linked Data project hosted by the Max Planck Institute for the Science of Human History.

Amerindian

North America

- Tlingit
- Haida
- Tsimshian
- Wakashan
- Nootka
- Salishan
- Bella Coola
- Chehalis
- Hokan?
- Karok
- Seri
- Zuni

- Nahuatl (Sierra de Zacapoaxtla, Puebla)
- Chatino, Zacatepec

Northern South America

- Chocoan
- Emberá
- Embera – Colombia
- Epena – Colombia
- Chibchan
- Muisca – Colombia
- Barí (Tairona) – Colombia / Venezuela
- Cofán – Colombia / Ecuador
- Barbacoan
- Cayapa (Cha'palaachi) – Ecuador
- Colorado (Tsafiki) – Ecuador
- Páez – Colombia
- Yanomaman
- Yanomami
- Ninam
- Yaruro – Venezuela
- Tucanoan
- Siona – Ecuador
- Tuyuca – Colombia / Brazil
- Jivaroan
- Aguaruna – Peru / Ecuador
- Waorani (Huaorani) – Ecuador

Amazonia

- Arawakan
- Goajiro (Wayuu) – Colombia

- Wapishana – Guyana / Brazil
- Yavitero – Venezuela (extinct)
- Mashco Piro (Yine) – Peru / Brazil
- Waurá – Brazil
- Baure – Bolivia
- Moxos – Bolivia
- Ignaciano – Bolivia
- Trinitario – Bolivia
- Macro-Gê
- Karajá
- Gê
- Kaingáng
- Canela
- Tupian
- Tupinambá – Brazil
- Guaraní – Paraguay
- Chiriguano – Bolivia
- Aché – Paraguay
- Mundurukú – Brazil
- Sirionó – Bolivia
- Wayampi – French Guiana
- Cariban
- Carib (De'kwana)
- Panare – Venezuela
- Macushi – Brazil / Guyana
- Wai Wai – Brazil / Guyana
- Panoan
- Cashibo – Peru
- Shipibo-Conibo – Peru
- Yaminahua – Peru
- Chácobo – Bolivia
- Pacahuara – Bolivia

- Tacanan
- Ese Ejja (Huarayo) – Peru / Bolivia
- Tacana – Bolivia
- Cavineña – Bolivia
- Araona – Bolivia
- Catuquina – Acre, Brazil
- Puinavean (Nadahup/Makú)
- Hup – Brazil / Colombia
- Yuwana (Hodï)? – Venezuela
- Peba-Yaguan
- Yagua – Brazil
- Chapacuran
- Pacaas Novos – Brazil
- Uru-Chipaya
- Chipaya – Bolivia
- Trumai – Brazil
- Aymara
- Cayuvava – Bolivia (extinct)
- Itonama – Bolivia
- Movima – Bolivia

Southern South America

- Guaicuruan
- Pilagá – Argentina
- Toba – Argentina / Paraguay
- Mocoví – Argentina
- Matacoan
- Chorote – Argentina
- Maká – Paraguay
- Nivaclé – Paraguay
- Wichi – Argentina

- Zamucoan
- Ayoreo – Paraguay / Bolivia
- Mascoian
- Sanapaná – Paraguay
- Moseten
- Mosetén (Tsimané) – Bolivia
- Chon
- Selknam
- Tehuelche
- Qawasqar
- Puelche (Gününa Küne) – Argentina Pampas
- Kunza – Chile (extinct)
- Mapudungun – Chile / Argentina
- Yagán (Yaghan)

Northeast Caucasian

- Northeast Caucasian
- Nakh
- Chechen
- Avar–Andic
- Avar
- Andi
- Botlikh
- Chamalal
- Ghodoberi
- Bagvalin (Bagvalal)
- Tindi
- Karata
- Akhvakh
- Tsezic

- Tsez
- Hinukh
- Bezhta
- Hunzib
- Khvarshi
- Lak (isolate)
- Khinalug (isolate)
- Dargi
- Dargwa
- Lezgic
- Archi
- Udi
- Lezgi
- Aghul
- Tabasaran
- Budukh
- Rutul
- Tsakhur

Indo-European

- Indo-European
- Hittite
- Tocharian A/B
- Armenian (Eastern, Western)
- Albanian, Tosk
- Greek (Ancient, Modern)
- Indo-Iranian
- Persian
- Avestan
- Tats (Judeo-Tat)

- Sanskrit
- Romani
- Celtic
- Irish (Old, Modern)
- Breton
- Welsh
- Germanic
- Core Germanic
- English (Old, Middle, Modern)
- German (Old, Middle, Modern)
- Yiddish
- Dutch
- Gothic
- Scandinavian
- Old Norse
- Danish
- Swedish
- Balto-Slavic
- Baltic
- Lithuanian
- Latvian
- Prussian
- Slavic
- Russian
- Old Church Slavonic
- Bulgarian
- Serbo-Croatian
- Polish
- Czech
- Romance
- Latin
- Spanish

- Portuguese
- Catalan
- French
- Italian
- Romanian

Uralic

- Uralic
- Finnic languages
- Finnish
- Estonian
- Hungarian
- Mordvinic languages
- Erzya-Mordvin
- Komi
- Khanty
- Udmurt
- Mansi
- Mari
- Samic languages
- Northern Sami
- Samoyedic
- Nenets
- Selkup

Tai-Kadai

- Tai-Kadai
- Kra
- Gelao (Qau)

- Gelao (Hakei)
- Buyang (Langjia)
- Buyang (Ecun)
- Hlai
- Li (Baoting)
- Kam-Sui
- Lakkja
- Mulam
- Maonan
- Chadong
- Kam, Southern
- Sui
- Tai
- Zhuang (Longzhou)
- Nung (Fengshan)
- Nung (Lazhai)
- Nung (Ningbei)
- Tai Khuen
- Tai Lue
- Dehong
- Shan
- Thai (standard)
- Thai (central)
- Thai (Khorat)
- Thai (Songkhla)

Others

- Basque
- Elamite

- Turkic
- Azerbaijan
- Nogai
- Kumyk
- Chulym
- Austronesian
- Proto Austronesian
- Proto Polynesian
- Rotuman – Fiji
- Tongan
- Marquesan
- Tuamotuan
- Hawaiian
- Māori
- Rapa Nui
- Afro-Asiatic
- Semitic
- Arabic
- Aramaic
- Chadic
- Hausa
- Polci
- Nilo-Saharan
- Ghulfan
- Creoles
- Negerhollands (Dutch-based) – U.S. Virgin Islands
- Limonese Creole (English-based) – Costa Rica
- Lengua (Quechua-based) – Ecuador (mixed)

Lexicostatistics

Lexicostatistics is a method of comparative linguistics that involves comparing the percentage of lexical cognates between languages to determine their relationship. Lexicostatistics is related to the comparative method but does not reconstruct a proto-language. It is to be distinguished from glottochronology, which attempts to use lexicostatistical methods to estimate the length of time since two or more languages diverged from a common earlier proto-language. This is merely one application of lexicostatistics, however; other applications of it may not share the assumption of a constant rate of change for basic lexical items.

The term "lexicostatistics" is misleading in that mathematical equations are used but not statistics. Other features of a language may be used other than the lexicon, though this is unusual. Whereas the comparative method used shared identified innovations to determine sub-groups, lexicostatistics does not identify these. Lexicostatistics is a distance-based method, whereas the comparative method considers language characters directly. The lexicostatistics method is a simple and fast technique relative to the comparative method but has limitations (discussed below). It can be validated by cross-checking the trees produced by both methods.

History

Lexicostatistics was developed by Morris Swadesh in a series of articles in the 1950s, based on earlier ideas. The concept's first known use was by Dumont d'Urville in 1834 who

compared various "Oceanic" languages and proposed a method for calculating a coefficient of relationship. Hymes (1960) and Embleton (1986) both review the history of lexicostatistics.

Method

Create word list

The aim is to generate a list of universally used meanings (hand, mouth, sky, I). Words are then collected for these meaning slots for each language being considered. Swadesh reduced a larger set of meanings down to 200 originally. He later found that it was necessary to reduce it further but that he could include some meanings that were not in his original list, giving his later 100-item list. The Swadesh list in Wiktionary gives the total 207 meanings in a number of languages. Alternative lists that apply more rigorous criteria have been generated, e.g. the Dolgopolsky list and the Leipzig–Jakarta list, as well as lists with a more specific scope; for example, Dyen, Kruskal and Black have 200 meanings for 84 Indo-European languages in digital form.

Determine cognacies

A trained and experienced linguist is needed to make cognacy decisions. However, the decisions may need to be refined as the state of knowledge increases. However, lexicostatistics does not rely on all the decisions being correct. For each pair of lists the cognacy of a form could be positive, negative or indeterminate. Sometimes a language has multiple words for one meaning, e.g. *small* and *little* for *not big*.

Calculate lexicostatistic percentages

This percentage is related to the proportion of meanings for a particular language pair that are cognate, i.e. relative to the total without indeterminacy. This value is entered into aN x N table of distances, where N is the number of languages being compared. When complete this table is half-filled in triangular form. The higher the proportion of cognacy the closer the languages are related.

Create family tree

Creation of the language tree is based solely on the table found above. Various sub-grouping methods can be used but that adopted by Dyen, Krustal and Black was:

- all lists are placed in a pool
- the two closest members are removed and form a nucleus which is placed in the pool
- this step is repeated
- under certain conditions a nucleus becomes a group
- this is repeated until the pool only contains one group.

Calculations have to be of nucleus and group lexical percentages.

Applications

A leading exponent of lexicostatistics application has been Isidore Dyen. He used lexicostatistics to classify Austronesian languages as well as Indo-European ones. A major study of the

latter was reported by Dyen, Kruskal and Black (1992). Studies have also been carried out on Amerindian and African languages.

Pama-Nyungan

The question of internal branching within the Pama-Nyungan language family has been a long-standing issue within Australianist linguistics, and general consensus held that internal connections between the 25+ different subgroups of Pama-Nyungan were either impossible to reconstruct or that the subgroups were not in fact genetically related at all. In 2012,

Claire Bowern and Quentin Atkinson published the results from their application of computational phylogenetic methods on 194 doculects representing all major subgroups and isolates of Pama-Nyungan.

Their model "recovered" many of the branches and divisions that had erstwhile been proposed and accepted by many other Australianists, while also providing some insight into the more problematic branches, such as Paman (which is complicated by the lack of data) and Ngumpin-Yapa (where the genetic picture is obscured by very high rates of borrowing between languages). Their dataset forms the largest of its kind for a hunter-gatherer language family, and the second largest overall after Austronesian (Greenhill et al. 2008). They conclude that Pama-Nyungan languages are in fact not exceptional to lexicostatistical methods, which have successfully been applied to other language families of the world.

Criticisms

People such as Hoijer (1956) have showed that there were difficulties in finding equivalents to the meaning items while many have found it necessary to modify Swadesh's lists. Gudschinsky (1956) questioned whether it was possible to obtain a universal list.

Factors such as borrowing, tradition and taboo can skew the results, as with other methods. Sometimes lexicostatistics has been used with lexical similarity being used rather than cognacy to find resemblances. This is then equivalent to mass comparison.

The choice of meaning slots is subjective, as is the choice of synonyms.

Improved methods

Some of the modern computational statistical hypothesis testing methods can be regarded as improvements of lexicostatistics in that they use similar word lists and distance measures.

Mass comparison

Mass comparison is a method developed by Joseph Greenberg to determine the level of genetic relatedness between languages. It is now usually called**multilateral comparison**. The method is rejected by most linguists (Campbell 2001, p. 45), though not all.

Some of the top-level relationships Greenberg named are now generally accepted, though they had already been posited by others (e.g. Afro-Asiatic and Niger–Congo). Others are accepted by many though disputed by some prominent specialists (e.g. Nilo-Saharan), others are predominantly rejected but have some defenders (e.g. Eurasiatic), while others are almost universally rejected (e.g. Khoisan and Amerind).

Methodology

The thesis of mass comparison is that a group of languages is related when they show numerous resemblances in vocabulary, including pronouns, and morphemes, forming an interlocking pattern common to the group. Unlike the comparative method, mass comparison does not require any regular or systematic correspondences between the languages compared; all that is required is an impressionistic feeling of similarity. Greenberg does not establish a clear standard for determining relatedness; he does not set a standard for what he considers a "resemblance" or how many resemblances are needed to prove relationship.

Mass comparison is done by setting up a table of basic vocabulary items and their forms in the languages to be compared for resemblances. The table can also include common morphemes. The following table was used by Greenberg (1957, p. 41) to illustrate the technique. It shows the forms of six items of basic vocabulary in nine different languages, identified by letters.

According to Greenberg, basic relationships can be determined without any experience in the case of languages that are fairly

closely related, though knowledge of probable paths of sound change acquired through typology allows one to go farther faster. For instance, the path *p*>*f* is extremely frequent, but the path *f*>*p* is much less so, enabling one to hypothesize that *fi* :*pi* and *fik* : *pix* are indeed related and go back to protoforms **pi* and **pik*/*x*. Similarly, while knowledge that *k*>*x* is extremely frequent, *x*>*k* much less so enables one to choose **pik* over **pix*. Thus, according to Greenberg (2005:318), phonological considerations come into play from the very beginning, even though mass comparison does not attempt to produce reconstructions of protolanguages as these belong to a later phase of study. The tables used in actual mass comparison involve much larger numbers of items and languages. The items included may be either lexical, such as 'hand', 'sky', and 'go', or morphological, such as PLURAL and MASCULINE (Ruhlen 1987, p. 120). For Greenberg, the results achieved through mass comparison approached certainty (Greenberg 1957, p. 39): "The presence of fundamental vocabulary resemblances and resemblances in items with grammatical function, particularly if recurrent through a number of languages, is a sure indication of genetic relationship."

Relation to the comparative method

As a tool for identifying genetic relationships between languages, mass comparison is an alternative to the comparative method. Proponents of mass comparison, such as Greenberg, claim that the comparative method is unnecessary to identify genetic relationships; furthermore, they claim that it can only be used once relationships are identified using mass comparison, making mass comparison the "first step" in

determining relationships (1957:44). This contrasts with mainstream comparative linguistics, which relies on the comparative method to aid in identifying genetic relationships; specifically, it involves comparing data from two or more languages. If sets of recurrent sound correspondences are found, the languages are most likely related; if further investigation confirms the potential relationship, reconstructed ancestral forms can be set up using the collated sound correspondences.

