Toward a phylogenetic chronology of ancient Gaulish, Celtic, and Indo-European

Peter Forster*[†‡] and Alfred Toth[§]

*McDonald Institute for Archaeological Research, University of Cambridge, Cambridge CB2 3ER, United Kingdom; [†]Junge Akademie, 10117 Berlin, Germany; and [§]Phonogrammarchiv der Universität Zürich, CH-8032 Zürich, Switzerland

Edited by Henry C. Harpending, University of Utah, Salt Lake City, UT, and approved May 6, 2003 (received for review February 27, 2003)

Indo-European is the largest and best-documented language family in the world, yet the reconstruction of the Indo-European tree, first proposed in 1863, has remained controversial. Complications may include ascertainment bias when choosing the linguistic data, and disregard for the wave model of 1872 when attempting to reconstruct the tree. Essentially analogous problems were solved in evolutionary genetics by DNA sequencing and phylogenetic network methods, respectively. We now adapt these tools to linguistics, and analyze Indo-European language data, focusing on Celtic and in particular on the ancient Celtic language of Gaul (modern France), by using bilingual Gaulish–Latin inscriptions. Our phylogenetic network reveals an early split of Celtic within Indo-European. Interestingly, the next branching event separates Gaulish (Continental Celtic) from the British (Insular Celtic) languages, with Insular Celtic subsequently splitting into Brythonic (Welsh, Breton) and Goidelic (Irish and Scottish Gaelic). Taken together, the network thus suggests that the Celtic language arrived in the British Isles as a single wave (and then differentiated locally), rather than in the traditional two-wave scenario ("P-Celtic" to Britain and "Q-Celtic" to Ireland). The phylogenetic network furthermore permits the estimation of time in analogy to genetics, and we obtain tentative dates for Indo-European at 8100 BC ± 1,900 years, and for the arrival of Celtic in Britain at 3200 BC ± 1,500 years. The phylogenetic method is easily executed by hand and promises to be an informative approach for many problems in historical linguistics.

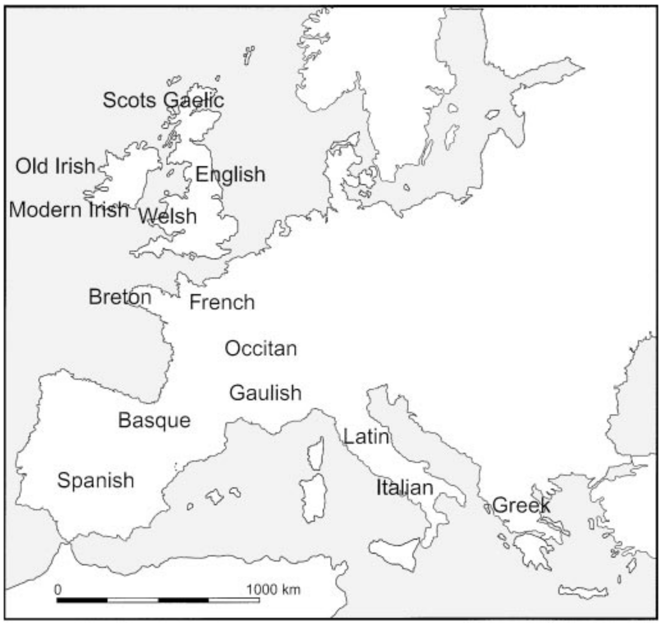

Fig. 1. Living and extinct languages referred to in this study.

The quest for reconstructing the prehistory of the Indo-European language family commenced in 1786 with the discovery by Sir William Jones of the remarkable similarities between Sanskrit, Greek, Latin, Gothic, Celtic, and Persian, indicating a "common source" for these languages (1). The next major step occurred in 1863, when Schleicher proposed an evolutionary tree of descent for the Indo-European language family (2), shortly after Charles Darwin had introduced the evolutionary tree concept to the descent of species. Further insight into language evolution was supplied by Schmidt in 1872 (3), who published the wave model according to which initially distinct languages increasingly acquire similarities through borrowing. More recently, we proposed a method for uniting these two models into a single network diagram of language evolution (4), which displays a tree if the analyzed languages have evolved in a strict branching process, but degenerates into a reticulate network if the data indicate borrowing and convergence. In Forster *et al.* (4), we tested this method on vocabulary lists of Alpine Romance languages, producing a network that revealed language subclusters in close agreement with the geographic locations of the Alpine valleys in which the languages are spoken. Moreover, the hypothetical ancestral language proposed by the method was directly validated by comparison with Latin. Our Alpine analysis reconstructed the past from synchronic data, in the sense that all of the used Romance languages either were current or went extinct only very recently. But the network method is equally applicable to reconstructing prehistoric evolutionary relationships from diachronic data, i.e., from languages of quite different time levels. We now exploit this feature to tackle afresh the reconstruction of the prehistoric tree (or network?) of Indo-European languages, whose ages of attestation span several millennia, from ancient Greek to, for example, modern English. We shall see that the network infers a tree, and therefore a hypothetical ancestral Indo-European language for which we provide a tentative phylogenetic age estimate.

