Data in computational historical linguistics

Gerhard Jäger

December 2, 2016

Background

- comparative method strongly focuses on two types of data:
    - morphological paradigms
    - regular sound correspondences
- both are not very suitable for computational approaches, because
    - morphological categories are not easily comparable across languages, especially if we look individual language families
    - also, isolating languages have no morphology
    - identifying regular sound correspondences automatically is a surprisingly hard problem, due to data sparseness
    - currently one of the hot topics, far from resolved (List, 2014; Hruschka et al., 2015; Bouchard-Côté et al., 2013)

Background

- what we need (especially if we apply statistical methods):
    - data types which are applicable to all natural languages
    - ideally **lots** of data
- current practice:
    - word lists + expert annotations about cognacy (currently the dominant paradigm)
    - unannotated word lists in phonetic transcriptions
    - discrete grammatical categorizations (compiled by human experts)

Cognate-coded Swadesh lists

Swadesh lists

- collections of 100 – 200 concepts (there are different versions)
- *core vocabulary:*
  - not culture dependent
  - diachronically stable, i.e. resistant both against semantic change and aginst borrowing
- proposed by Morris Swadesh (Swadesh, 1955, 1971) to facilitate an early attempt to automatize certain tasks in historical linguistics
- popular among computational historical linguistics because *it is a standard*
- see (List, 2016) for a thoughtful discussion of the notion of cognacy

Cognates

- *Cognates* are words that have the same origin

  Latin *filius* ⇒ French *fils*, Italian *figlio*

- traditionally, cognacy excludes loanwords, but terminology among computationalists is sometimes less strict:

  Latin *persona* ⇒ English *person*

  would also qualify as cognate pair

- on average, the closer two languages are related, the more cognate pairs they share

Cognates

- during language change, the word for a given concept is sometimes replaced by a non-cognate one
- causes: semantic change, borrowing, morphological word formation
    - 'bone': Old High German *Bein* (cognate to Engl. *bone* ⇒ New High German *Knochen*
    - *Bein* is still part of the German lexicon, but it now means *leg*
- *cognate replacement* is comparable to a mutation in biological evolution

Cognates

Caveats

- cognacy is not binary, but a matter of degree
  - English *woman* ⇐ Old English *wiff-man*
  - first component is cognate to *wife*, German *Weib* etc., and second component to *man*, German *Mann* etc. Are *woman* and *Weib* cognate or not?
- for distantly related languages, experts often disagree about cognacy

Ancient Greek ὕλη/Latin silva 'woods'

IELex

- *Indo-European Lexical Cognacy Database*
- freely available online at http://ielex.mpi.nl/
- based on Dyen et al. (1992)
- current version curated by group at MPI Nijmegen
- recently migrated to new MPI Jena; new version not public yet

IELex

- 207-item Swadesh lists for 135 Indo-European languages
- words in orthographic and partially in phonetic transcription (IPA)
- entries are assigned to *cognate classes*
- sample entries:

language	*iso_code*	*gloss*	*global_id*	*local_id*	*transcription*	*cognate_class*
ELFDALIAN	qov	woman	962	woman	ˈkɛ̀lɪŋg	woman:Ag
DUTCH	nld	woman	962	woman	vrɑu	woman:B
GERMAN	deu	woman	962	woman	fraŭ	woman:B
DANISH	dan	woman	962	woman	ˈg̊ʰvenə	woman:D
DANISH_FJOLDE		woman	962	woman	kvinʲ	woman:D
GUTNISH_LAU		woman	962	woman	ˈkvɪnːˌfolk	woman:D
LATIN	lat	woman	962	woman	ˈmulier	woman:E
LATIN	lat	woman	962	woman	ˈfeːmina	woman:G
ENGLISH	eng	woman	962	woman	wʊmən	woman:H
GERMAN	deu	woman	962	woman	vaĭp	woman:H
DANISH	dan	woman	962	woman	ˈd̥ɛːmə	woman:K

Other publicly available cognacy data sources

- Austronesian Basic Vocabulary Database
  http://language.psy.auckland.ac.nz/austronesian/
- ten collections of cognate-coded Swadesh lists from various language families collected by Johann-Mattis List[1]
- ten collections of short (40-100 items) cognate-coded Swadesh lists from various language families collected by Søren Wichman and Eric Holman[2]
- 88 cognate-coded Swadesh lists from Central-Asian languages[3]

[1] List, J.-M. (2014): Data from: Sequence comparison in historical linguistics. GitHub Repository. http://github.com/SequenceComparison/SupplementaryMaterial. Release: 1.0.