However, Greenberg did not entirely disavow the comparative method; he stated that "once we have a well-established stock I go about comparing and reconstructing just like anyone else, as can be seen in my various contributions to historical linguistics" (1990, quoted in Ruhlen 1994:285) and accused mainstream linguists of spreading "the strange and widely disseminated notion that I seek to replace the comparative method with a new and strange invention of my own" (2002:2). Earlier in his career, before he fully developed mass comparison, he even stated that his methodology did not "conflict in any fashion with the traditional comparative method" (1957:44). However, Greenberg sees the comparative method as playing no role in determining relationships, significantly reducing its importance compared to traditional methods of linguistic comparison. In effect, his approach of mass comparison sidelined the comparative method with a "new and strange invention of his own".

Reflecting the methodological empiricism also present in his typological work, he viewed facts as of greater weight than their interpretations, stating (1957:45):

- [R]econstruction of an original sound system has the status of an explanatory theory to account for etymologies already strong on other grounds. Between the **vaida* of Bopp and the **γwoidxe* of Sturtevant lie more than a hundred years of the intensive development of Indo-European phonological reconstruction. What has remained constant has been the validity of the etymologic relationship among Sanskrit *veda*, Greek *woida*, Gothic *wita*, all meaning "I know", and many other unshakable etymologies both of root and of non-root morphemes recognized at the outset. And who will be bold enough to conjecture from what original the Indo-Europeanist one hundred years from now will derive these same forms?

Criticism

Errors in application

The presence of frequent errors in Greenberg's data has been pointed out by linguists such as Lyle Campbell and Alexander Vovin, who see it as fatally undermining Greenberg's attempt to demonstrate the reliability of mass comparison. Campbell notes in his discussion of Greenberg's Amerind proposal that "nearly every specialist finds extensive distortions and inaccuracies in Greenberg's data"; for example, Willem Adelaar, a specialist in Andean languages, has stated that "the number of erroneous forms [in Greenberg's data probably exceeds that of the correct forms". Some forms in Greenberg's data even appear to be attributed to the wrong language. Greenberg also neglects known sound changes that languages have undergone;

once these are taken into account, many of the resemblances he points out vanish. Greenberg's data also contains errors of a more systematic sort: for instance, he groups unrelated languages together based on outdated classifications or because they have similar names.

Greenberg also arbitrarily deems certain portions of a word to be affixes when affixes of the requisite phonological shape are unknown to make words cohere better with his data. Conversely, Greenberg frequently employs affixed forms in his data, failing to recognise actual morphemic boundaries; when affixes are removed, the words often no longer bear any resemblance to his "Amerind" reconstructions. Greenberg has responded to this criticism by claiming that "the method of multilateral comparison is so powerful that it will give reliable results even with the poorest of data. Incorrect material should merely have a randomizing effect". This has hardly reassured critics of the method, who are far from convinced of the method's "power".

Borrowing

A prominent criticism of mass comparison is that it cannot distinguish borrowed forms from inherited ones, unlike comparative reconstruction, which is able to do so through regular sound correspondences. Undetected borrowings within Greenberg's data support this claim; for instance, he lists "cognates" of Uwa*baxita* "machete", even though it is a borrowing from Spanish*machete*. Greenberg (1957, p. 39) admits that "in particular and infrequent instances the question of borrowing may be doubtful" when using mass comparison, but claims that basic vocabulary is unlikely to be

borrowed compared to cultural vocabulary, stating that "where a mass of resemblances is due to borrowing, they will tend to appear in cultural vocabulary and to cluster in certain semantic areas which reflect the cultural nature of the contact." Mainstream linguists accept this premise, but claim that it does not suffice for distinguish borrowings from inherited vocabulary.

According to him, any type of linguistic item may be borrowed "on occasion", but "fundamental vocabulary is proof against mass borrowing". However, languages can and do borrow basic vocabulary. For instance, in the words of Campbell, Finnish has borrowed "from its Baltic and Germanic neighbors various terms for basic kinship and body parts, including 'mother', 'daughter', 'sister', 'tooth', 'navel', 'neck', 'thigh', and 'fur'". Greenberg continues by stating that "[D]erivational, inflectional, and pronominal morphemes and morph alternations are the least subject of all to borrowing"; he does incorporate morphological and pronominal correlations when performing mass comparison, but they are peripheral and few in number compared to his lexical comparisons. Greenberg himself acknowledges the peripheral role they play in his data by saying that they are "not really necessary". Furthermore, the correlations he lists are neither exclusive to or universally found within the languages which he compares. Greenberg is correct in pointing out that borrowing of pronouns or morphology is rare, but it cannot be ruled out without recourse to a method more sophisticated than mass comparison.

Greenberg continues by claiming that "[R]ecurrent sound correspondences" do not suffice to detect borrowing, since "where loans are numerous, they often show such

correspondences" (Greenberg 1957, pp. 39–40). However, Greenberg misrepresents the practices of mainstream comparative linguistics here; few linguists advocate using sound correspondences to the exclusion of all other kinds of evidence. This additional evidence often helps separate borrowings from inherited vocabulary; for instance, Campbell mentions how "[c]ertain sorts of patterned grammatical evidence (that which resists explanation from borrowing, accident, or typology and universals) can be important testimony, independent of the issue of sound correspondences". It may not always be possible to separate borrowed and inherited material, but any method has its limits; in the vast majority of cases, the difference can discerned.

Chance resemblances

Cross-linguistically, chance resemblances between unrelated lexical items are common, due to the large amount of lexemes present across the world's languages; for instance, English *much* and Spanish *mucho* are demonstrably unrelated, despite their similar phonological shape. This means that many of the resemblances found through mass comparison finds are likely to be coincidental. Greenberg worsens this issue by reconstructing a common ancestor when only a small proportion of the languages he compares actually display a match for any given lexical item, effectively allowing him to cherry-pick similar-looking lexical items from a wide array of languages. Though they are less susceptible to borrowing, pronouns and morphology also typically display a restricted subset of a language's phonemic inventory, making cross-linguistic chance resemblances more likely.

Greenberg also allows for a wide semantic latitude when comparing items; while widely-accepted linguistic comparisons do allow for a degree of semantic latitude, what he allows for is incommensurably greater; for instance, he one of his comparisons involves words for "night", "excrement", and "grass".

Sound symbolism and onomatopoeia

Proponents of mass comparison often neglect to exclude classes of words that are usually considered to be unreliable for proving linguistic relationships. For instance, Greenberg made no attempt to exclude onomatopoeic words from his data. Onomatopoeic words are often excluded from linguistic comparison, as similar-sounding onomatopoeic words can easily evolve in parallel. Though it is impossible to make a definite judgement as to whether a word is onomatopoeic, certain semantic fields, such as "blow" and "suck", show a cross-linguistic tendency to be onomatopoeic; making such a judgement may require deep analysis of a type that mass comparison makes difficult. Similarly, Greenberg neglected to exclude items affected by sound symbolism, which often distorts the original shape of lexical items, from his data. Finally, "nursery words", such as "mama" and "papa" lack evidential value in linguistic comparison, as they are usually thought to derive from the sounds infants make when beginning to acquire languages. Advocates of mass comparison often avoid taking sufficient care to exclude nursery words; one, Merritt Ruhlen has even attempted to downplay the problems inherent in using them in linguistic comparison. The fact that many of indigenous languages of the Americas have pronouns that begin with nasal stops, which Greenberg sees as

evidence of common ancestry, may ultimately also be linked to early speech development; Algonquian specialist Ives Goddard notes that "A gesture equivalent to that used to articulate the sound *n* is the single most important voluntary muscular activity of a nursing infant".

Disputed legacy of the comparative method

The conflict over mass comparison can be seen as a dispute over the legacy of the comparative method, developed in the 19th century, primarily by Danish and German linguists, in the study of Indo-European languages.

Position of Greenberg's detractors

Since the development of comparative linguistics in the 19th century, a linguist who claims that two languages are related, whether or not there exists historical evidence, is expected to back up that claim by presenting general rules that describe the differences between their lexicons, morphologies, and grammars. The procedure is described in detail in the comparative method article.

For instance, one could demonstrate that Spanish is related to Italian by showing that many words of the former can be mapped to corresponding words of the latter by a relatively small set of replacement rules—such as the correspondence of initial *es-* and *s-*, final *-os* and *-i*, etc.

Many similar correspondences exist between the grammars of the two languages. Since those systematic correspondences are extremely unlikely to be random coincidences, the most likely explanation by far is that the two languages have evolved from a single ancestral tongue (Latin, in this case).

All pre-historical language groupings that are widely accepted today—such as the Indo-European, Uralic, Algonquian, and Bantu families—have been established this way.

Response of Greenberg's defenders

The actual development of the comparative method was a more gradual process than Greenberg's detractors suppose. It has three decisive moments. The first was Rasmus Rask's observation in 1818 of a possible regular sound change in Germanic consonants. The second was Jacob Grimm's extension of this observation into a general principle (Grimm's law) in 1822.

The third was Karl Verner's resolution of an irregularity in this sound change (Verner's law) in 1875. Only in 1861 did August Schleicher, for the first time, present systematic reconstructions of Indo-European proto-forms (Lehmann 1993:26). Schleicher, however, viewed these reconstructions as extremely tentative (1874:8). He never claimed that they proved the existence of the Indo-European family, which he accepted as a given from previous research—primarily that of Franz Bopp, his great predecessor in Indo-European studies.

Karl Brugmann, who succeeded Schleicher as the leading authority on Indo-European, and the other Neogrammarians of the late 19th century, distilled the work of these scholars into

the famous (if often disputed) principle that "every sound change, insofar as it occurs automatically, takes place according to laws that admit of no exception" (Brugmann 1878).

The Neogrammarians did not, however, regard regular sound correspondences or comparative reconstructions as relevant to the proof of genetic relationship between languages. In fact, they made almost no statements on how languages are to be classified (Greenberg 2005:158). The only Neogrammarian to deal with this question was Berthold Delbrück, Brugmann's collaborator on the *Grundriß der vergleichenden Grammatik der indogermanischen Sprachen* (Greenberg 2005:158-159, 288). According to Delbrück (1904:121-122, quoted in Greenberg 2005:159), Bopp had claimed to prove the existence of Indo-European in the following way:

The proof was produced by juxtaposing words and forms of similar meanings. When one considers that in these languages the formation of the inflectional forms of the verb, noun and pronoun agrees in essentials and likewise that an extraordinary number of inflected words agree in their lexical parts, the assumption of chance agreement must appear absurd.

Furthermore, Delbrück took the position later enunciated by Greenberg on the priority of etymologies to sound laws (1884:47, quoted in Greenberg 2005:288): "obvious etymologies are the material from which sound laws are drawn."

The opinion that sound correspondences or, in another version of the opinion, reconstruction of a proto-language are necessary to show relationship between languages thus dates

from the 20th, not the 19th century, and was never a position of the Neogrammarians. Indo-European was recognized by scholars such as William Jones (1786) and Franz Bopp (1816) long before the development of the comparative method.

Furthermore, Indo-European was not the first language family to be recognized by students of language. Semitic had been recognized by European scholars in the 17th century, Finno-Ugric in the 18th. Dravidian was recognized in the mid-19th century by Robert Caldwell (1856), well before the publication of Schleicher's comparative reconstructions.

Finally, the supposition that all of the language families generally accepted by linguists today have been established by the comparative method is untrue. For example, although Eskimo–Aleut has long been accepted as a valid family, "Proto-Eskimo–Aleut has not yet been reconstructed" (Bomhard 2008:209). Other families were accepted for decades before comparative reconstructions of them were put forward, for example Afro-Asiatic and Sino-Tibetan. Many languages are generally accepted as belonging to a language family even though no comparative reconstruction exists, often because the languages are only attested in fragmentary form, such as the Anatolian language Lydian (Greenberg 2005:161). Conversely, detailed comparative reconstructions exist for some language families which nonetheless remain controversial, such as Altaic and Nostratic (however, a specification is needed here: Nostratic is a proposed proto-proto-language, while Altaic is a "simple" proto-language - with Altaic languages widely accepted as typologically related. Detractors of both proposals simply claim that the data collected to show by comparativism the existence of both families is scarce, wrong and non

sufficient. Keep in mind that regular phonological correspondences need thousands of lexicon lists to be prepared and compared before being established. These lists are lacking for both the proposed families. Furthermore, other specific problems affect "comparative" lists of both proposals, like the late attestation for Altaic languages, or the comparison of not certain proto-forms, like proto-Kartvelian, for Nostratic.).

A continuation of earlier methods?

Greenberg claimed that he was at bottom merely continuing the simple but effective method of language classification that had resulted in the discovery of numerous language families prior to the elaboration of the comparative method (1955:1-2, 2005:75) and that had continued to do so thereafter, as in the classification of Hittite as Indo-European in 1917 (Greenberg 2005:160-161).

This method consists in essentially two things: resemblances in basic vocabulary and resemblances in inflectional morphemes. If mass comparison differs from it in any obvious way, it would seem to be in the theoretization of an approach that had previously been applied in a relatively ad hoc manner and in the following additions:

- The explicit preference for basic vocabulary over cultural vocabulary.
- The explicit emphasis on comparison of multiple languages rather than bilateral comparisons.
- The very large number of languages simultaneously compared (up to several hundred).

- The introduction of typologically based paths of sound change.

The positions of Greenberg and his critics therefore appear to provide a starkly contrasted alternative:

- According to Greenberg, the identification of sound correspondences and the reconstruction of protolanguages arise from genetic classification.
- According to Greenberg's critics, genetic classification arises from the identification of sound correspondences or (others state) the reconstruction of protolanguages.

Time limits of the comparative method

Besides systematic changes, languages are also subject to random mutations (such as borrowings from other languages, irregular inflections, compounding, and abbreviation) that affect one word at a time, or small subsets of words. For example, Spanish *perro* (dog), which does not come from Latin, cannot be rule-mapped to its Italian equivalent *cane* (the Spanish word *can* is the Latin-derived equivalent but is much less used in everyday conversations, being reserved for more formal purposes). As those sporadic changes accumulate, they will increasingly obscure the systematic ones—just as enough dirt and scratches on a photograph will eventually make the face unrecognisable.