Our particular focus will be the Celtic languages, including ancient Gaulish, formerly spoken in what is today France and northern Italy (Fig. 1). In western Europe, Gaulish is the only pre-Roman language with a significant bilingual corpus, and knowledge of its time depth and relationship to other languages would enable valuable comparisons with the time depth and landscape of western European archaeology and genetics. In AD 98, Tacitus recorded that between Britain and Gaul "the language differs but little" (*Agricola* 11). Nevertheless, classical sources excluded Britain from the "Celtic" designation bestowed on Gaul, compare Strabo in AD 18: "The men of Britain are taller than the Celti, and not so yellow-haired" (Strabo, *Geography* 4,5,2). Buchanan (5) and Lhuyd (6) proposed a relationship between Gaulish and the British languages, hence British languages are now conventionally termed "Insular Celtic," as opposed to "Continental Celtic" formerly spoken on the European mainland. Insular Celtic is subdivided into Brythonic (e.g., Welsh and Breton) and Goidelic (e.g., Irish and Scots Gaelic). However, there are conflicting proposals on the branching order and on relative and absolute dates of language splits in the Celtic language tree, if it is a tree at all. The underlying problems largely consist in the limited number and often uncertain translations of surviving Continental Celtic records (7). Extensive

This paper was submitted directly (Track II) to the PNAS office.

[‡]To whom correspondence should be addressed. E-mail: pf223@cam.ac.uk.

Table 1. Minimal glossary of Gaulish translated into European languages

Gaulish	English	Latin	Classical Greek	Old Irish	Mod. Irish	Mod. Scots Gaelic
Syntax: SV (*a*)	SV (*a*)	SV (*a*)	SV (*a*)	VS (*b*)	VS (*b*)	VS (*b*)
-OS (*a*)	(nom.sg.masc.suffix) (*b*)	-us (*a*)	−ος (*a*)	Absent (*b*)	Absent (*b*)	Absent (*b*)
-I (*a*)	(gen.sg.masc.suffix) -s (*b*)	-i (*a*)	−ου (*a*)	ICM, vowel change (*c*)	ICM, vowel change (*c*)	ICM, vowel change (*c*)
-V (*a*)	(dat.sg.masc. suffix) (*b*)	-o, -u (*a*)	−ω (*a*)	ICM, vowel change (*c*)	Absent (*b*)	Absent (*b*)
-A (*a*)	(nom.sg.fem.suffix) (*b*)	-a (*a*)	−η, −α (*a*)	ICM (*c*)	ICM (*c*)	ICM (*c*)
-AS (*a*)	(gen.sg.fem.suffix) -s (*a*)	-ae (*b*)	−ης, −ας (*a*)	Vowel change (*c*)	Vowel change/add. (*c*)	Vowel change (*c*)
parapsidi>paraxidi (*a*)	ps frequent (*b*)	ps frequent (*b*)	ps frequent (*b*)	ps rare (*a*)	ps rare (*a*)	ps rare (*a*)
TEUO- (*a*)	to gods (*b*)	deis (*a*)	θεοισι(ν) (*a*)	do déib (*a*)	do dhéithe (*a*)	do dhiadhan (*a*)
-XTONION (*a*)	and to men (*b*)	et hominibus (*c*)	και ανθρωποισι(ν) (*d*)	ocus do daínib (*a*)	agus do dhaoine (*a*)	agus do dhaoinean (*a*)
IEVRV, IOVRVS. . . (*a*)	has offered (*b*)	obtulit (*b*)	διδοναι,παρεχειν (*c*)	ro ír (*a*)	tá sé tar éis a íobairt (*a*)	thairgse (*d*)
-IKNOS (*a*)	(patronymic suffix) son of (*b*)	fil. + gen. (*b*)	gen. (*b*)	mac + gen. (*b*)	(mac) + gen. (*b*)	mac + gen. (*b*)
TARVOS (*a*)	bull (*b*)	taurus (*a*)	ταυρος (*a*)	tarb (*a*)	tarbh (*a*)	tarbh (*a*)
TRI- (*a*)	three- (*a*)	tri- (*a*)	τρι- (*a*)	trí- (*a*)	trí- (*a*)	tri- (*a*)
GARANVS (*a*)	crane (*a*)	grus (*a*)	γερανος (*a*)	corr (*a*)	corr mhóna (*a*)	absent (*b*)
Tu*θ*os (*a*)	oven (*b*)	furnus (*c*)	ιπνος (*d*)	sorn (*e*)	sorn (*e*)	abhan (*b*)
LUXTODOS (*a*)	loaded (*a*)	oneratus (*b*)	γεμιστος (*c*)	lán (*a*)	lán (*a*)	lionta (*a*)
SUMMA UXSEDIA (*a*)	grand total (*b*)	summa summarum (*c*)	πας αριθμος (*d*)	Not determined	an t-iomlán (*e*)	cunntas (*f*)
ETI (*a*)	*thing* as well as *thing* (*b*)	item (*c*)	και (*d*)	ocus (*e*)	agus (*e*)	agus (*e*)
DUCI (*a*)	*person* and *person* (*b*)	et (*c*)	και (*d*)	ocus (*e*)	agus (*e*)	agus (*e*)
. . .DUCI . . . TONI. . . (*a*)	*person* (and), *p.* and *p.* (*b*)	*p.* et *p.* et *p.* (*c*)	*p.* και *p.* και *p.* (*d*)	*p.* ocus *p.* ocus *p.* (*e*)	*p.*, *p.* agus *p.* (*e*)	*p.* agus *p.* agus *p.* (*e*)
AVVOT, etc. (*a*)	has made (*b*)	fecit (*c*)	ποιειν, δραν (*d*)	do-rigni (*e*)	dhein sé, rinne sé (*e*)	rinn (*e*)
CINTUX (*a*)	first (*b*)	primus (*c*)	πρωτος (*c*)	cétnae (*a*)	céad− (*a*)	ceud (*a*)
ALLOS (*a*)	second (*b*)	secundus (*b*)	δευτερος (*c*)	tánaise/aile (*a*)	dara (*d*)	darna (*d*)
TR[ ] (*a*)	third (*a*)	tertius (*a*)	τριτος (*a*)	triss (*a*)	tríú (*a*)	treas (*a*)
PETUAR[ ] (*a*)	fourth (*b*)	quartus (*c*)	τεταρτος (*d*)	cethramad (*c*)	ceathrú (*c*)	ceithreamh (*c*)
PINPETOS (*a*)	fifth (*b*)	quintus (*c*)	πεμπτος (*a*)	cóiced (*c*)	cúigiú (*c*)	coigeamh (*c*)
SUEXOS (*a*)	sixth (*a*)	sextus (*a*)	εκτος (*a*)	seissed (*a*)	séú (*a*)	siathamh (*a*)
SEXTAMETOS (*a*)	seventh (*a*)	septimus (*a*)	εβδομος (*a*)	sechtmad (*a*)	seachtú (*a*)	seachdamh (*a*)
OXTUMETO[ ] (*a*)	eighth (*a*)	octauius (*a*)	ογδοος (*a*)	ochtmad (*a*)	ochtú (*a*)	ochdamh (*a*)
NAMET[ ] (*a*)	ninth (*a*)	nonius (*a*)	ενατος (*a*)	nómad (*a*)	naoú (*a*)	naoidheamh (*a*)
DECAMETOS (*a*)	tenth (*b*)	decimus (*a*)	δεκατος (*a*)	dechmad (*a*)	deichiú (*a*)	deicheamh (*a*)
M, MID (*a*)	month (*a*)	mensis (*a*)	μην (*a*)	mí (*a*)	mí (*a*)	mios (*a*)
LAT (*a*)	day (*b*)	dies (*b*)	ημερα (*c*)	laithe (*a*)	lá (*a*)	latha (*a*)
MATIR (*a*)	mother (*a*)	mater (*a*)	μητηρ (*a*)	máthair (*a*)	máthair (*a*)	màthair (*a*)
DUXTIR (*a*)	daughter (*a*)	filia (*b*)	θυγατηρ(*a*)	ingen (*c*)	iníon (*c*)	nighean (*c*)