[2] Supplementary material to Wichmann and Holman (2013)

[3] Supplementary material to Mennecier et al. (2016)

Phonetically transcribed Swadesh lists

The Automatic Similarity Judgment Program

- Project originally hosted at MPI EVA in Leipzig around Søren Wichmann
- since 2009; currently version 17 (2016)
- covers more than 7,000 languages and dialects (4.574 languages with iso code)
- basic vocabulary of 40 words for each language, in uniform phonetic transcription
- freely available at `http://asjp.clld.org/`

used concepts: *I, you, we, one, two, person, fish, dog, louse, tree, leaf, skin, blood, bone, horn, ear, eye, nose, tooth, tongue, knee, hand, breast, liver, drink, see, hear, die, come, sun, star, water, stone, fire, path, mountain, night, full, new, name*

The Automatic Similarity Judgment Program

Phonetic transcription

- 41 sound classes, all coded as ASCII characters
- various diacritics to capture finer phonetic distinctions, e.g.
  - `ph~`: aspirated `p`
  - `a*`: nasalized `a`
  - `hkw$`: pre-aspirated labalized `k`

Metadata

- language family, language genus, classifcation according to Ethnologue and Glottolog
- geographic location
- population size

The Automatic Similarity Judgment Program

ASJP sound classes (from Brown et al. 2013)

ASJP code symbol	*Description*	*IPA symbols*
p	voiceless bilabial stop and fricative	p,ɸ
b	voiced bilabial stop and fricative	b, β
f	voiceless labiodental fricative	f
v	voiced labiodental fricative	v
m	bilabial nasal	m
w	voiced bilabial-velar approximant	w
8	voiceless and voiced dental fricative	θ, ð
4	dental nasal	n̪
t	voiceless alveolar stop	t
d	voiced alveolar stop	d
s	voiceless alveolar fricative	s
z	voiced alveolar fricative	z
c	voiceless and voiced alveolar affricate	ts, ʤ
n	alveolar nasal	n
r	voiced apico-alveolar flap and all other varieties of "r-sounds"	ɾ, r, ʀ, ɽ
l	voiced alveolar lateral approximant	l
S	voiceless post-alveolar fricative	ʃ
Z	voiced post-alveolar fricative	ʒ
C	voiceless palato-alveolar affricate	ʧ
j	voiced palato-alveolar affricate	ʤ
T	voiceless and voiced palatal stop	c, ɟ
5	palatal nasal	ɲ
y	palatal approximant	j
k	voiceless velar stop	k
g	voiced velar stop	g
x	voiceless and voiced velar fricative	x,
N	velar nasal	ŋ

ASJP code symbol	*Description*	*IPA symbols*
q	voiceless uvular stop	q
G	voiced uvular stop	ɢ
X	voiceless and voiced uvular fricative, voiceless and voiced pharyngeal fricative	χ, ʁ, ħ, ʕ
h	voiceless and voiced glottal fricative	h, ɦ
7	voiceless glottal stop	ʔ
L	all other laterals	ʟ, ɭ, ʎ
!	all varieties of "click-sounds"	!, ǀ, ǁ, ǂ
i	high front vowel, rounded and unrounded	i, ɪ, y, ʏ
e	mid front vowel, rounded and unrounded	e, ø
E	low front vowel, rounded and unrounded	æ, ɛ, œ, ɶ
3	high and mid central vowel, rounded and unrounded	ɨ, ɘ, ə,ɜ, ʉ, ɵ, ɞ
a	low central vowel, unrounded	a, ɐ
u	high back vowel, rounded and unrounded	ɯ, u
o	mid and low back vowel, rounded and unrounded	ɣ, ʌ, ɑ, o, ɔ, ɒ

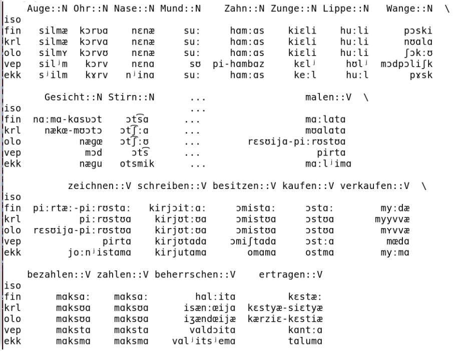
Automated Similarity Judgment Project

concept	Latin	English
I	ego	Ei
you	tu	yu
we	nos	wi
one	unus	w3n
two	duo	tu
person	persona, homo	%pers3n
fish	piskis	fiS
dog	kanis	dag
louse	pedikulus	laus
tree	arbor	tri
leaf	foly∼u*	lif
skin	kutis	%skin
blood	saNgw∼is	bl3d
bone	os	bon
horn	kornu	horn
ear	auris	ir
eye	okulus	Ei
nose	nasus	nos
tooth	dens	tu8
tongue	liNgw∼E	t3N

concept	Latin	English
knee	genu	ni
hand	manus	hEnd
breast	pektus, mama	brest
liver	yekur	liv3r
drink	bibere	drink
see	widere	si
hear	audire	hir
die	mori	dEi
come	wenire	k3m
sun	sol	s3n
star	stela	star
water	akw∼a	wat3r
stone	lapis	ston
fire	iNnis	fEir
path	viya	pE8
mountain	mons	%maunt3n
night	noks	nEit
full	plenus	ful
new	nowus	nu
name	nomen	nem