On this point, Greenberg and his critics agree, as over against the Moscow school, but they draw contrasting conclusions:

- Greenberg's critics argue that the comparative method has an inherent limit of 6,000 – 10,000 years (depending on the author), and that beyond this too many irregularities of sound change have accumulated for the method to function. Since according to them the identification of regular sound correspondences is necessary to establish genetic relationship, they conclude that genetic relationships older than 10,000 years (or less) cannot be determined. In consequence, it is not possible to go much beyond those genetic classifications that have already been arrived at (e.g. Ringe 1992:1).
- Greenberg argued that cognates often remain recognizable even when recurrent sound changes have been overlaid by idiosyncratic ones or interrupted by analogy, citing the cases of English *brother* (2002:4), which is easily recognizable as a cognate of German *Bruder* even though it violates Verner's law, and Latin *quattuor* (1957:45), easily recognizable as a reflex of Proto-Indo-European *k^{w}*etwor* even though the changes *e*>*a* and *t*>*tt* violate the usual sound changes from Proto-Indo-European to Latin. (In the case of *brother*, the sound changes are actually known, but intricate, and are only decipherable because the language is heavily documented from an early date. In the case of *quattuor*, the changes are genuinely irregular, and the form of the word can only be explained through means other than regular sound change, such as the operation of analogy.)

In contrast, the "Moscow school" of linguists, perhaps best known for its advocacy of the Nostratic hypothesis (though active in many other areas), has confidence in the traceability of regular sound changes at very great time depths, and believes that reconstructed proto-languages can be pyramided on top of each other so as to attain still earlier proto-languages, without violating the principles of the standard comparative method.

Toward a resolution of the conflict?

In spite of the apparently intractable nature of the conflict between Greenberg and his critics, a few linguists have begun to argue for its resolution. Edward Vajda, noted for his recent proposal of Dené–Yeniseian, attempts to stake out a position that is sympathetic to both Greenberg's approach and that of its critics, such as Lyle Campbell and Johanna Nichols. George Starostin, a member of the Moscow school, argues that Greenberg's work, while perhaps not going beyond inspection, presents interesting sets of forms that call for further scrutiny by comparative reconstruction, specifically with regard to the proposed Khoisan and Amerind families.

Moscow School of Comparative Linguistics

The **Moscow School of Comparative Linguistics** (also called the **Nostratic School**) is a school of linguistics based in Moscow, Russia that is known for its work in long-range comparative linguistics [ru]. Formerly based at Moscow State University, it is currently centered at the RSUH Institute of Linguistics [ru] (Institute of Linguistics of the Russian State

University for the Humanities), and also the Institute of Linguistics of the Russian Academy of Sciences in Moscow, Russia.

History

The founders of the school are Vladislav Illich-Svitych and Aharon Dolgopolsky. Vladimir Dybo, Vyacheslav Ivanov, and Andrey Zaliznyak also played key roles in the founding of the school.

In 1962, Illich-Svitych began working on a book about Nostrative comparative linguistics called *A Tentative Comparison of the Nostratic Languages* (Russian: Опытсравненияностратическихязыков). After the death of Illich-Svitych, his colleagues, Vladimir Dybo and Aharon Dolgopolsky, completed and published the book. Initially, Illich-Svitych and Dolgopolsky were probably doing research on Nostratic independently of each other. Both Dolgopolsky and Illich-Svitych published their first articles on Nostratic linguistics in 1964.

Conferences and seminars on Nostratic comparative linguistics were organized in the 1960s. Originally, they were held informally at Vladimir Dybo's apartment, and were only formally held at the Russian State University for the Humanities since 1992.

The first generation of the school includes the Russian linguists Sergei Starostin, Sergei Nikolaev, Alexander Militarev, Ilia Peiros, Anna Dybo, Oleg Mudrak [ru], Olga Stolbova [ru], and Eugene Helimski. The second generation mainly consists of

graduates of the Faculty of Linguistics (now the Institute of Linguistics) of the Russian State University for the Humanities from the 1990s. During the 1960s and 1970s, the school was centered at the Department of Theoretical and Applied Linguistics [ru] of the Faculty of Philology at Moscow State University. In the 1990s, the school moved to the Faculty of Linguistics of the Russian State University for the Humanities (RSUH). Some members are also affiliated with the Institute of Linguistics of the Russian Academy of Sciences.

From the 1970s until his death in 2005, Sergei Starostin was the informal "head of the Moscow school," and since 1999, the formal head of the Center of Comparative Linguistics [ru] at the RSUH Institute for Oriental and Classical Studies [ru].

Other linguists closely associated with the Moscow School of Comparative Linguistics include Václav Blažek, Irén Hegedűs, Murray Gell-Mann, John Bengtson, and Allan Bomhard.

Projects

The Tower of Babel website is the main lexical database for the Moscow School of Comparative Linguistics. The website runs on the Starling database management system and was originally developed by Sergei Starostin. After Sergei Starostin's death in 2005, the website has since been run by his son Georgiy Starostin. In 2011, the Global Lexicostatistical Database was launched.

Members of the Moscow School of Comparative Linguistics play key roles in the Evolution of Human Languages project at the Santa Fe Institute in the United States. In addition to regularly

hosting the Nostratic Seminar, the school also hosts the Annual Sergei Starostin Memorial Conference on Comparative-Historical Linguistics. Academic journals include the *Journal of Language Relationship*.

Pseudoscientific language comparison

Pseudoscientific language comparison is a form of pseudo-scholarship that has the objective of establishing historical associations between languages by naïve postulations of similarities between them.

While comparative linguistics also studies the historical relationships of languages, linguistic comparisons are considered pseudoscientific by linguists when they are not based on the established practices of comparative linguistics, or on the more general principles of the scientific method. Pseudoscientific language comparison is usually performed by people with little or no specialization in the field of comparative linguistics. It is a widespread type of linguistic pseudoscience.

The most common method applied in pseudoscientific language comparisons is to search two or more languages for words that seem similar in their sound and meaning. While similarities of this kind often seem convincing to laypeople, linguistic scientists consider this kind of comparison to be unreliable for two primary reasons. First, the method applied is not well-defined: the criterion of similarity is subjective and thus not subject to verification or falsification, which is contrary to the principles of the scientific method. Second, the large size of all

languages' vocabulary makes it easy to find coincidentally similar words between languages.

Because of its unreliability, the method of searching for isolated similarities is rejected by nearly all comparative linguists (however, see mass comparison for a controversial method that operates by similarity). Instead of noting isolated similarities, comparative linguists use a technique called the comparative method to search for regular (i.e. recurring) correspondences between the languages' phonology, grammar and core vocabulary in order to test hypotheses of relatedness.

Certain types of languages seem to attract much more attention in pseudoscientific comparisons than others. These include languages of ancient civilizations such as Egyptian, Etruscan or Sumerian; language isolates or near-isolates such as Basque, Japanese and Ainu; and languages that are unrelated to their geographical neighbors such as Hungarian.

Political or religious implications

In some cases, languages are associated with one another for political or religious reasons, despite a lack of support from accepted methods of science or historical linguistics:

For example, it was argued by Niclas Wahlgren that Herman Lundborg encouraged that the posited Ural-Altaic or Turanian, language family, which seeks to relate Sami to the Mongolian language, was used to justify Swedishracism towards the Sami people in particular. (There are also strong, albeit areal not genetic, similarities between the Uralic and Altaic languages, which provide a more benign but nonetheless incorrect basis

for this theory.) Some believers in Abrahamic religions have sought to derive their native languages from Classical Hebrew. For example, Herbert W. Armstrong (1892–1986), a proponent of British Israelism, claimed that the word 'British' comes from Hebrew בְּרִית (Hebrew pronunciation: [brit], meaning 'covenant') and אִישׁ (Hebrew pronunciation: [iʃ], meaning 'man'), as supposed proof that the British people are the 'covenant people' of God. Pre-modern scholars of the Hebrew Bible, debating the language spoken by Adam and Eve, often relied on belief in the literal truth of Genesis and of the accuracy of the names transcribed therein. On the other hand, the sixteenth-century Renaissance scholars Johannes Goropius Becanus (1519–1572) and Simon Stevin (1548–1620) argued that the Adamic language had been a dialect of their own native language, Dutch.

The Sun Language Theory, positing a proto-Turkic language as the ancestor of all human languages, was motivated by Turkish nationalism.

The Israeli-American linguist Paul Wexleris known for his fringe theories about the origin of Jewish populations and Jewish languages:

- that most Ashkenazi Jews are of Turkic origin, and that their language, Yiddish, is ultimately derived from Judaeo-Slavic
- that most Sephardi Jews are of Berber origin, as is their language, Ladino

The Lithuanian–American archaeologistMarija Gimbutas argued during the mid-1900s that Basque is clearly related to the extinct Pictish and Etruscan languages, even though at least the comparison had earlier been rejected within a decade

of being proposed in 1892 by Sir John Rhys. Her motivation was to show Basque was a remnant of an "Old European culture".

Traits and characteristics

There is no universal way to identify pseudoscientific language comparisons; indeed, it is not clear that all pseudoscientific language comparisons form a single group. However, the following characteristics tend to be more common among pseudoscientific theories (and their advocates) than among scientific ones:

Failure to apply an accepted, or at least systematic, method to demonstrate regular correspondences between the languages. Unsystematic comparisons are effectively unfalsifiable.

Failure to present grammatical evidence for relatedness: claims are based exclusively on word comparisons, even though in comparative linguistics grammatical evidence is also required to confirm relatedness.

Arbitrary segmentation of compared forms: comparisons are based on the similarity of only a part of the words compared (usually the first syllable), whereas the rest of the word is ignored.

Disregard for the effects of morphology on word structure: uninflected root forms may be compared with fully inflected forms, or marked forms may be used in preference to lesser- or unmarked forms.

Failure to consider the possibility of borrowing and areal features. Neighboring languages may share much vocabulary and many grammatical features as a result of language contact, and adequate application of the comparative method is required to determine whether the similarities result from contact or from relatedness.

Relying on typological similarities between languages: the morphological type of the language is claimed to provide evidence for relatedness, but in comparative linguistics only material parallels are accepted as evidence of a historical connection.

Neglect of known history: present-day forms of words are used in comparisons, neglecting either the attested or the reconstructed history of the language in question, or words of varying time depths (such as current, archaic, and reconstructed words) and reliability of reconstruction are used interchangeably.

Advocation of geographically far-fetched connections, such as comparing Finnish (in Finland) to Quechua (in Peru), or Basque (in Spain and France) to Ainu (in Japan), or Castilian (in Spain) to Japanese (in Japan). This criterion is only suggestive, though, as a long distance does not exclude the possibility of a relationship: English is demonstrably related to Hindi (in India), and Hawaiian to Malagasy (on Madagascar).

Advocacy of fanciful historical scenarios on the basis of the purported linguistic findings, e.g. claims of unknown civilizations or ancient migrations across oceans.

Proponents of pseudoscientific language comparisons also tend to share some common characteristics with cranks in other fields of science:

Overestimation of their own knowledge or competence in one or more of the languages under comparison, or their historical development, and underestimation of experts' knowledge. For example, assigning of incorrect meanings to words or sentences, quoting of rare or even spurious lexemes, morphs or meanings or of obscure dialect forms, misinterpretation of explanations in linguistic literature, or failure to take well-known developments or facts into account.

When forms and meanings are simply compiled and quoted from dictionaries (or even only a single source), inaccuracies creep in very easily. Even linguistically trained native speakers are not necessarily linguistic experts in their own language, its dialectology, and its history; and even professional linguists are not necessarily experts in large numbers of diverse languages and families.

Claims that the purported remote linguistic relationship is obvious and easy to perceive. A distant relationship between languages is usually not obvious on a superficial examination, and can only be uncovered via a successful application of the comparative method.

Failure to submit results to peer reviewed linguistic journals.

Assertion that criticism towards the theory is motivated by traditionalism, ideological factors or conspiracy on behalf of the linguistic community.

Quantitative comparative linguistics

- **Quantitative comparative linguistics** is the use of quantitative analysis as applied to comparative linguistics. Examples include the statistical fields of lexicostatistics and glottochronology, and the borrowing of phylogenetics from biology.

History

Statistical methods have been used for the purpose of quantitative analysis in comparative linguistics for more than a century. During the 1950s, the Swadesh list emerged: a standardised set of lexical concepts found in most languages, as words or phrases, that allow two or more languages to be compared and contrasted empirically.

Probably the first published quantitative historical linguistics study was by Sapir in 1916, while Kroeber and Chretien in 1937 investigated nine Indo-European (IE) languages using 74 morphological and phonological features (extended in 1939 by the inclusion of Hittite).

Ross in 1950 carried out an investigation into the theoretical basis for such studies. Swadesh, using word lists, developed lexicostatistics and glottochronology in a series of papers published in the early 1950s but these methods were widely criticised though some of the criticisms were seen as unjustified by other scholars.

Embleton published a book on "Statistics in Historical Linguistics" in 1986 which reviewed previous work and

extended the glottochronological method. Dyen, Kruskal and Black carried out a study of the lexicostatistical method on a large IE database in 1992. During the 1990s, there was renewed interest in the topic, based on the application of methods of computational phylogenetics and cladistics. Such projects often involved collaboration by linguistic scholars, and colleagues with expertise in information science and/or biological anthropology.

These projects often sought to arrive at an optimal phylogenetic tree (or network), to represent a hypothesis about the evolutionary ancestry and perhaps its language contacts. Pioneers in these methods included the founders of CPHL: computational phylogenetics in historical linguistics (CPHL project): Donald Ringe, Tandy Warnow, Luay Nakhleh and Steven N. Evans.

In the mid-1990s a group at Pennsylvania University computerised the comparative method and used a different IE database with 20 ancient languages. In the biological field several software programs were then developed which could have application to historical linguistics. In particular a group at the University of Auckland developed a method that gave controversially old dates for IE languages.

A conference on "Time-depth in Historical Linguistics" was held in August 1999 at which many applications of quantitative methods were discussed. Subsequently many papers have been published on studies of various language groups as well as comparisons of the methods.

Greater media attention was generated in 2003 after the publication by anthropologists Russell Gray and Quentin

Atkinson of a short study on Indo-European languages in *Nature*. Gray and Atkinson attempted to quantify, in a probabilistic sense, the age and relatedness of modern Indo-European languages and, sometimes, the preceding proto-languages.