uncertainty in the primary data would seriously affect a phylogenetic analysis, so we decided to minimize this risk by consulting bilingual Gaulish–Latin inscriptions.

Methods

Construction of the Indo-European Network. We used the linguistic network approach of Forster *et al.* (4). A phylogenetic network displays differences of items between language lists (or DNA molecules) with links and branches like a tree does, except that a network may contain reticulations when convergence (i.e., through historical loan events and chance parallel changes, or even through data misassignments incurred by the researcher) has obscured the evolutionary tree. The linguistic network approach is therefore expressly intended to search for treelike structure in potentially "messy" data. A technical problem that needs to be overcome is the multitude of irrelevant trees that a network may contain, in the form of reticulations, with even modestly noisy data. The first phase of the linguistic network approach is to remove characters (items) with more than a certain number of states (e.g., lexemes) across the translations according to an empirically determined threshold, because high variability is a sign either of inherent instability of that item or of unreconstructable ancient changes. The second phase is to process the binary characters (binary items) in the data to keep initial complexity to a minimum, with the rationale that a less variable, geographically widespread character state (e.g., a lexeme or phoneme) is more likely to reflect genetic relationships between languages. Characters that include states that are suffix losses are initially disregarded. In the third phase, the multistate characters are processed by splitting them into binary characters dictated by the current network: for each multistate character, that binary split which partitions the largest group of closely linked nodes or taxa (i.e., languages) is introduced first, following the same rationale as in phase 2. The other states of the multistate character are then split off the enlarged network. The fourth phase is to process the suffix losses. We found these to be least reliable for tree construction, because independent losses of a suffix frequently occur, causing convergence and thus reticulation. In all four phases, the processing of a certain character may contribute disproportionately to an increase in reticulations. If this occurs, the step should be reverted and other characters should be chosen iteratively for their ability to enlarge the network at low complexity. The temporarily excluded item is reintroduced in the next round of the phase. In the final phase, the uninformative binary items (i.e., those that differ in only one language) are added to the branch tips.

We compiled 35 Indo-European items with a view to avoiding ascertainment bias as explained in the supporting information, which is published on the PNAS web site, www.pnas.org, and listed these items in Table 1. Applying the phylogenetic procedure to Table 1, trials in phase one indicate that characters with more than five states (IEVRV, TUΘOS, SUMMA UXSEDIA, ETI, DUCI, DUCI/TONI, and AVVOT) contribute disproportionately to network complexity and are thus omitted. This reduces the item list to 29 less variable items. The subsequent phases are illustrated in Fig. 2. The first character to be processed is the binary syntactical item, which takes on the state VS in the Insular Celtic languages, and SV in all other languages. Next, the binary phonetic items -ps- and MATIR split English, Latin, and Greek from all others, and Welsh and Breton from all of the

Table 1. (continued)