NorthEuraLex

- Massive data collection effort of the Tübingen EVOLAEMP project
- (currently) translations of 1,017 concepts into 103 (mostly) Northern Eurasian languages (cf. Dellert, 2015)
- everything transcribed in IPA
- (so far) no manual cognate coding



```
      Auge::N Ohr::N Nase::N Mund::N    Zahn::N Zunge::N Lippe::N   Wange::N  \
iso
fin   silmæ  kɔrʋɑ    nɛnæ     su:     hɑm:ɑs    kiɛli    hu:li      pɔski
krl   silmæ  kɔrvɑ    nɛnæ     su:     hɑm:ɑs    kiɛli    hu:li      nʊɑlɑ
olo   silmʏ  kɔrvʊ    nɛnæ     su:     hɑm:ɑs    kiɛli    hu:li      ʃɔk:ʊ
vep   silʲm   kɔrv    nɛnɑ      sʊ  pi-hɑmbɑz     kɛlʲ     hʊlʲ  mɔdpɔliʃk
ekk   sʲilm   kɤrv   nʲinɑ     su:     hɑm:ɑs     ke:l     hu:l       pɤsk

         Gesicht::N Stirn::N      ...                  malen::V  \
iso                               ...
fin   nɑ:mɑ-kɑsʋɔt     ɔt͡sɑ      ...                  mɑ:lɑtɑ
krl     nækœ-mʊɔtɔ    ɔt͡ʃ:ɑ      ...                  mʊɑlɑtɑ
olo           nægœ    ɔt͡ʃ:ʊ      ...        rɛsʊijɑ-pi:rʊstʊɑ
vep            mɔd      ɔt͡s      ...                    pirtɑ
ekk           nægu   otsmik       ...                 mɑ:lʲimɑ

           zeichnen::V schreiben::V besitzen::V kaufen::V verkaufen::V  \
iso
fin   pi:rtæ:-pi:rʊstɑ:   kirjɔit:ɑ:     ɔmistɑ:     ɔstɑ:         my:dæ
krl          pi:rʊstʊɑ    kirjʊt:ʊɑ     ɔmistʊɑ     ɔstʊɑ        myyvvæ
olo   rɛsʊijɑ-pi:rʊstʊɑ    kirjʊt:ʊɑ     ɔmistʊɑ     ɔstʊɑ         mʏvvæ
vep              pirtɑ    kirjʊtɑdɑ    ɔmiʃtɑdɑ     ɔst:ɑ          mœdɑ
ekk         jo:nʲistɑmɑ    kirjutɑmɑ       omɑmɑ     ostmɑ         my:mɑ

      bezahlen::V zahlen::V beherrschen::V    ertragen::V
iso
fin       mɑksɑ:    mɑksɑ:        hɑl:itɑ          kɛstæ:
krl       mɑksʊɑ    mɑksʊɑ        isæn:œijɑ  kɛstyæ-siɛtyæ
olo       mɑksʊɑ    mɑksʊɑ        iʒændœijæ  kærziɛ-kɛstiæ
vep       mɑkstɑ    mɑkstɑ        vɑldɔitɑ          kɑnt:ɑ
ekk       mɑksmɑ    mɑksmɑ      vɑlʲitsʲemɑ         tɑlumɑ
```

Grammatical classifications

Grammatical classification databases

- **World Atlas of Language Structure** (WALS) http://wals.info/
- **Syntactic Structures of the World's Languages** (SSWL) http://sswl.railsplayground.net/
- collection of syntactic parameters (in the Chomskyan sense) for a few dozen languages collected in the LanGeLin project (Giuseppe Longobardi)

Expert family trees

Expert family trees

- Ethnologue https://www.ethnologue.com/
- Glottolog http://glottolog.org/
  - in many ways improved version of Ethnologue
  - strives to apply uniform standards across all languages
  - rather conservative in accepting family status

Running example

Running example

- 25 living Indo-European languages
- three types of data
  - Swadesh lists in IPA transcription, taken from IELex
  - expert cognate classifications of Swadesh list entries (likewise taken from IELex),[4] and
  - phonological, grammatical and semantic classifications of languages (taken from WALS)

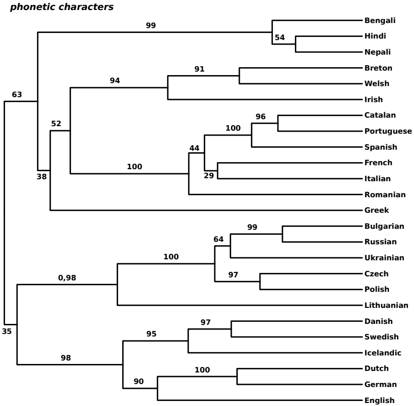
[4] I only included those entries from IELex where both an IPA transcription and a cognate classification is given.