The proceedings of an influential 2004 conference, *Phylogenetic Methods and the Prehistory of Languages*were published in 2006, edited by Peter Forster and Colin Renfrew.

Studied language families

Computational phylogenetic analyses have been performed for:

- Indo-European languages: Bouckaert (2012)
- Uralic languages: Honkola (2013)
- Turkic languages: Hruschka (2014)
- Dravidian languages: Kolipakam (2018)
- Austroasiatic languages: Sidwell (2015)
- Austronesian languages: Gray (2009)
- Pama-Nyungan languages: Bowern & Atkinson (2012), Bouckaert, Bowern and Atkinson (2018)
- Bantu languages: Currie (2013), Grollemund (2015)
- Semitic languages: Kitchen (2009)
- Dené–Yeniseian languages: Sicoli & Holton (2014)
- Uto-Aztecan languages: Wheeler & Whiteley (2014)
- Mayan languages: Atkinson (2006)
- Arawakan languages: Walker & Ribeiro (2011)
- Tupi-Guarani languages: Michael (2015)
- Sino-Tibetan languages: Zhang et al. (2019), Sagart et al. (2019)

Background

The standard method for assessing language relationships has been the comparative method. However this has a number of limitations. Not all linguistic material is suitable as input and there are issues of the linguistic levels on which the method operates. The reconstructed languages are idealized and different scholars can produce different results. Language family trees are often used in conjunction with the method and "borrowings" must be excluded from the data, which is difficult when borrowing is within a family. It is often claimed that the method is limited in the time depth over which it can operate. The method is difficult to apply and there is no independent test. Thus alternative methods have been sought that have a formalised method, quantify the relationships and can be tested.

A goal of comparative historical linguistics is to identify instances of genetic relatedness amongst languages. The steps in quantitative analysis are (i) to devise a procedure based on theoretical grounds, on a particular model or on past experience, etc. (ii) to verify the procedure by applying it to some data where there exists a large body of linguistic opinion for comparison (this may lead to a revision of the procedure of stage (i) or at the extreme of its total abandonment) (iii) to apply the procedure to data where linguistic opinions have not yet been produced, have not yet been firmly established or perhaps are even in conflict.

Applying phylogenetic methods to languages is a multi-stage process: (a) the encoding stage - getting from real languages to some expression of the relationships between them in the form

of numerical or state data, so that those data can then be used as input to phylogenetic methods (b) the representation stage - applying phylogenetic methods to extract from those numerical and/or state data a signal that is converted into some useful form of representation, usually two dimensional graphical ones such as trees or networks, which synthesise and "collapse" what are often highly complex multi dimensional relationships in the signal (c) the interpretation stage - assessing those tree and network representations to extract from them what they actually mean for real languages and their relationships through time.

Types of trees and networks

An output of a quantitative historical linguistic analysis is normally a tree or a network diagram. This allows summary visualisation of the output data but is not the complete result. A tree is a connected acyclic graph, consisting of a set of vertices (also known as "nodes") and a set of edges ("branches") each of which connects a pair of vertices. An internal node represents a linguistic ancestor in a phylogenic tree or network. Each language is represented by a path, the paths showing the different states as it evolves. There is only one path between every pair of vertices. Unrooted trees plot the relationship between the input data without assumptions regarding their descent. A rooted tree explicitly identifies a common ancestor, often by specifying a direction of evolution or by including an "outgroup" that is known to be only distantly related to the set of languages being classified. Most trees are binary, that is a parent has two children. A tree can always be produced even though it is not always appropriate. A different sort of tree is that only based on language similarities

/ differences. In this case the internal nodes of the graph do not represent ancestors but are introduced to represent the conflict between the different splits ("bipartitions") in the data analysis. The "phenetic distance" is the sum of the weights (often represented as lengths) along the path between languages. Sometimes an additional assumption is made that these internal nodes do represent ancestors.

When languages converge, usually with word adoption ("borrowing"), a network model is more appropriate. There will be additional edges to reflect the dual parentage of a language. These edges will be bidirectional if both languages borrow from one another. A tree is thus a simple network,however there are many other types of network. A phylogentic network is one where the taxa are represented by nodes and their evolutionary relationships are represented by branches. Another type is that based on splits, and is a combinatorial generalisation of the split tree.

A given set of splits can have more than one representation thus internal nodes may not be ancestors and are only an "implicit" representation of evolutionary history as distinct from the "explicit" representation of phylogenetic networks. In a splits network the phrenetic distance is that of the shortest path between two languages. A further type is the reticular network which shows incompatibilities (due to for example to contact) as reticulations and its internal nodes do represent ancestors. A network may also be constructed by adding contact edges to a tree. The last main type is the consensus network formed from trees. These trees may be as a result of bootstrap analysis or samples from a posterior distribution.

Language change

Change happens continually to languages, but not usually at a constant rate, with its cumulative effect producing splits into dialects, languages and language families. It is generally thought that morphology changes slowest and phonology the quickest. As change happens, less and less evidence of the original language remains. Finally there could be loss of any evidence of relatedness. Changes of one type may not affect other types, for example sound changes do not affect cognancy. Unlike biology, it cannot be assumed that languages all have a common origin and establishing relatedness is necessary. In modelling it is often assumed for simplicity that the characters change independently but this may not be the case. Besides borrowing, there can also be semantic shifts and polymorphism.

Analysis input

Data

Analysis can be carried out on the "characters" of languages or on the "distances" of the languages. In the former case the input to a language classification generally takes the form of a data matrix where the rows correspond to the various languages being analysed and the columns correspond to different features or characters by which each language may be described. These features are of two types cognates or typological data. Characters can take one or more forms (homoplasy) and can be lexical, morphological or phonological. Cognates are morphemes (lexical or grammatical) or larger

constructions. Typological characters can come from any part of the grammar or lexicon. If there are gaps in the data these have to be coded.

In addition to the original database of (unscreened) data, in many studies subsets are formed for particular purposes (screened data).

In lexicostatistics the features are the meanings of words, or rather semantic slots. Thus the matrix entries are a series of glosses. As originally devised by Swadesh the single most common word for a slot was to be chosen, which can be difficult and subjective because of semantic shift. Later methods may allow more than one meaning to be incorporated.

Constraints

Some methods allow constraints to be placed on language contact geography (isolation by distance) and on sub-group split times.

Databases

Swadesh originally published a 200 word list but later refined it into a 100 word one. A commonly used IE database is that by Dyen, Kruskal and Black which contains data for 95 languages, though the original is known to contain a few errors. Besides the raw data it also contains cognacy judgements. This is available online. The database of Ringe, Warnow and Taylor has information on 24 IE languages, with 22 phonological characters, 15 morphological characters and 333 lexical characters. Gray and Atkinson used a database of

87 languages with 2449 lexical items, based on the Dyen set with the addition of three ancient languages. They incorporated the cognacy judgements of a number of scholars. Other databases have been drawn up for African, Australian and Andean language families, amongst others.

Coding of the data may be in binary form or in multistate form. The former is often used but does result in a bias. It has been claimed that there is a constant scale factor between the two coding methods, and that allowance can be made for this. However, another study suggests that the topology may change

Word lists

The word slots are chosen to be as culture- and borrowing- free as possible. The original Swadesh lists are most commonly used but many others have been devised for particular purposes. Often these are shorter than Swadesh's preferred 100 item list.

Kessler has written a book on "The Significance of Word Lists while McMahon and McMahon carried out studies on the effects of reconstructability and retentiveness. The effect of increasing the number of slots has been studied and a law of diminishing returns found, with about 80 being found satisfactory. However some studies have used less than half this number.

Generally each cognate set is represented as a different character but differences between words can also be measured as a distance measurement by sound changes. Distances may also be measured letter by letter.

Morphological features

Traditionally these have been seen as more important than lexical ones and so some studies have put additional weighting on this type of character. Such features were included in the Ringe, Warnow and Taylor IE database for example. However other studies have omitted them.

Typological features

Examples of these features include glottalised constants, tone systems, accusative alignment in nouns, dual number, case number correspondence, object-verb order, and first person singular pronouns. These will be listed in the WALS database, though this is only sparsely populated for many languages yet.

Probabilistic models

Some analysis methods incorporate a statistical model of language evolution and use the properties of the model to estimate the evolution history. Statistical models are also used for simulation of data for testing purposes. A stochastic process can be used to describe how a set of characters evolves within a language.

The probability with which a character will change can depend on the branch but not all characters evolve together, nor is the rate identical on all branches. It is often assumed that each character evolves independently but this is not always the case. Within a model borrowing and parallel development (homoplasy) may also be modelled, as well as polymorphisms.

Effects of chance

Chance resemblances produce a level of noise against which the required signal of relatedness has to be found. A study was carried out by Ringe into the effects of chance on the mass comparison method. This showed that chance resemblances were critical to the technique and that Greenberg's conclusions could not be justified, though the mathematical procedure used by Rimge was later criticised.

With small databases sampling errors can be important.

In some cases with a large database and exhaustive search of all possible trees or networks is not feasible because of running time limitations. Thus there is a chance that the optimum solution is not found by heuristic solution-space search methods.

Detection of borrowing

Loanwords can severely affect the topology of a tree so efforts are made to exclude borrowings. However, undetected ones sometimes still exist. McMahon and McMahon showed that around 5% borrowing can affect the topology while 10% has significant effects. In networks borrowing produces reticulations. Minett and Wang examined ways of detecting borrowing automatically.

Split dating

Dating of language splits can be determined if it is known how the characters evolve along each branch of a tree. The simplest

assumption is that all characters evolve at a single constant rate with time and that this is independent of the tree branch. This was the assumption made in glottochronology. However, studies soon showed that there was variation between languages, some probably due to the presence of unrecognised borrowing. A better approach is to allow rate variation, and the gamma distribution is usually used because of its mathematical convenience. Studies have also been carried out that show that the character replacement rate depends on the frequency of use. Widespread borrowing can bias divergence time estimates by making languages seem more similar and hence younger. However, this also makes the ancestor's branch length longer so that the root is unaffected. This aspect is the most controversial part of quantitative comparative linguistics.

Types of analysis

There is a need to understand how a language classification method works in order to determine its assumptions and limitations. It may only be valid under certain conditions or be suitable for small databases. The methods differ in their data requirements, their complexity and running time. The methods also differ in their optimisation criteria.

Character based models

Maximum parsimony and maximum compatibility

These two methods are similar but the maximum parsimony method's objective is to find the tree (or network) in which the minimum number of evolutionary changes occurs. In some

implementations the characters can be given weights and then the objective is to minimise the total weighted sum of the changes. The analysis produces unrooted trees unless an outgroup is used or directed characters. Heuristics are used to find the best tree but optimisation is not guaranteed. The method is often implemented using the programs PAUP or TNT.

Maximum compatibility also uses characters, with the objective of finding the tree on which the maximum number of characters evolve without homoplasy. Again the characters can be weighted and when this occurs the objective is to maximise the sum of the weights of compatible characters. It also produces unrooted trees unless additional information is incorporated. There are no readily available heuristics available that are accurate with large databases.This method has only been used by Ringe's group.

In these two methods there are often several trees found with the same score so the usual practice is to find a consensus tree via an algorithm. A majority consensus has bipartitions in more than half of the input trees while a greedy consensus adds bipartitions to the majority tree. The strict consensus tree is the least resolved and contains those splits that are in every tree.

Bootstrapping (a statistical resampling strategy) is used to provide branch support values. The technique randomly picks characters from the input data matrix and then the same analysis is used. The support value is the fraction of the runs with that bipartition in the observed tree. However, bootstrapping is very time consuming.

Maximum likelihood and Bayesian analysis

Both of these methods use explicit evolution models. The maximum likelihood method optimises the probability of producing the observed data, while Bayesian analysis estimates the probability of each tree and so produces a probability distribution. A random walk is made through the "model-tree space". Both take an indeterminate time to run, and stopping may be arbitrary so a decision is a problem. However, both produce support information for each branch.

The assumptions of these methods are overt and are verifiable. The complexity of the model can be increased if required. The model parameters are estimated directly from the input data so assumptions about evolutionary rate are avoided.

Perfect Phylogenetic Networks

This method produces an explicit phylogenic network having an underlying tree with additional contact edges. Characters can be borrowed but evolve without homoplasy. To produce such networks, a graph-theoretic algorithm has been used.

Gray and Atkinson's method

The input lexical data is coded in binary form, with one character for each state of the original multi-state character. The method allows homoplasy and constraints on split times. A likelihood-based analysis method is used, with evolution expressed as a rate matrix. Cognate gain and loss is modelled with a gamma distribution to allow rate variation and with rate smoothing. Because of the vast number of possible trees with

many languages, Bayesian inference is used to search for the optimal tree. A Markov ChainMonte Carlo algorithm generates a sample of trees as an approximation to the posterior probability distribution. A summary of this distribution can be provided as a greedy consensus tree or network with support values. The method also provides date estimates.

The method is accurate when the original characters are binary, and evolve identically and independently of each other under a rates-across-sites model with gamma distributed rates; the dates are accurate when the rate of change is constant. Understanding the performance of the method when the original characters are multi-state is more complicated, since the binary encoding produces characters that are not independent, while the method assumes independence.

Nicholls and Gray's method

This method is an outgrowth of Gray and Atkinson's. Rather than having two parameters for a character, this method uses three. The birth rate, death rate of a cognate are specified and its borrowing rate. The birth rate is a Poisson random variable with a single birth of a cognate class but separate deaths of branches are allowed (Dollo parsimony). The method does not allow homoplasy but allows polymorphism and constraints. Its major problem is that it cannot handle missing data (this issue has since been resolved by Ryder and Nicholls. Statistical techniques are used to fit the model to the data. Prior information may be incorporated and an MCMC research is made of possible reconstructions. The method has been applied to Gray and Nichol's database and seems to give similar results.

Distance based models

These use a triangular matrix of pairwise language comparisons. The input character matrix is used to compute the distance matrix either using the Hamming distance or the Levenshtein distance. The former measures the proportion of matching characters while the latter allows costs of the various possible transforms to be included. These methods are fast compared with wholly character based ones. However, these methods do result in information loss.

UPGMA

The "Unweighted Pairwise Group Method with Arithmetic-mean" (UPGMA) is a clustering technique which operates by repeatedly joining the two languages that have the smallest distance between them. It operates accurately with clock-like evolution but otherwise it can be in error. This is the method used in Swadesh's original lexicostatistics.