Mod. Welsh	Mod. Breton	Mod. French	Mod. Occitan	Mod. Italian	Mod. Spanish	Mod. Basque
VS (*b*)	VS (*b*)	SV (*a*)	SV (*a*)	SV (*a*)	SV (*a*)	SX (*z*)
Absent (*b*)	Absent (*b*)	Absent (*b*)	Absent (*b*)	-o (*a*)	-o (*a*)	-a (*z*)
Absent (*d*)	Absent (*d*)	Absent (*d*)	Absent (*d*)	Absent (*d*)	Absent (*d*)	-(r)en (*z*)
Absent (*b*)	Absent (*b*)	Absent (*b*)	Absent (*b*)	Absent (*b*)	Absent (*b*)	-(r)entzat(ke) (*z*)
ICM (*c*)	ICM (*c*)	-e (*a*)	-a (*a*)	-a (*a*)	-a (*a*)	-a (*a*)
Absent (*d*)	Absent (*d*)	Absent (*d*)	Absent (*d*)	Absent (*d*)	Absent (*d*)	-(r)en (*z*)
ps rare (*a*)	ps rare (*a*)	ps rare (*a*)	ps rare (*a*)	ps rare (*a*)	ps rare (*a*)	ps rare (*a*)
i dduwian (*a*)	d'an Doueed (*a*)	aux dieux (*a*)	als dius (*a*)	agli die (*a*)	a los dioses (*a*)	jainkoei (*z*)
ac i ddynion (*a*)	ha d'an dud (*a*)	et aux hommes (*c*)	e als òmes (*c*)	ed agli uomini (*c*)	y a los hombres (*c*)	eta gizakiei (*z*)
mae e wedi cynnig (*e*)	deus kinniget (*e*)	a offert (*b*)	a ofèrt (*b*)	ha offerto (*b*)	ha ofrecido (*b*)	eman die (*z*)
ap/ab + name (*b*)	mab + name (*b*)	fils de + name (*b*)	filh de + name (*b*)	(*b*)	(*b*)	(*b*)
tarw (*a*)	tarv (*a*)	taureau (*a*)	taure/taur (*a*)	toro (*a*)	toro (*a*)	zezen (*z*)
tri- (*a*)	tri- (*a*)	tri- (*a*)	tri- (*a*)	tri- (*a*)	tri- (*a*)	hiru- (*z*)
garan, crychedd (*a*,*c*)	garan (*a*)	grue (*a*)	grua (*a*)	gru (*a*)	grulla (*a*)	kurrilo (*a*)
ffwrn (*c*)	forn (*c*)	four (*c*)	forn (*c*)	forno (*c*)	horno (*c*)	labea (*z*)
llawn (*a*)	karget (*d*)	chargé (*d*)	cargat (*d*)	carico (*d*)	cargado (*d*)	beteta (*z*)
cyfanswm (*g*)	absent (*h*)	total général (*i*)	en tot (*j*)	importo totale (*k*)	total (*j*)	guztira (*z*)
a/ac (*e*)	ha/hag (*e*)	et (*a*)	e (*a*)	e (*a*)	y (*f*)	eta (*a*)
a/ac (*e*)	ha/hag (*e*)	et (*c*)	e (*c*)	e (*c*)	y (*f*)	eta (*c*)
p., *p*. a/ac *p*. (*e*)	*p*., *p*. ha/hag *p*. (*e*)	*p*., *p*. et *p*. (*c*)	*p*., *p*. e *p*. (*c*)	*p*., *p* e *p*. (*c*)	*p*., *p*. y *p*. (*f*)	*p*., eta *p* eta *p*. (*c*)
mae e wedi gwneud (*f*)	deus graet (*g*)	a fait (*c*)	a fach (*c*)	ha fatto (*c*)	ha hecho (*h*)	egin du (*z*)
cyntaf (*a*)	kentan (*a*)	premier (*c*)	primièr (*c*)	primo (*c*)	primero (*c*)	lehenengoa (*z*)
ail (*a*)	eil (*a*)	seconde (*b*)	segond (*b*)	secondo (*b*)	segundo (*b*)	bigarrena (*z*)
trydydd (*a*)	trede (*a*)	troisième (*a*)	tresen (*a*)	terzo (*a*)	tercero (*a*)	hirugarrena (*z*)
pedwerydd (*a*)	pevare (*a*)	quatrième (*c*)	quatren (*c*)	quarto (*c*)	cuarto (*c*)	laugarrena (*z*)
pumed (*a*)	pempvet (*a*)	cinquième (*c*/*d*)	cinquen (*c*/*d*)	quinto (*c*)	quinto (*c*)	bostgarrena (*z*)
chweched (*a*)	c'hwec'hvet (*a*)	sixième (*a*)	seisen (*a*)	sesto (*a*)	sexto (*a*)	seigarrena (*a*)
seithfed (*a*)	seizhvet (*a*)	septième (*a*)	seten (*a*)	sèttimo (*a*)	septimo (*a*)	zazpigarrena (*z*)
wythfed (*a*)	eizhvet (*a*)	huitième (*a*)	ochen (*a*)	ottavo (*a*)	octavo (*a*)	zortzigarrena (*z*)
nawfed (*a*)	navet (*a*)	neuvième (*a*)	noven (*a*)	nono (*a*)	noveno (*a*)	bederatzigarrena (*z*)
degfed (*a*)	dekvet (*a*)	dixième (*a*)	desen (*a*)	dècimo (*a*)	décimo (*a*)	hamargarrena (*z*)
mis (*a*)	miz (*a*)	mois (*a*)	mes (*a*)	mese (*a*)	mes (*a*)	hilea (*z*)
dydd (*b*)	deiz (*b*)	jour (*d*)	jorn, (dia) (*d*,*b*)	giorno (*d*)	día (*b*)	eguna (*z*)
mam (*b*)	mamm (*b*)	mère (*a*)	maire (*a*)	madre (*a*)	madre (*a*)	ama (*b*)
merch (*d*)	merc'h (*d*)	fille (*b*)	filha (*b*)	figlia (*b*)	hija (*e*)	alaba (*z*)

Mod., modern; SV, subject–verb; SX, subject–any part of sentence; nom., nominative; gen., genitive; dat., dative; sg., singular; masc., masculine; fem., feminine; ICM, initial consonant mutation; (*p*), person. The letters in parentheses indicate character states, which are explained in the supporting information.

others, respectively. This completes the binary torso, which happens to be one-dimensional, in this case. Then the multistate character CINTUX is introduced, as it comes in only three states (a, b, and c). The nodes and languages with state c are split from the torso, after which the minority state b (found only in English) is split. The introduction of the next multistate character PINPETOS creates the first reticulations. The further splits are shown until LUXTODOS in Fig. 2, whereafter the suffix losses are split off, and finally the uninformative characters (those with only one deviant language, such as TEUO) are simply added to the branches, yielding the final network of Fig. 3.

Dating of Language Splits. Phylogeny-based time estimates using the average mutational distance (8–10), when applied to a tree of a biomolecule or word list, are only valid (*i*) if the molecules/word lists are related through treelike descent, and (*ii*) if an overall acceleration or deceleration of replacements across all branches has not occurred. Concerning *i*, we can trace a plausible tree through the network of Fig. 3 as indicated by the unbroken lines; concerning *ii*, only an "outgroup" (in genetics, an additional taxon definitely branching off before any of the observed branchings) would inform us of acceleration or deceleration, but a feasible outgroup is not yet clear for Indo-European. Note that uniformity in the retention rates of items is not required when average mutation rates are used (11), nor need all branches mutate at the same speed (12) as long as there are enough branches at an ancestral node to provide a reliable average, nor is mutational saturation a problem (encountered in pairwise glottochronological dating, when lexemes have been repeatedly replaced) if the tree reconstructs all recurrent changes. Dates and their errors are calculated by hand following Saillard *et al.* (8), or with the NETWORK 3.111 software available free at www.fluxus-engineering.com. Input data are coded item lists from a sample of languages: these can be extant languages, extinct languages, or any mixture of the two.

Results and Discussion

We have constructed a network (Fig. 3) of 13 ancient and modern Indo-European languages (Table 1) by using language items from ancient Gaulish–Latin inscriptions as explained in the supporting information. The phylogenetic network is largely treelike, indicating that the languages have exchanged few of these items in their prehistory. The consensus node of the six branches (Latin, Greek, English, Gaulish, Brythonic, and Goidelic) implies an Indo-European root for the sampled languages, possibly close to the hypothetical Proto-Indo-European language, thought to be at least 4,000 years old. The network method has correctly placed the ancient languages (Greek, Latin, Old Irish, and Gaulish) closer to the root of the tree than the modern languages. As a general point, it is interesting that the first five ordinal numbers (cintux, allos, tri, petuar, and pinpetos in Gaulish) are sufficient to subdivide the languages into known relationships (see table 7 in ref. 13), whereas the grammatical suffixes are less informative because they are frequently lost along independent branches of the tree.

The network strongly supports a Common Celtic branch, unambiguously distinguished from Proto-Indo-European by "cintux," "allos," and "xtonion." Within the Celtic branch, the