Split Decomposition

This is a technique for dividing data into natural groups. The data could be characters but is more usually distance measures. The character counts or distances are used to generate the splits and to compute weights (branch lengths) for the splits.

The weighted splits are then represented in a tree or network based on minimising the number of changes between each pair of taxa. There are fast algorithms for generating the collection of splits. The weights are determined from the taxon to taxon

distances. Split decomposition is effective when the number of taxa is small or when the signal is not too complicated.

Neighbor joining

This method operates on distance data, computes a transformation of the input matrix and then computes the minimum distance of the pairs of languages. It operates correctly even if the languages do not evolve with a lexical clock. A weighted version of the method may also be used. The method produces an output tree. It is claimed to be the closest method to manual techniques for tree construction.

Neighbor-net

It uses a similar algorithm to neighbor joining. Unlike Split Decomposition it does not fuse nodes immediately but waits until a node has been paired a second time. The tree nodes are then replaced by two and the distance matrix reduced. It can handle large and complicated data sets. However, the output is a phenogram rather than a phylogram. This is the most popular network method.

Network

This was an early network method that has been used for some language analysis. It was originally developed for genetic sequences with more than one possible origin. Network collapses the alternative trees into a single network. Where there are multiple histories a reticulation (a box shape) is drawn. It generates a list of characters incompatible with a tree.

ASP

This uses a declarative knowledge representation formalism and the methods of Answer Set Programming. One such solver is CMODELS which can be used for small problems but larger ones require heuristics.

Preprocessing is used to determine the informative characters. CMODELS transforms them into a propositional theory that uses a SAT solver to compute the models of this theory.

Fitch/Kitch

Fitch and Kitch are maximum likelihood based programs in PHYLIP that allow a tree to be rearranged after each addition, unlike NJ. Kitch differs from Fitch in assuming a constant rate of change throughout the tree while Fitch allows for different rates down each branch.

Separation level method

Holm introduced a method in 2000 to deal with some known problems of lexicostatistical analysis.

These are the "symplesiomorphy trap", where shared archaisms are difficult to distinguish from shared innovations, and the "proportionality "trap" when later changes can obscure early ones. Later he introduced a refined method, called SLD, to take account of the variable word distribution across languages. The method does not assume aconstant rate of change.

Fast convergence methods

A number of fast converging analysis methods have been developed for use with large databases (>200 languages). One of these is the Disk Covering Method (DCM). This has been combined with existing methods to give improved performance. A paper on the DCM-NJ+MP method is given by the same authors in "The performance of Phylogenetic Methods on Trees of Bounded Diameter", where it is compared with the NJ method.

Resemblance based models

These models compare the letters of words rather than their phonetics. Dunn *et al.*studied 125 typological characters across 16 Austronesian and 15 Papuan languages. They compared their results to an MP tree and one constructed by traditional analysis. Significant differences were found. Similarly Wichmann and Saunders used 96 characters to study 63 American languages.

Computerised mass comparison

A method that has been suggested for initial inspection of a set of languages to see if they are related was mass comparison. However, this has been severely criticised and fell into disuse. Recently Kessler has resurrected a computerised version of the method but using rigorous hypothesis testing.

The aim is to make use of similarities across more than two languages at a time. In another paper various criteria for comparing word lists are evaluated. It was found that the IE

and Uralic families could be reconstructed but there was no evidence for a joint super-family.

Nichol's method

This method uses stable lexical fields, such as stance verbs, to try to establish long-distance relationships. Account is taken of convergence and semantic shifts to search for ancient cognates. A model is outlined and the results of a pilot study are presented.

ASJP

The Automated Similarity Judgment Program (ASJP) is similar to lexicostatistics, but the judgement of similarities is done by a computer program following a consistent set of rules. Trees are generated using standard phylogenetic methods. ASJP uses 7 vowel symbols and 34 consonant symbols. There are also various modifiers.

Two words are judged similar if at least two consecutive consonants in the respective words are identical while vowels are also taken into account. The proportion of words with the same meaning judged to be similar for a pair of languages is the Lexical Similarity Percentage (LSP). The Phonological Similarity Percentage (PSP) is also calculated. PSP is then subtracted from the LSP yielding the Subtracted Similarity Percentage (SSP) and the ASJP distance is 100-SSP. Currently there are data on over 4,500 languages and dialects in the ASJP database from which a tree of the world's languages was generated.

Serva and Petroni's method

This measures the orthographical distance between words to avoid the subjectivity of cognacy judgements. It determines the minimum number of operations needed to transform one word into another, normalised by the length of the longer word. A tree is constructed from the distance data by the UPGMA technique.

Phonetic evaluation methods

Heggarty has proposed a means of providing a measure of the degrees of difference between cognates, rather than just yes/no answers. This is based on examining many (>30) features of the phonetics of the glosses in comparison with the protolanguage. This could require a large amount of work but Heggarty claims that only a representative sample of sounds is necessary. He also examined the rate of change of the phonetics and found a large rate variation, so that it was unsuitable for glottochronology. A similar evaluation of the phonetics had earlier been carried out by Grimes and Agard for Romance languages, but this used only six points of comparison.

Evaluation of methods

Metrics

Standard mathematical techniques are available for measuring the similarity/difference of two trees. For consensus trees the Consistency Index (CI) is a measure of homoplasy. For one character it is the ratio of the minimimum conceivable number

of steps on any one tree (= 1 for binary trees) divided by the number of reconstructed steps on the tree. The CI of a tree is the sum of the character CIs divided by the number of characters. It represents the proportion of patterns correctly assigned.

The Retention Index (RI) measures the amount of similarity in a character. It is the ratio (g - s) / (g - m) where **g** is the greatest number of steps of a character on any tree, **m** is the minimum number of steps on any tree, and **s** is the minimum steps on a particular tree. There is also a Rescaled CI which is the product of the CI and RI.

For binary trees the standard way of comparing their topology is to use the Robinson-Foulds metric. This distance is the average of the number of false positives and false negatives in terms of branch occurrence. R-F rates above 10% are considered poor matches. For other sorts of trees and for networks there is yet no standard method of comparison.

Lists of incompatible characters are produced by some tree producing methods. These can be extremely helpful in analysing the output. Where heuristic methods are used repeatability is an issue. However, standard mathematical techniques are used to overcome this problem.

Comparison with previous analyses

In order to evaluate the methods a well understood family of languages is chosen, with a reliable dataset. This family is often the IE one but others have been used. After applying the methods to be compared to the database, the resulting trees are compared with the reference tree determined by traditional

linguistic methods. The aim is to have no conflicts in topology, for example no missing sub-groups, and compatible dates. The families suggested for this analysis by Nichols and Warnow are Germanic, Romance, Slavic, Common Turkic, Chinese, and Mixe Zoque as well as older groups such as Oceanic and IE.

Use of simulations

Although the use of real languages does add realism and provides real problems, the above method of validation suffers from the fact that the true evolution of the languages is unknown. By generating a set of data from a simulated evolution correct tree is known. However it will be a simplified version of reality. Thus both evaluation techniques should be used.

Sensitivity analysis

To assess the robustness of a solution it is desirable to vary the input data and constraints, and observe the output. Each variable is changed slightly in turn. This analysis has been carried out in a number of cases and the methods found to be robust, for example by Atkinson and Gray.

Studies comparing methods

During the early 1990s, linguist Donald Ringe, with computer scientists Luay Nakhleh and Tandy Warnow, statistician Steven N. Evans and others, began collaborating on research in quantitative comparative linguistic projects. They later founded the CHPL project, the goals of which include: "producing and maintaining real linguistic datasets, in particular of Indo-

European languages", "formulating statistical models that capture the evolution of historical linguistic data", "designing simulation tools and accuracy measures for generating synthetic data for studying the performance of reconstruction methods", and "developing and implementing statistically-based as well as combinatorial methods for reconstructing language phylogenies, including phylogenetic networks".

A comparison of coding methods was carried out by Rexova *et al.* (2003). They created a reduced data set from the Dyen database but with the addition of Hittite. They produced a standard multistate matrix where the 141 character states corresponds to individual cognate classes, allowing polymorphism. They also joined some cognate classes, to reduce subjectivity and polymorphic states were not allowed. Lastly they produced a binary matrix where each class of words was treated as a separate character. The matrices were analysed by PAUP. It was found that using the binary matrix produced changes near the root of the tree.

McMahon and McMahon (2003) used three PHYLIP programs (NJ, Fitch and Kitch) on the DKB dataset. They found that the results produced were very similar. Bootstrapping was used to test the robustness of any part of the tree. Later they used subsets of the data to assess its retentiveness and reconstructability. The outputs showed topological differences which were attributed to borrowing. They then also used Network, Split Decomposition, Neighbor-net and SplitsTree on several data sets. Significant differences were found between the latter two methods. Neighbor-net was considered optimal for discerning language contact.

In 2005, Nakhleh, Warnow, Ringe and Evans carried out a comparison of six analysis methods using an Indo-European database. The methods compared were UPGMA, NJ MP, MC, WMC and GA. The PAUP software package was used for UPGMA, NJ, and MC as well as computing the majority consensus trees. The RWT database was used but 40 characters were removed due to evidence of polymorphism. Then a screened database was produced excluding all characters that clearly exhibited parallel development, so eliminating 38 features. The trees were evaluated on the basis of the number of incompatible characters and on agreement with established sub-grouping results. They found that UPGMA was clearly worst but there was not a lot of difference between the other methods. The results depended on the data set used. It was found that weighting the characters was important, which requires linguistic judgement.

Saunders (2005) compared NJ, MP, GA and Neighbor-Net on a combination of lexical and typological data. He recommended use of the GA method but Nichols and Warnow have some concerns about the study methodology.

Cysouw *et al.* (2006) compared Holm's original method with NJ, Fitch, MP and SD. They found Holm's method to be less accurate than the others.

In 2013, François Barbancon, Warnow, Evans, Ringe and Nakleh (2013) studied various tree reconstruction methods using simulated data. Their simulated data varied in the number of contact edges, the degree of homoplasy, the deviation from a lexical clock, and the deviation from the rates-across-sites assumption. It was found that the accuracy

of the unweighted methods (MP, NJ, UPGMA, and GA) were consistent in all the conditions studied, with MP being the best. The accuracy of the two weighted methods (WMC and WMP) depended on the appropriateness of the weighting scheme. With low homoplasy the weighted methods generally produced the more accurate results but inappropriate weighting could make these worse than MP or GA under moderate or high homoplasy levels.

Choosing the best model

Choice of an appropriate model is critical for the production of good phylogenetic analyses. Both underparameterised or overly restrictive models may produce aberrant behaviour when their underlying assumptions are violated, while overly complex or overparameterised models require long run times and their parameters may be overfit. The most common method of model selection is the "Likelihood Ratio Test" which produces an estimate of the fit between the model and the data, but as an alternative the Akaike Information Criterion or the Bayesian Information Criterion can be used. Model selection computer programs are available.

Sound change

A **sound change**, in historical linguistics, is a change in the pronunciation of a language over time. A sound change can involve the replacement of one speech sound (or, more generally, one phonetic feature value) by a different one (called **phonetic change**) or a more general change to the speech

sounds that exist (**phonological change**), such as the merger of two sounds or the creation of a new sound.

A sound change can eliminate the affected sound, or a new sound can be added. Sound changes can be **environmentally conditioned** if the change occurs in only some sound environments, and not others.

The term "sound change" refers to diachronic changes, which occur in a language's sound system over time. On the other hand, "alternation" refers to changes that happen synchronically (within the language of an individual speaker, depending on the neighbouring sounds) and do not change the language's underlying system (for example, the *-s* in the English plural can be pronounced differently depending on the preceding sound, as in *bet*[s], *bed*[z], which is a form of alternation, rather than sound change).However, since "sound change" can refer to the historical introduction of an alternation (such as postvocalic /k/ in the Tuscan dialect, which was once [k] as in *di* [k]*arlo* 'of Carlo' but is now [h] *di* [h]*arlo* and alternates with [k] in other positions: *con* [k]*arlo* 'with Carlo'), that label is inherently imprecise and must often be clarified as referring to either phonemic change or restructuring.

Research on sound change is usually conducted under the working assumption that it is *regular*, which means that it is expected to apply mechanically whenever its structural conditions are met, irrespective of any non-phonological factors like the meaning of the words that are affected. However, apparent exceptions to regular change can occur because of dialect borrowing, grammatical analogy, or other

causes known and unknown, and some changes are described as "sporadic" and so they affect only one or a few particular words, without any apparent regularity.

The Neogrammarian linguists of the 19th century introduced the term "sound law" to refer to rules of regular change, perhaps in imitation of the laws of physics, and the term "law" is still used in referring to specific sound rules that are named after their authors like Grimm's Law, Grassmann's Lawetc.. Real-world sound changes often admit exceptions, but the expectation of their regularity or absence of exceptions is of great heuristic value by allowing historical linguists to define the notion of *regular correspondence* by the comparative method.

Each sound change is limited in space and time and so it functions in a limited area (within certain dialects) and for a limited period of time. For those and other reasons, the term "sound law" has been criticized for implying a universality that is unrealistic for to sound change.

A sound change that affects the phonological system or the number or the distribution of its phonemes is a phonological change.

Principles

The following statements are used as heuristics in formulating sound changes as understood within the Neogrammarian model. However, for modern linguistics, they are not taken as inviolable rules but are seen as guidelines.

Sound change has no memory: sound change does not discriminate between the sources of a sound. If a previous sound change causes X,Y>Y (features X and Y merge as Y), a new one cannot affect only an original X.

Sound change ignores grammar: a sound change can have only phonological constraints, like X > Z in unstressed syllables. For example, it cannot only affect adjectives. The only exception to this is that a sound change may or may not recognise word boundaries, even when they are not indicated by prosodic clues. Also, sound changes may be regularized in inflectional paradigms (such as verbal inflection), in which case the change is no longer phonological but morphological in nature.

Sound change is exceptionless: if a sound change can happen at a place, it will. It affects all sounds that meet the criteria for change. Apparent exceptions are possible, because of analogy and other regularization processes, another sound change, or an unrecognized conditioning factor.