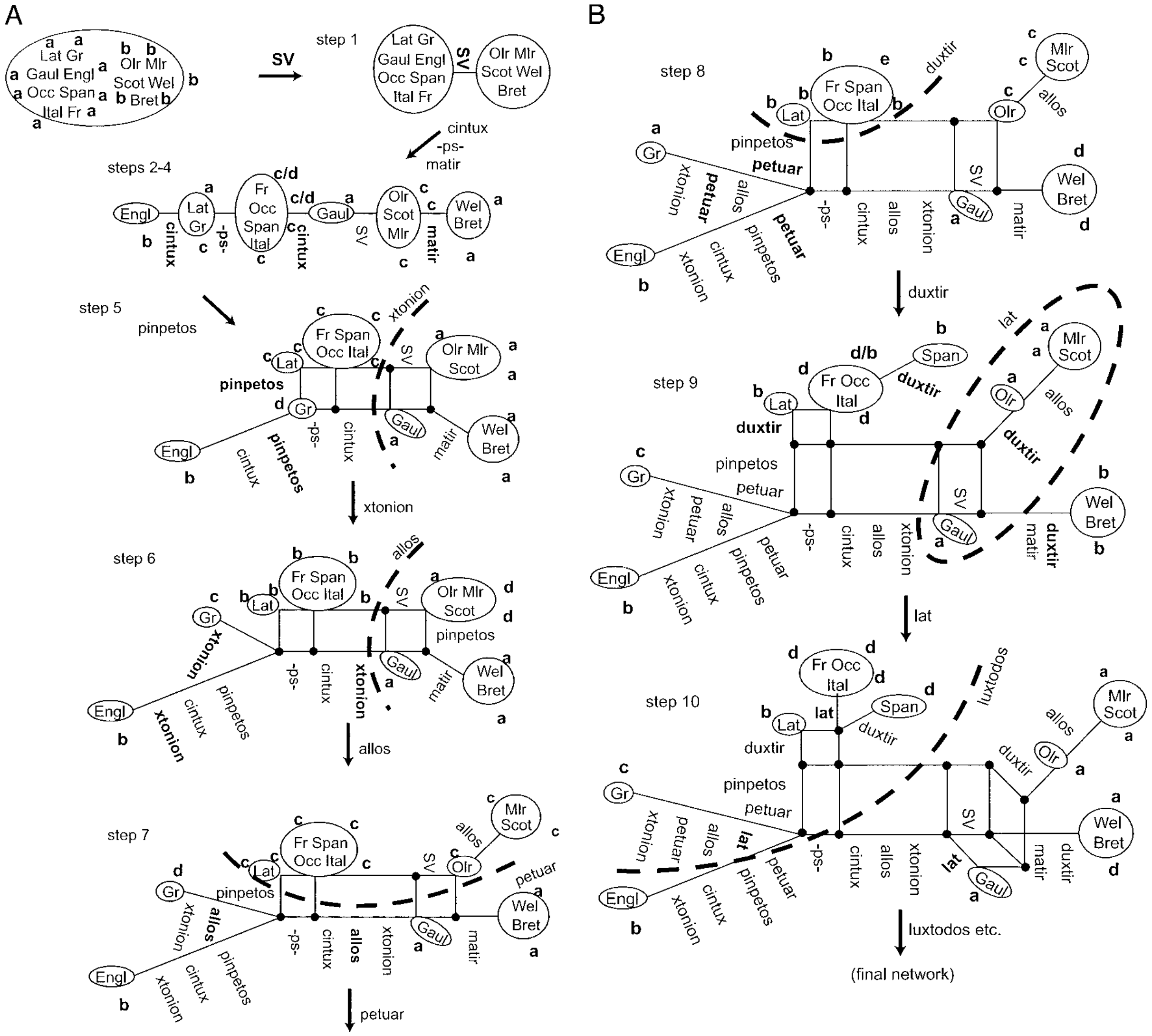

Fig. 2. Construction of the linguistic network. Thick broken lines indicate splits to be introduced in the following step. Character states *a–z* are taken from Table 1. Characters (e.g., SV, cintux) are entered perpendicularly to their links. Parallel links in a reticulation (here, a square or cube) signify the same character. The end result is shown in Fig. 3.

earliest split separates Gaulish (Continental Celtic) from the Insular Celtic languages, explaining Lambert's (7) paradoxical observation that "other [Gaulish] words are archaisms which can only be explained by calling upon the other Indo-European languages [. . .]: in this case, the other Celtic languages are of no help at all." The network method subsequently splits Insular Celtic into Brythonic (Welsh and Breton) and Goidelic (Irish and Scots Gaelic), in agreement with the traditional P/Q subclassification of Insular Celtic. The Celtic branching pattern evident in the network possibly reflects the prehistoric migration route of the ancient Celtic language: the split between Continental and Insular Celtic would then correspond to the arrival in the British Isles, and the split between Goidelic and Brythonic would correspond to their subsequent isolation in Ireland and Britain, respectively. Furthermore, the recent (circa 6th century) migrations of Irish to Scotland and of British to France are reflected in the short Scots Gaelic and Breton tips of the Celtic branches in Fig. 3.

Delving deeper into time, what can the network tell us about the often suspected existence of an ancestral Italo-Celtic branch (14, 15) within Indo-European? The network displays a multifurcation of Indo-European branches rather than a common Italo-Celtic branch. But it would be mistaken to conclude that the network disproves the hypothetical Italo-Celtic relationship. Because our item list is short, the network would be unlikely to distinguish brief periods of common ancestry. In other words, either Italo-Celtic never existed as a language, or it did exist but split into Italic and Celtic at a relatively early date.

A discussion of the Indo-European network would be incomplete without reference to its reticulations, expressing the "wave" aspect of nontreelike language evolution. For example, Celtic is unambiguously defined by "cintux," "allos," and "xtonion," while the "ps" loss (Graeco-Latin "parapsidi" becomes "paraxidi" in Gaulish inscriptions) is also shared by the modern Romance languages, which are otherwise quite distant. This character conflict is represented by the perpendicular reticulation in Fig. 3 and is thought (7) to indicate the survival in modern Romance languages of the Celtic tendency to eliminate "ps;" compare Latin "capsa" vs. Italian "cassa;" French "caisse;" and Spanish "caja."

Phylogenetic time estimates have not previously been attempted in linguistics to our knowledge, but are statistically