That is the traditional view expressed by the Neogrammarians. In past decades, however, it has been shown that sound change does not necessarily affect all possible words. However, when a sound change is initiated, it often eventually expands to the whole lexicon. For example, the Spanish fronting of the Vulgar Latin [g] (voiced velar stop) before [i e ɛ] seems to have reached every possible word. By contrast, the voicing of word-initial Latin [k] to [g] occurred in *colaphus>golpe* and *cattus>gato* but not in *canna>caña*. See also lexical diffusion.

Sound change is inevitable: All languages vary from place to place and time to time, and neither writing nor media prevents that change.

Formal notation

A statement of the form

- A > B

is to be read, "Sound A changes into (or is replaced by, is reflected as, etc) sound B". Therefore, A belongs to an older stage of the language in question, and B belongs to a more recent stage. The symbol ">" can be reversed, B < A, which also means that the (more recent) B derives from the (older) A":

- POc. *t > Rot. f
- means that "Proto-Oceanic (POc.) *t is reflected as [f] in the Rotuman (Rot.)".

The two sides of such a statement indicate only the start and the end of the change, but additional intermediate stages may have occurred. The example above is actually a compressed account of a *sequence* of changes: *[t] first changed to [θ] (like the initial consonant of English*thin*), which has since yielded [f] and can be represented more fully:

- t >θ> f

Unless a change operates unconditionally (in all environments), the context in which it applies must be specified:

- A > B /X__Y
- = "A changes to B when it is preceded by X and followed by Y."

For example:

- It. b > v /[vowel]__[vowel], which can be simplified to just
- It. b > v /V__V (in which the V stands for any vowel)
- = "Intervocalic [b] (inherited from Latin) became [v] in Italian" (such as in *caballum, dēbet>cavallo* 'horse', *deve* 'owe (3rd pers. sing.)'

Here is a second example:

- PIr. [–cont][–voi] > [+cont]/__[C][+cont]
- = "A preconsonantal voiceless non-continuant (voiceless stop) changed into corresponding a voiceless continuant (fricative) in Proto-Iranian (PIr.)" when it was immediately followed by a continuant consonant (a resonant or a fricative): Proto-Indo-Iranian **pra* 'forth' >Avestan*fra*; **trayas* "three" (masc. nom. pl.)> Av. *θrayō*; **čatwāras* "four" (masc. nom. pl.) > Av. *čaθwārō*; **pśaws* "of a cow" (nom. **paśu*) > Av. *fšāoš* (nom. *pasu*). Note that the fricativization did not occur before stops and so **sapta* "seven" > Av. *hapta*. (However, in the variety of Iranian that led to Old Persian, fricativization occurred in all clusters: Old Persian *hafta* "seven".)

The symbol "#" stands for a word boundary (initial or final) and so the notation "/__#" means "word-finally", and "/#__" means "word-initially":

- Gk. [stop] >Ø /__#
- = "Word-final stops were deleted in Greek (*Gk.*)".

That can be simplified to

- Gk. P >Ø / __#

in which P stands for any plosive.

Terms for changes in pronunciation

In historical linguistics, a number of traditional terms designate types of phonetic change, either by nature or result. A number of such types are often (or usually) sporadic, that is, more or less accidents that happen to a specific form. Others affect a whole phonological system. Sound changes that affect a whole phonological system are also classified according to how they affect the overall shape of the system; see *phonological change*.

- Assimilation: One sound becomes more like another, or (much more rarely) two sounds become more like each other. Example: in Latin the prefix **kom-* becomes *con-* before an apical stop ([t d]) or [n]: *contactus* "touched", *condere* "to found, establish", *connūbium* "legal marriage". The great majority of assimilations take place between contiguous segments, and the great majority involve the earlier sound becoming more like the later one (e.g. in *connūbium*,*m-* + *n* becomes *-nn-* rather than *-mm-*). Assimilation between contiguous segments are

(diachronically speaking) exceptionless sound laws rather than sporadic, isolated changes.

- Dissimilation: The opposite of assimilation. One sound becomes less like another, or (much more rarely) two sounds become less like each other. Examples: Classical Latin *quīnque*/kʷiːnkʷe/ "five" > Vulgar Latin **kinkʷe* (whence French *cinq*, Italian *cinque*, etc.);Old Spanish*omne* "man" > Spanish *hombre*. The great majority of dissimilations involve segments that are **not** contiguous, but, as with assimilations, the great majority involve an earlier sound changing with reference to a later one. Dissimilation is usually a sporadic phenomenon, but Grassmann's Law (in Sanskrit and Greek) exemplifies a systematic dissimilation. If the change of a sequence of fricatives such that one becomes a stop is dissimilation, then such changes as Proto-Germanic *hs to /ks/ (spelled *x*) in English would count as a regular sound law: PGmc. **sehs* "six" >Old English*siex*, etc.
- Metathesis: Two sounds switch places. Example: Old English *th**ri**dda* became Middle English *th**ir**d*. Most such changes are sporadic, but occasionally a sound law is involved, as Romance **tl*> Spanish *ld*, thus **kapitlu, *titlu* "chapter (of a cathedral)", "tittle" > Spanish *cabildo, tilde*. Metathesis can take place between non-contiguous segments, as Greek *amélgō* "I milk" > Modern Greek *armégō*.
- Lenition, softening of a consonant, e.g. stop consonant to affricate or fricative; and its antonymfortition, hardening of a consonant.

- Tonogenesis: Syllables come to have distinctive pitch contours.
- Sandhi: conditioned changes that take place at word-boundaries but not elsewhere. It can be morpheme-specific, as in the loss of the vowel in the enclitic forms of English *is*/ɪz/, with subsequent change of /z/ to /s/ adjacent to a voiceless consonant *Frank's not here*/ˈfræŋksnɒtˈhɪər/. Or a small class of elements, such as the assimilation of the /ð/ of English *the, this* and *that* to a preceding /n/ (including the /n/ of *and* when the /d/ is elided) or /l/: *all the* often /ɔːllə/, *in the* often /ɪnnə/, and so on. As in these examples, such features are rarely indicated in standard orthography. In a striking exception, Sanskrit orthography reflects a wide variety of such features; thus, *tat* "that" is written *tat*,*tac*, *taj*, *tad*,or *tan* depending on what the first sound of the next word is. These are all assimilations, but medial sequences do not assimilate the same way.
- Haplology: The loss of a syllable when an adjacent syllable is similar or (rarely) identical. Example: Old English *Englaland* became Modern English *England*, or the common pronunciation of *probably* as [ˈprɒbli]. This change usually affects commonly used words. The word haplology itself is sometimes jokingly pronounced "haplogy".
- Elision, aphaeresis, syncope, and apocope: all losses of sounds. Elision is the loss of unstressed sounds, aphaeresis the loss of initial sounds, syncope is the loss of medial sounds, and apocope is the loss of final sounds.

- Elision examples: in the southeastern United States, unstressed schwas tend to drop, so "American" is not /əˈmɛɹəkən/ but /ˈmɚkən/. Standard English is *possum*<*opossum*.
- Syncope examples: the Old French word for "state" is *estat*, but the *s* disappeared, yielding *état*. Similarly, the loss of /t/ in English *soften, hasten, castle*, etc.
- Apocope examples: the final -*e*[ə] in Middle English words was pronounced, but is only retained in spelling as a silent E. In English /b/ and /ɡ/ were apocopated in final position after nasals: *lamb, long*/læm/, /lɒŋ ~ lɔːŋ/.
- Epenthesis (also known as anaptyxis): The introduction of a sound between two adjacent sounds. Examples: Latin *humilis*> English *humble*; in Slavic an -l- intrudes between a labial and a following yod, as **zemya* "land" > Russian *zemlya* (земля). Most commonly, epenthesis is in the nature of a "transitional" consonant, but vowels may be epenthetic: non-standard English *film* in two syllables, *athlete* in three. Epenthesis can be regular, as when the Indo-European "tool" suffix *-*tlom* everywhere becomes Latin -*culum* (so *speculum* "mirror" < **speē tlom*, *pōculum* "drinking cup" < **poH*$_3$*-tlom*). Some scholars reserve the term *epenthesis* for "intrusive" vowels and use *excrescence* for intrusive consonants.
- Prothesis: The addition of a sound at the beginning of a word. Example: word-initial /s/ + stop clusters in Latin gained a preceding /e/ in Old Spanish and Old French; hence, the Spanish word for "state" is *estado*, deriving from Latin *status*.

- Nasalization: Vowels followed by nasal consonants can become nasalized. If the nasal consonant is lost but the vowel retains its nasalized pronunciation, nasalization becomes phonemic, that is, distinctive. Example: French "-in" words used to be pronounced [in], but are now pronounced [ɛ̃], and the [n] is no longer pronounced (except in cases of liaison).

Examples of specific historical sound changes

- Anglo-Frisian nasal spirant law
- Canaanite shift
- Dahl's law
- Grassmann's law
- Great Vowel Shift (English)
- Grimm's law
- High German consonant shift
- Kluge's law
- Ruki sound law
- Slavic palatalization
- Sound change in Japanese
- Umlaut
- Verner's law

Comparative literature

Comparative literature is an academic field dealing with the study of literature and cultural expression across linguistic, national, geographic, and disciplinary boundaries. Comparative literature "performs a role similar to that of the study of

international relations, but works with languages and artistic traditions, so as to understand cultures 'from the inside'". While most frequently practised with works of different languages, comparative literature may also be performed on works of the same language if the works originate from different nations or cultures among which that language is spoken.

The characteristically intercultural and transnational field of comparative literature concerns itself with the relation between literature, broadly defined, and other spheres of human activity, including history, politics, philosophy, art, and science. Unlike other forms of literary study, comparative literature places its emphasis on the interdisciplinary analysis of social and cultural production within the "economy, political dynamics, cultural movements, historical shifts, religious differences, the urban environment, international relations, public policy, and the sciences".

Overview

Students and instructors in the field, usually called "comparatists", have traditionally been proficient in several languages and acquainted with the literary traditions, literary criticism, and major literary texts of those languages. Many of the newer sub-fields, however, are more influenced by critical theory and literary theory, stressing theoretical acumen and the ability to consider different types of art concurrently, over proficiency in multiple languages.

The interdisciplinary nature of the field means that comparatists typically exhibit acquaintance with sociology,

history, anthropology, translation studies, critical theory, cultural studies, and religious studies. As a result, comparative literature programs within universities may be designed by scholars drawn from several such departments. This eclecticism has led critics (from within and without) to charge that Comparative Literature is insufficiently well-defined, or that comparatists too easily fall into dilettantism, because the scope of their work is, of necessity, broad. Some question whether this breadth affects the ability of Ph.D.s to find employment in the highly specialized environment of academia and the career market at large, although such concerns do not seem to be borne out by placement data that shows comparative literature graduates to be hired at similar or higher rates than their peers in English.

The terms "comparative literature" and "world literature" are often used to designate a similar course of study and scholarship. Comparative Literature is the more widely used term in the United States, with many universities having Comparative Literature departments or Comparative Literature programs.

Comparative literature is an interdisciplinary field whose practitioners study literature across national borders, across time periods, across languages, across genres, across boundaries between literature and the other arts (music, painting, dance, film, etc.), across disciplines (literature and psychology, philosophy, science, history, architecture, sociology, politics, etc.). Defined most broadly, comparative literature is the study of "literature without borders". Scholarship in comparative literature include, for example, studying literacy and social status in the Americas, studying

medieval epic and romance, studying the links of literature to folklore and mythology, studying colonial and postcolonial writings in different parts of the world, asking fundamental questions about definitions of literature itself. What scholars in comparative literature share is a desire to study literature beyond national boundaries and an interest in languages so that they can read foreign texts in their original form. Many comparatists also share the desire to integrate literary experience with other cultural phenomena such as historical change, philosophical concepts, and social movements.

The discipline of comparative literature has scholarly associations such as the International Comparative Literature Association (ICLA) and comparative literature associations exist in many countries. There are many learned journals that publish scholarship in comparative literature: see "Selected Comparative Literature and Comparative Humanities Journals" and for a list of books in comparative literature see "Bibliography of (Text)Books in Comparative Literature".

Early work

Work considered foundational to the discipline of comparative literature include Spanish humanist Juan Andrés's work, Transylvanian Hungarian Hugo Meltzl de Lomnitz's scholarship, also the founding editor of the journal *Acta Comparationis Litterarum Universarum* (1877) and Irish scholar H.M. Posnett's *Comparative Literature* (1886). However, antecedents can be found in the ideas of Johann Wolfgang von Goethe in his vision of "world literature" (*Weltliteratur)* and Russian Formalists credited Alexander Veselovsky with laying the groundwork for the discipline. Viktor Zhirmunsky, for

instance, referred to Veselovsky as "the most remarkable representative of comparative literary study in Russian and European scholarship of the nineteenth century" (Zhirmunsky qtd. in Rachel Polonsky, *English Literature and the Russian Aesthetic Renaissance* [Cambridge UP, 1998. 17]; see also David Damrosch During the late 19th century, comparatists such as Fyodor Buslaev were chiefly concerned with deducing the purported Zeitgeist or "spirit of the times", which they assumed to be embodied in the literary output of each nation. Although many comparative works from this period would be judged chauvinistic, Eurocentric, or even racist by present-day standards, the intention of most scholars during this period was to increase the understanding of other cultures, not to assert superiority over them (although politicians and others from outside the field sometimes used their works for this purpose).

French School

From the early part of the 20th century until WWII, the field was characterised by a notably empiricist and positivist approach, termed the "French School", in which scholars like Paul Van Tiegham examined works forensically, looking for evidence of "origins" and "influences" between works from different nations often termed "rapport des faits". Thus a scholar might attempt to trace how a particular literary idea or motif traveled between nations over time. In the French School of Comparative Literature, the study of influences and mentalities dominates. Today, the French School practices the nation-state approach of the discipline although it also promotes the approach of a "European Comparative Literature". The publications from this school include, *La*

Littérature Comparée (1967) by C. Pichois and A.M. Rousseau, *La Critique Littéraire* (1969) by J.-C. Carloni and Jean Filloux and *La Littérature Comparée* (1989) by Yves Cheverel, translated into English as *Comparative Literature Today: Methods & Perspectives* (1995).