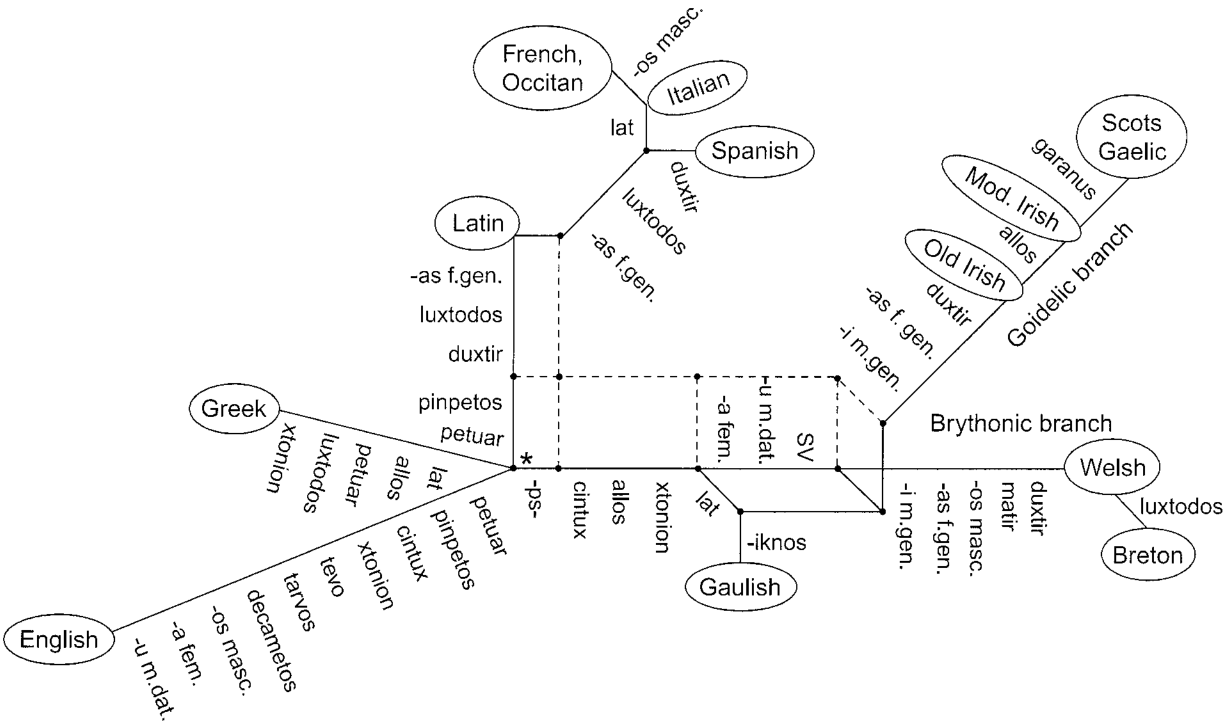

Fig. 3. Phylogenetic network of ancient and modern Indo-European languages. The items are noted in Gaulish, for translations consult Table 1. The asterisked node denotes the putative Indo-European rooting in the network. The network method produced the full network, from which the broken lines have been subsequently removed by hand. The remaining unbroken lines indicate the putative Indo-European tree that we chose for time estimations. As an illustration of how to read the tree, consider Spanish, which differs from Latin by the items ''duxtir'' (daughter), ''luxtodos'' (loaded), ''-as'' (genitive singular feminine suffix), and the phoneme ''ps.'' In Latin, the state for duxtir is ''filia,'' which according to the network has mutated to the state ''hija'' in Spanish, in agreement with etymological considerations. The genitive suffix has mutated from Latin to Spanish by being lost, likewise the phonetic item. The luxtodos item (''oneratus'' in Latin and ''cargado'' in Spanish) on the other hand has mutated by outright replacement.

feasible (8) once the language tree has been correctly reconstructed, by uncovering any recurrent changes of the items. For the time estimates we take only lexeme (word) changes and discard grammatical and phonological (pronunciation) items because the latter appear to be less reliable for reconstructing trees, as we observed above. Then, we can obtain a rough calibration from known language splits in Fig. 3: French/Occitan (0.5 lexemes spanning two branches of 1,000 years each), Latin/Romance (2 lexemes in 2,000 years), Old Irish/Modern Irish (1 lexeme in 1,000 years), Welsh/Breton (1.5 lexemes spanning two branches of 1,500 years each), Old Irish/Scots Gaelic (2 lexemes in 1,500 years), giving an average of ≈1 lexeme mutation in 1,350 years. The calibration takes into account that a pair consisting of two living languages encompasses two lines of descent from their ancestor, whereas a pair consisting of a dead and a descendent living language encompasses only one line of descent. The rooted lexeme tree of Fig. 4, which normalizes the six dead and living language branch lengths to AD 2000, yields a date for Indo-European fragmentation in Europe at 8100 BC ± 1,900 years. Note that the standard deviation of 1,900 years does not include uncertainty in the calibration, but it does express the uncertainty caused by mutation rate fluctuation (both in items and in languages), unlike "pairwise" glottochronology as advanced by Swadesh (16). For the fragmentation of Gaulish, Goidelic, and Brythonic from their most recent common ancestor, the lexeme tree yields a date of 3200 BC ± 1,500 years, but this date should be regarded as exploratory because it is based on only three estimators, i.e., three descendent branches. The date of 3200 BC ± 1,500 years would represent an oldest feasible estimate for the arrival of Celtic in the British Isles, and indeed is expected to be close to the actual date if the phylogenetic split between Gaulish and Insular Celtic was caused by the migration of the Celtic language to Britain and subsequent independent development in Britain.

How do previous approaches for reconstructing the Indo-European tree differ? Our approach combines three well known factors. First, we have chosen data that are unambiguously comparable across languages because we restrict ourselves to bilingual inscriptions. Second, our network method does not "force" a tree solution on the data if the data are not treelike, e.g., if the languages have evolved with wavelike spreads of loan

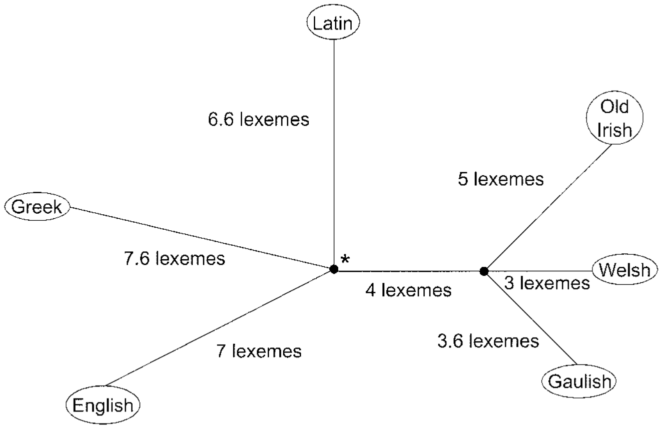

Fig. 4. Lexeme distance tree extracted from the Indo-European network of Fig. 3. The branch lengths are normalized to AD 2000 for the dead languages, assuming 1 lexeme exchange per 1,350 years on average.

features. Third, our method combines individual items and looks at their collective inheritance (akin to DNA sequence inheritance) rather than on their individual inheritance (akin to the chemical change of a single DNA nucleotide, or the etymological change of a word). It has long been recognized that any single item (e.g., the centum/satem criterion for subdividing Indo-European, or the P/Q criterion for classifying Celtic within Indo-European) can be unsatisfactory for language classification, and accordingly, historical linguists now secure more extensive item lists. These lists are not trivial to evaluate, and the phylogenetic approach we present here can assist in exploiting such combined information to the fullest extent.

Outlook

The present analysis excludes a number of interesting ancient Indo-European languages such as Hittite, Tocharian, etc. These omissions are an inevitable side effect of including the fragmentary corpus of ancient Gaulish: other ancient and fragmentary corpora would have little or no overlap with the Gaulish items, thus preventing any comprehensive phylogenetic analysis. To circumnavigate this difficulty and to arrive at a complete tree of ancient and modern Indo-European languages, future analyses may focus on the phylogenetic placement of a specific fragmentary language, as we have performed here for Gaulish, and may then piece together the resulting partial phylogenetic networks into a unified Indo-European language network. The unified network would yield improved age estimates for Indo-European, which in turn would assist in confirming or weakening the case for an early (possibly Neolithic) arrival and fragmentation of Proto-Indo-European in Europe (17) as suggested in this study.

We thank the following native speakers for compiling and/or correcting our language data (native language and where the individual learned that native language are given in parentheses): Mrs. Marion and Mr. John Angus McLeod (Scots Gaelic, Isle of Harris, Scotland), Dr. Mari Jones (southeast standard Welsh, Cwmbrân, Wales), Mr. Bernard Moullec (Breton, Guemene, Central Brittany, France), Dr. Maire Ní Mhaonaigh (Old and Modern Irish, Munster, Ireland), Dr. Izagirre Neskuts (Basque, Guernica, Spain), Mr. Miquèu Grosclaude (Occitan, Sauvelade, France), and Mrs. Claude Simion (Occitan, Colomiers, Haute-Garonne, France). We thank Michael, Iris, Andrew, Martin, and Paul Forster for valuable comments on an earlier version of the manuscript.

1. Jones, W. (1799) in *The Works of Sir William Jones* (Robinson and Evans, London), Vol. I, pp. 19–34.
2. Schleicher, A. (1863) *Die Darwinsche Theorie und die Sprachwissenschaft: Offenes Sendschreiben an Herrn Dr. Ernst Haeckel, o. Professor der Zoologie und Direktor des zoologischen Museums an der Universität Jena* (Böhlau, Weimar, Germany).
3. Schmidt, J. (1872) *Die Verwantschaftsverhältnisse der Indogermanischen Sprachen* (Böhlau, Weimar, Germany).
4. Forster, P., Toth, A. & Bandelt, H.-J. (1998) *J. Quant. Linguist.* **5,** 174–187.
5. Buchanan, G. (1582) *Rerum Scoticarum Historia*, revised and corrected (1722) (J. Bettenham, London), 2nd Ed.
6. Lhuyd, E. (1707) *Archaeologia Britannica* (Edward Lhuyd, Oxford); reprinted (1969) (Scholar Press, Menston, U.K.).
7. Lambert, P.-Y. (1994) *La Langue Gauloise* (Editions Errance, Paris).
8. Saillard, J., Forster, P., Lynnerup, N., Bandelt, H.-J. & Nørby, S. (2000) *Am. J. Hum. Genet.* **67,** 718–726.
9. Morral, N., Bertranpetit, J., Estivill, X., Nunes, V., Casals, T., Gimenez, J., Reis, A., Varon-Mateeva, R., Macek, M., Jr., Kalaydjieva, L., *et al.* (1994) *Nat. Genet.* **7,** 169–175.
10. Forster, P., Harding, R., Torroni, A. & Bandelt, H.-J. (1996) *Am. J. Hum. Genet.* **59,** 935–945.
11. Sigurðardóttir, S., Helgason, A., Gulcher, J. R., Stefansson, K. & Donnelly, P. (2000) *Am. J. Hum. Genet.* **66,** 1599–1609.
12. Bergsland, K. & Vogt, H. (1962) *Curr. Anthropol.* **3,** 115–153.
13. Greenberg, J. H. (1987) *Language in the Americas* (Stanford Univ. Press, Stanford, CA).
14. Russell, P. (1995) *An Introduction to the Celtic Languages* (Longman, Essex, U.K.).
15. Warnow, T. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 6585–6590.
16. Swadesh, M. (1955) *Int. J. Am. Linguist.* **21,** 121–137.
17. Renfrew, C. (1987) *Archaeology and Language* (Jonathan Cape, London).