German School

Like the French School, German Comparative Literature has its origins in the late 19th century. After World War II, the discipline developed to a large extent owing to one scholar in particular, Peter Szondi (1929–1971), a Hungarian who taught at the Free University Berlin. Szondi's work in Allgemeine und Vergleichende Literaturwissenschaft (German for "General and Comparative Literary Studies") included the genre of drama, lyric (in particular hermetic) poetry, and hermeneutics: "Szondi's vision of Allgemeine und Vergleichende Literaturwissenschaft became evident in both his policy of inviting international guest speakers to Berlin and his introductions to their talks. Szondi welcomed, among others, Jacques Derrida (before he attained worldwide recognition), Pierre Bourdieu and Lucien Goldman from France, Paul de Man from Zürich, Gershom Sholem from Jerusalem, Theodor W. Adorno from Frankfurt, Hans Robert Jauss from the then young University of Konstanz, and from the US René Wellek, Geoffrey Hartman and Peter Demetz (all at Yale), along with the liberal publicist Lionel Trilling. The names of these visiting scholars, who form a programmatic network and a methodological canon, epitomise Szondi's conception of comparative literature. German comparatists working in East Germany, however, were not invited, nor were recognised colleagues from France or the Netherlands. Yet while he was

oriented towards the West and the new allies of West Germany and paid little attention to comparatists in Eastern Europe, his conception of a transnational (and transatlantic) comparative literature was very much influenced by East European literary theorists of the Russian and Prague schools of structuralism, from whose works René Wellek, too, derived many of his concepts, concepts that continue to have profound implications for comparative literary theory today" ... A manual published by the department of comparative literature at the LMU Munich lists 31 German departments which offer a diploma in comparative literature in Germany, albeit some only as a 'minor'. These are: Augsburg, Bayreuth, Free University Berlin, Technical University Berlin, Bochum, Bonn, Chemnitz-Zwickau, Erfurt, Erlangen-Nürnberg, Essen, Frankfurt am Main, Frankfurt an der Oder,

Gießen, Göttingen, Jena, Karlsruhe, Kassel, Konstanz, Leipzig, Mainz, München, Münster, Osnabrück, Paderborn, Potsdam, Rostock, Saarbrücken, Siegen, Stuttgart, Tübingen, Wuppertal. (Der kleine Komparatist [2003]). This situation is undergoing rapid change, however, since many universities are adapting to the new requirements of the recently introduced Bachelor and Master of Arts.

German comparative literature is being squeezed by the traditional philologies on the one hand and more vocational programmes of study on the other which seek to offer students the practical knowledge they need for the working world (e.g., 'Applied Literature'). With German universities no longer educating their students primarily for an academic market, the necessity of a more vocational approach is becoming ever more evident".

American (US) School

Reacting to the French School, postwar scholars, collectively termed the "American School", sought to return the field to matters more directly concerned with literary criticism, de-emphasising the detective work and detailed historical research that the French School had demanded. The American School was more closely aligned with the original internationalist visions of Goethe and Posnett (arguably reflecting the postwar desire for international cooperation), looking for examples of universal human "truths" based on the literary archetypes that appeared throughout literatures from all times and places.

Prior to the advent of the American School, the scope of comparative literature in the West was typically limited to the literatures of Western Europe and Anglo-America, predominantly literature in English, German and French literature, with occasional forays into Italian literature (primarily for Dante) and Spanish literature (primarily for Miguel de Cervantes). One monument to the approach of this period is Erich Auerbach's book *Mimesis: The Representation of Reality in Western Literature*, a survey of techniques of realism in texts whose origins span several continents and three thousand years.

The approach of the American School would be familiar to current practitioners of cultural studies and is even claimed by some to be the forerunner of the Cultural Studies boom in universities during the 1970s and 1980s. The field today is highly diverse: for example, comparatists routinely study Chinese literature, Arabic literature and the literatures of most

other major world languages and regions as well as English and continental European literatures.

Current developments

There is a movement among comparativists in the United States and elsewhere to re-focus the discipline away from the nation-based approach with which it has previously been associated towards a cross-cultural approach that pays no heed to national borders. Works of this nature include Alamgir Hashmi's *The Commonwealth, Comparative Literature and the World*, Gayatri Chakravorty Spivak's *Death of a Discipline*, David Damrosch's *What is World Literature?*, Steven Tötösy de Zepetnek's concept of "comparative cultural studies", and Pascale Casanova's *The World Republic of Letters*. It remains to be seen whether this approach will prove successful given that comparative literature had its roots in nation-based thinking and much of the literature under study still concerns issues of the nation-state. Given developments in the studies of globalization and interculturalism, comparative literature, already representing a wider study than the single-language nation-state approach, may be well suited to move away from the paradigm of the nation-state. While in the West comparative literature is experiencing institutional constriction, there are signs that in many parts of the world the discipline is thriving, especially in Asia, Latin America, the Caribbean, and the Mediterranean. Current trends in Transnational studies also reflect the growing importance of post-colonial literary figures such as J. M. Coetzee, Maryse Condé, Earl Lovelace, V. S. Naipaul, Michael Ondaatje, Wole Soyinka, Derek Walcott, and Lasana M. Sekou. For recent post-colonial studies in North America see George Elliott Clarke.

Directions Home: Approaches to African-Canadian Literature. (University of Toronto Press, 2011), Joseph Pivato. *Echo: Essays in Other Literatures*. (Guernica Editions, 2003), and "The Sherbrooke School of Comparative Canadian Literature". (*Inquire*, 2011). In the area of comparative studies of literature and the other arts see Linda Hutcheon's work on Opera and her *A Theory of Adaptation*. 2nd. ed. (Routledge, 2012). Canadian scholar Joseph Pivato is carrying on a campaign to revitalize comparative study with his book, *Comparative Literature for the New Century* eds. Giulia De Gasperi & Joseph Pivato (2018). In response to Pivato Canadian comparatists Susan Ingram and Irene Sywenky co-edited *Comparative Literature in Canada: Contemporary Scholarship, Pedagogy, and Publishing in Review* (2019), an initiative of the Canadian Comparative Literature Association.

Chapter 5

International Auxiliary Language

An **international auxiliary language** (sometimes abbreviated as **IAL** or **auxlang**) is a language meant for communication between people from different nations who do not share a common first language. An auxiliary language is primarily a foreign language. It usually takes words from widely spoken languages.

Languages of dominant societies over the centuries have served as lingua francas that have sometimes approached the international level. Latin, Greek, Sanskrit, Persian, Old Tamil and the Mediterranean Lingua Franca were used in the past, and Standard Arabic, Standard Chinese, English, French, Portuguese, Hindustani (Hindi-Urdu), Russian and Spanish have been used as such in recent times in many parts of the world. However, as lingua francas are traditionally associated with the very dominance—cultural, political, and economic—that made them popular, they are often also met with resistance. For this and other reasons, some have turned to the idea of promoting an artificial or constructed language as a possible solution, by way of an "auxiliary" language. The term "auxiliary" implies that it is intended to be an additional language for the people of the world, rather than to replace their native languages. Often, the term is used to refer to planned or constructed languages proposed specifically to ease international communication, such as Esperanto, Ido and Interlingua. However, it can also refer to the concept of such a language being determined by international consensus,

including even a standardized natural language (e.g., International English), and has also been connected to the project of constructing a universal language.

History

The use of an intermediary auxiliary language (also called a "working language", "bridge language", "vehicular language" or "unifying language") to make communication possible between people not sharing a first language, in particular when it is a third language, distinct from both mother tongues, may be almost as old as language itself. Certainly they have existed since antiquity. Latin and Greek (or Koine Greek) were the intermediary language of all areas of the Mediterraneum; Akkadian, and then Aramaic, remained the common languages of a large part of Western Asia through several earlier empires. Such natural languages used for communication between people not sharing the same mother tongue are called*lingua francas*.

Natural international languages: Lingua francas

Lingua francas have arisen around the globe throughout human history, sometimes for commercial reasons (so-called "trade languages") but also for diplomatic and administrative convenience, and as a means of exchanging information between scientists and other scholars of different nationalities. The term originates with one such language, Mediterranean Lingua Franca, a pidgin language used as a trade language in the Mediterranean area from the 11th to the 19th century. Examples of lingua francas remain numerous, and exist on every continent. The most obvious example as of the early 21st

century is English. Moreover, a special case of English is that of Basic English, a simplified version of English which shares the same grammar (though simplified) and a reduced vocabulary of only 1,000 words, with the intention that anyone with a basic knowledge of English should be able to understand even quite complex texts. There are many other lingua francas centralized on particular regions, such as Arabic, Chinese, French, Greek, Hindi, Portuguese, Russian and Spanish.

Constructed languages

Since all natural languages display a number of irregularities in grammar that make them more difficult to learn, and they are also associated with the national and cultural dominance of the nation that speaks it as its mother tongue, attention began to focus on the idea of creating an artificial or constructed language as a possible solution. The concept of simplifying an existing language to make it an auxiliary language was already in the *Encyclopédie* of the 18th century, where Joachim Faiguet de Villeneuve, in the article on *Langue*, wrote a short proposition of a "laconic" or regularized grammar of French.

Some of the philosophical languages of the 17th–18th centuries could be regarded as proto-auxlangs, as they were intended by their creators to serve as bridges among people of different languages as well as to disambiguate and clarify thought. However, most or all of these languages were, as far as can be told from the surviving publications about them, too incomplete and unfinished to serve as auxlangs (or for any other practical purpose). The first fully developed constructed

languages we know of, as well as the first constructed languages devised primarily as auxlangs, originated in the 19th century; Solresol by François Sudre, a language based on musical notes, was the first to gain widespread attention although not, apparently, fluent speakers.

Volapük

During the 19th century, a bewildering variety of such constructed international auxiliary languages (IALs) were proposed, so Louis Couturatand Léopold Leau in *Histoire de la langue universelle* (1903) reviewed 38 projects.

Volapük, first described in an article in 1879 by Johann Martin Schleyer and in book form the following year, was the first to garner a widespread international speaker community. Three major Volapük conventions were held, in 1884, 1887, and 1889; the last of them used Volapük as its working language. André Cherpillod writes of the third Volapük convention,

In August 1889 the third convention was held in Paris. About two hundred people from many countries attended. And, unlike in the first two conventions, people spoke only Volapük. For the first time in the history of mankind, sixteen years before the Boulogne convention, an international convention spoke an international language.

However, not long after, the Volapük speaker community broke up due to various factors including controversies between Schleyer and other prominent Volapük speakers, and the appearance of newer, easier-to-learnconstructed languages, primarily Esperanto.

From *Kadem bevünetik volapüka* to *Academia pro Interlingua*

Answering the needs of the first successful artificial language community, the Volapükists established the regulatory body of their language, under the name International Volapük Academy (*Kadem bevünetik volapüka*) at the second Volapük congress in Munich in August 1887. The Academy was set up to conserve and perfect the auxiliary language Volapük, but soon conflicts arose between conservative Volapükists and those who wanted to reform Volapük to make it a more naturalistic language based on the grammar and vocabulary of major world languages. In 1890 Schleyer himself left the original Academy and created a new Volapük Academy with the same name, from people completely loyal to him, which continues to this day.

Under Waldemar Rosenberger, who became the director in 1892, the original Academy began to make considerable changes in the grammar and vocabulary of Volapük. The vocabulary and the grammatical forms unfamiliar to Western Europeans were completely discarded, so that the changes effectively resulted in the creation of a new language, which was named "Idiom Neutral". The name of the Academy was changed to *Akademi Internasional de Lingu Universal* in 1898 and the circulars of the Academy were written in the new language from that year.

In 1903, the mathematician Giuseppe Peano published his completely new approach to language construction. Inspired by the idea of philosopher Gottfried Wilhelm Leibniz, instead of inventing schematic structures and an *a priori* language, he chose to simplify an existing and once widely used

international language, Latin. This simplified Latin, devoid of inflections and declensions, was named*Interlingua* by Peano but is usually referred to as "Latino sine flexione".

Impressed by Peano's Interlingua, the *Akademi Internasional de Lingu Universal* effectively chose to abandon Idiom Neutral in favor of Peano's Interlingua in 1908, and it elected Peano as its director. The name of the group was subsequently changed to *Academia pro Interlingua* (where *Interlingua* stands for Peano's language). The *Academia pro Interlingua* survived until about 1939. It was Peano's Interlingua that partly inspired the better-known Interlingua presented in 1951 by the International Auxiliary Language Association (IALA).

Esperanto

After the emergence of Volapük, a wide variety of other auxiliary languages were devised and proposed in the 1880s–1900s, but none except Esperanto gathered a significant speaker community. Esperanto was developed from about 1873–1887 (a first version was ready in 1878), and finally published in 1887, by L. L. Zamenhof, as a primarily schematic language; the word-stems are borrowed from Romance, West Germanic and Slavic languages. The key to the relative success of Esperanto was probably the highly productive and elastic system of derivational word formation which allowed speakers to derive hundreds of other words by learning one word root. Moreover, Esperanto is quicker to learn than other languages, usually in a third up to a fifth of the time. From early on, Esperantists created their own culture which helped to form the Esperanto language community.

Within a few years this language had thousands of fluent speakers, primarily in eastern Europe. In 1905 its first world convention was held in Boulogne-sur-Mer. Since then world congresses have been held in different countries every year, except during the two World Wars. Esperanto has become "the most outlandishly successful invented language ever" and the most widely spoken constructed international auxiliary language. Esperanto is probably among the fifty languages which are most used internationally.

In 1922 a proposal by Iran and several other countries in the League of Nations to have Esperanto taught in member nations' schools failed. Esperanto speakers were subject to persecution under Stalin's regime. In Germany under Hitler, in Spain under Franco for about a decade, in Portugal under Salazar, in Romania under Ceaușescu, and in half a dozen Eastern European countries during the late forties and part of the fifties, Esperanto activities and the formation of Esperanto associations were forbidden.

In spite of these factors more people continued to learn Esperanto, and significant literary work (both poetry and novels) appeared in Esperanto in the period between the World Wars and after them.

Esperanto is spoken today in a growing number of countries and it has multiple generations of native speakers, although it is primarily used as a second language. Of the various constructed language projects, it is Esperanto that has so far come closest to becoming an officially recognized international auxiliary language; China publishes daily news in Esperanto.

Ido and the Esperantidos

The Delegation for the Adoption of an International Auxiliary Language was founded in 1900 by Louis Couturat and others; it tried to get the International Association of Academies to take up the question of an international auxiliary language, study the existing ones and pick one or design a new one. However, the meta-academy declining to do so, the Delegation decided to do the job itself. Among Esperanto speakers there was a general impression that the Delegation would of course choose Esperanto, as it was the only auxlang with a sizable speaker community at the time; it was felt as a betrayal by many Esperanto speakers when in 1907 the Delegation came up with its own reformed version of Esperanto, Ido. Ido drew a significant number of speakers away from Esperanto in the short term, but in the longer term most of these either returned to Esperanto or moved on to other new auxlangs. Besides Ido, a great number of simplified Esperantos, called Esperantidos, emerged as concurrent language projects; still, Ido remains today one of the three most widely spoken auxlangs.

Interlingue (Occidental)

Edgar de Wahl's Occidental of 1922 was in reaction against the perceived artificiality of some earlier auxlangs, particularly Esperanto. Inspired by Idiom Neutral and Latino sine flexione, de Wahl created a language whose words, including compound words, would have a high degree of recognizability for those who already know a Romance language. However, this design criterion was in conflict with the ease of coining new compound or derived words on the fly while speaking.

Occidental was most active from the 1920s to the 1950s, and supported some 80 publications by the 1930s, but had almost entirely died out by the 1980s. Its name was officially changed to Interlingue in 1949. More recently Interlingue has been revived on the Internet.

Novial

In 1928 Ido's major intellectual supporter, the Danish linguist Otto Jespersen, abandoned Ido, and published his own planned language, Novial. It was mostly inspired by Idiom Neutral and Occidental, yet it attempted a derivational formalism and schematism sought by Esperanto and Ido. The notability of its creator helped the growth of this auxiliary language, but a reform of the language was proposed by Jespersen in 1934 and not long after this Europe entered World War II, and its creator died in 1943 before Europe was at peace again.

Interlingua

The International Auxiliary Language Association (IALA) was founded in 1924 by Alice Vanderbilt Morris; like the earlier *Delegation for the Adoption of an International Auxiliary Language*, its mission was to study language problems and the existing auxlangs and proposals for auxlangs, and to negotiate some consensus between the supporters of various auxlangs. However, like the Delegation, it finally decided to create its own auxlang. Interlingua, published in 1951, was primarily the work of Alexander Gode, though he built on preliminary work by earlier IALA linguists including André Martinet, and relied on elements from previous naturalistic auxlang projects, like

Peano's Interlingua (Latino sine flexione), Jespersen's Novial, de Wahl's Interlingue, and the Academy's Idiom Neutral. Like Interlingue, Interlingua was designed to have words recognizable at sight by those who already know a Romance language or a language like English with much vocabulary borrowed from Romance languages; to attain this end the IALA accepted a degree of grammatical and orthographic complexity considerably greater than in Esperanto or Interlingue, though still less than in any natural language.

The theory underlying Interlingua posits an *international vocabulary*, a large number of words and affixes that are present in a wide range of languages. This already existing international vocabulary was shaped by social forces, science and technology, to "all corners of the world". The goal of the International Auxiliary Language Association was to accept into Interlingua every widely international word in whatever languages it occurred. They conducted studies to identify "the most generally international vocabulary possible", while still maintaining the unity of the language. This scientific approach of generating a language from selected source languages (called *control languages*) resulted in a vocabulary and grammar that can be called the highest common factor of each major European language.

Interlingua gained a significant speaker community, perhaps roughly the same size as that of Ido (considerably less than the size of Esperanto). Interlingua's success can be explained by the fact that it is the most widely *understood* international auxiliary language by virtue of its naturalistic (as opposed to schematic) grammar and vocabulary, allowing those familiar with a Romance language, and educated speakers of English,

to read and understand it without prior study. Interlingua has some active speakers currently on all continents, and the language is propagated by the Union Mundial pro Interlingua (UMI), and Interlingua is presented on CDs, radio, and television.

After the creation of Interlingua, the enthusiasm for constructed languages gradually decreased in the years between 1960 and 1990.

Internet age

All of the auxlangs with a surviving speaker community seem to have benefited from the advent of the Internet, Esperanto more than most. The CONLANG mailing list was founded in 1991; in its early years discussion focused on international auxiliary languages.

As people interested in artistic languages and engineered languages grew to be the majority of the list members, and flame-wars between proponents of particular auxlangs irritated these members, a separate AUXLANG mailing listwas created in 1997, which has been the primary venue for discussion of auxlangs since then. Besides giving the existing auxlangs with speaker communities a chance to interact rapidly online as well as slowly through postal mail or more rarely in personal meetings, the Internet has also made it easier to publicize new auxlang projects, and a handful of these have gained a small speaker community, including Kotava (published in 1978), Lingua Franca Nova (1998), Interslavic (2006), Pandunia (2007), Sambahsa (2007), Lingwa de Planeta (2010), and Globasa (2019).

Zonal constructed languages

Not every international auxiliary language is necessarily intended to be used on a global scale. A special subgroup are languages created to facilitate communication between speakers of related languages. The oldest known example is a Pan-Slavic language written in 1665 by the Croatian priest Juraj Križanić. He named this language Ruski ("Russian"), although in reality it was a mixture of the Russian edition of Church Slavonic, his own Southern Chakavian dialect of Serbo-Croatian, and, to a lesser degree, Polish. Most zonal constructed languages were created during the period of romantic nationalism at the end of the 19th century; some were created later. Particularly numerous are the Pan-Slavic language projects. However, similar efforts at creating umbrella languages have been made for other language families as well: Tutonish (1902), Folkspraak (1995) and other pan-Germanic languages for the Germanic languages; Romanid (1956) and several other pan-Romance languages for the Romance languages; and Afrihili (1973) for the African continent. Notable among modern examples is Interslavic, a project first published in 2006 as Slovianski and then established in its current form in 2011 after the merger of several other projects. In 2012 it was reported to have several hundred users.

Scholarly study

In the early 1900s auxlangs were already becoming a subject of academic study. Louis Couturat et al. described the controversy in the preface to their book *International Language and Science*:

- The question of a so-called world-language, or better expressed, an international auxiliary language, was during the now past Volapük period, and is still in the present Esperanto movement, so much in the hands of Utopians, fanatics and enthusiasts, that it is difficult to form an unbiased opinion concerning it, although a good idea lies at its basis. (1910, p. v).

Leopold Pfaundler wrote that an IAL was needed for more effective communication among scientists:

- All who are occupied with the reading or writing of scientific literature have assuredly very often felt the want of a common scientific language, and regretted the great loss of time and trouble caused by the multiplicity of languages employed in scientific literature.

For Couturat et al., Volapükists and Esperantists confounded the linguistic aspect of the question with many side issues, and they considered this a main reason why discussion about the idea of an international auxiliary language has appeared unpractical.

Some contemporaries of Couturat, notably Edward Sapir saw the challenge of an auxiliary language not as much as that of identifying a descriptive linguistic answer (of grammar and vocabulary) to global communicative concerns, but rather as one of promoting the notion of a linguistic platform for lasting international understanding. Though interest among scholars, and linguists in particular, waned greatly throughout the 20th century, such differences of approach persist today. Some scholars and interested laymen make concrete language

proposals. By contrast, Mario Pei and others place the broader societal issue first. Yet others argue in favor of a particular language while seeking to establish its social integration.

Writing systems

Whilst most IAL use the Latin script, some of them, such as LFN, also offer an alternative in the Cyrillic script.

Latin script

The vast majority of IALs use the Latin script. Several sounds, e.g. /n/, /m/, /t/, /f/ are written with the same letter as in IPA.

Some consonant sounds found in several Latin-script IAL alphabets are not represented by an ISO 646 letter in IPA. Three have a single letter in IPA, one has a widespread alternative taken from ISO 646:

- /ʃ/ (U+0283, IPA 134)
- /ʒ/ (U+0292, IPA 135)
- /ɡ/ (U+0261, IPA 110, single storey g) = g (U+0067, double storey g)

Four are affricates, each represented in IPA by two letters and a combining marker. They are often written decomposed:

- /t͡s/ = /ts/
- /tʃ/ = /tʃ/; Note: Polish distinguishes between them
- /d͡z/ = /dz/
- /d͡ʒ/ = /dʒ/

That means that two sounds that are one character in IPA and are not ISO 646, also have no common alternative in ISO 646: ʃ, ʒ.

Classification

The following classification of auxiliary languages was developed by Pierre Janton in 1993:

- *A priori* languagesare characterized by largely artificial morphemes (not borrowed from natural languages), schematic derivation, simple phonology, grammar and morphology. Some *a priori* languages are calledphilosophical languages, referring to their basis in philosophical ideas about thought and language. These include some of the earliest efforts at auxiliary language in the 17th century. Some more specific subcategories:
- Taxonomic languages form their words using a taxonomic hierarchy, with each phoneme of a word helping specify its position in a semantic hierarchy of some kind; for example, Solresol.
- Pasigraphiesare purely written languages without a spoken form, or with a spoken form left at the discretion of the reader; many of the 17th–18th century philosophical languages and auxlangs were pasigraphies. This set historically tends to overlap with taxonomic languages, though there is no inherent reason a pasigraphy needs to be taxonomic.
- *A posteriori* languages are based on existing natural languages. Nearly all the auxiliary languages with fluent speakers are in this category. Most of the *a*

posteriori auxiliary languages borrow their vocabulary primarily or solely from European languages, and base their grammar more or less on European models. (Sometimes these European-based languages are referred to as "euroclones", although this term has negative connotations and is not used in the academic literature.) Interlinguawas drawn originally from international scientific vocabulary, in turn based primarily on Greek and Latin roots. Glosa did likewise, with a stronger dependence of Greek roots. Although *a posteriori* languages have been based on most of the families of European languages, the most successful of these (notably Esperanto, Ido and Interlingua) have been based largely on Romance elements.

- Schematic (or "mixed") languages have some *a priori* qualities. Some have ethnic morphemes but alter them significantly to fit a simplified phonotactic pattern (e.g., Volapük) or both artificial and natural morphemes (e.g., Perio). Partly schematic languages have partly schematic and partly naturalistic derivation (e.g. Esperanto and Ido). Natural morphemes of languages in this group are rarely altered greatly from their source-language form, but compound and derived words are generally not recognizable at sight by people familiar with the source languages.
- Naturalistic languages resemble existing natural languages. For example, Interlingue, Interlingua, and Lingua Franca Nova were developed so that not only the root words but their compounds and derivations will often be immediately recognized by large

numbers of people. Some naturalistic languages do have a limited number of artificial morphemes or invented grammatical devices (e.g. Novial).

- Simplified, or controlled versions of natural languages reduce the full extent of the vocabulary and partially regularize the grammar of a natural language (e.g. Basic English and Special English).

Methods of propagation

As has been pointed out, the issue of an international language is not so much which, but how. Several approaches exist toward the eventual full expansion and consolidation of an international auxiliary language.

- Laissez-faire. This approach is taken in the belief that one language will eventually and inevitably "win out" as a world auxiliary language (e.g. International English) without any need for specific action.
- Institutional sponsorship and grass-roots promotion of language programs. This approach has taken various forms, depending on the language and language type, ranging from government promotion of a particular language to one-on-one encouragement to learn the language to instructional or marketing programs.
- National legislation. This approach seeks to have individual countries (or even localities) progressively endorse a given language as an official language (or to promote the concept of international legislation).
- International legislation. This approach involves promotion of the future holding of a binding

international convention (perhaps to be under the auspices of such international organizations as the United Nations or Inter-Parliamentary Union) to formally agree upon an official international auxiliary language which would then be taught in all schools around the world, beginning at the primary level. This approach, an official principle of the Bahá'í Faith, seeks to put a combination of international opinion, linguistic expertise, and law behind a to-be-selected language and thus expand or consolidate it as a full official world language, to be used in addition to local languages. This approach could either give more credibility to a natural language already serving this purpose to a certain degree (e.g. if English were chosen) or to give a greatly enhanced chance for a constructed language to take root. For constructed languages particularly, this approach has been seen by various individuals in the IAL movement as holding the most promise of ensuring that promotion of studies in the language would not be met with skepticism at its practicality by its would-be learners.

Pictorial languages

There have been a number of proposals for using pictures, ideograms, diagrams, and other pictorial representations for international communications. Examples range from the original Characteristica Universalis proposed by the philosopher Leibniz, to suggestions for the adoption of Chinese writing, to recent inventions such as Blissymbol.

Within the scientific community, there is already considerable agreement in the form of the schematics used to represent electronic circuits, chemical symbols, mathematical symbols, and the Energy Systems Language of systems ecology. We can also see the international efforts at regularizing symbols used to regulate traffic, to indicate resources for tourists, and in maps. Some symbols have become nearly universal through their consistent use in computers and on the Internet.

Sign languages

An international auxiliary sign language has been developed by deaf people who meet regularly at international forums such as sporting events or in political organisations. Previously referred to as Gestuno but now more commonly known simply as 'international sign', the language has continued to develop since the first signs were standardised in 1973, and it is now in widespread use. International sign is distinct in many ways from spoken IALs; many signs are iconic, and signers tend to insert these signs into the grammar of their own sign language, with an emphasis on visually intuitive gestures and mime. A simple sign language called Plains Indian Sign Language was used by indigenous peoples of the Americas.

Gestuno is not to be confused with the separate and unrelated sign language Signuno, which is essentially a Signed Exact Esperanto. Signuno is not in any significant use, and is based on the Esperanto community rather than based on the international Deaf community.

Criticism

There has been considerable criticism of international auxiliary languages, both in terms of individual proposals, types of proposals, and in more general terms.

Much criticism has been focused either on the artificiality of international auxiliary languages, or on the argumentativeness of proponents and their failure to agree on one language, or even on objective criteria by which to judge them. However, probably the most common criticism is that a constructed auxlang is unnecessary because natural languages such as English are already in wide use as auxlangs and work well enough for that purpose.

One criticism already prevalent in the late 19th century, and still sometimes heard today, is that an international language might hasten the extinction of minority languages. One response has been that, even if this happens, the benefits would outweigh the costs.

Although referred to as *international* languages, most of these languages have historically been constructed on the basis ofWestern European languages. } Esperanto and other languages such as Interlingua and Ido have been criticized for being too European and not global enough. The term "Euroclone" was coined to refer to such languages in contrast to "worldlangs" with global vocabulary sources.