Running example

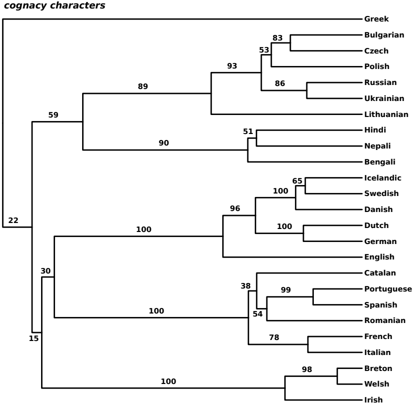
sample entries:

language	*phonological form* (IELex)	*cognate class* (IELex)	*order of subject, object and verb* (WALS)
Bengali	-	-	SOV
Breton	-	-	SVO
Bulgarian	muˈrɛ	sea:B	SVO
Catalan	mar; maɾ; ma	sea:B	SVO
Czech	ˈmɔr̝ɛ	sea:B	SVO
Danish	hɑw/søˀ	sea:K/sea:J	SVO
Dutch	ze	sea:J	no dominant order
English	si:	sea:J	SVO
French	mɛʀ	sea:B	SVO
German	ze:/ʼo:t͜sea:n/me:ɐ̯	sea:J/sea:E/sea:B	no dominant order
Greek	ˈθalaˌsa	sea:F	no dominant order
Hindi	-	-	SOV
Icelandic	haːv/sjouːr	sea:K/sea:J	SVO
Irish	ˈfˠæɾˠɟɪ	sea:G	VSO
Italian	ˈmare	sea:B	SVO
Lithuanian	ˈjuːrɐ	sea:H	SVO
Nepali	-	-	SOV
Polish	ˈmɔʐɛ	sea:B	SVO
Portuguese	maɾ	sea:B	SVO
Romanian	ˈmare	sea:B	SVO
Russian	ˈmɔrʲɛ	sea:B	SVO
Spanish	maɾ	sea:B	SVO
Swedish	hɑːv/ɧøː	sea:K/sea:J	SVO
Ukrainian	ˈmɔrɛ	sea:B	SVO
Welsh	-	-	VSO

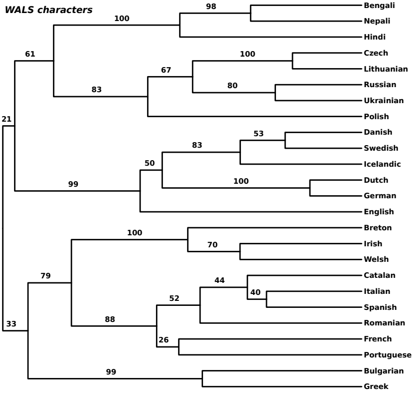
Automatic phylogenetic inference



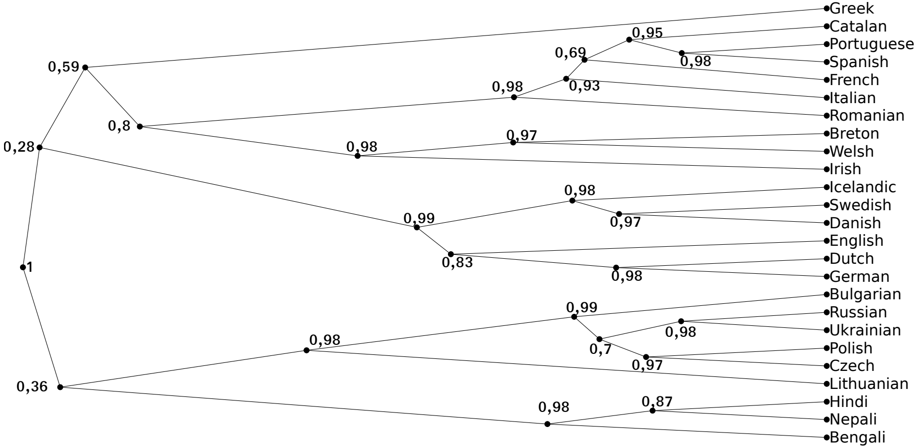
Automatic phylogenetic inference



Automatic phylogenetic inference



Automatic phylogenetic inference



Bouchard-Côté, A., D. Hall, T. L. Griffiths, and D. Klein (2013). Automated reconstruction of ancient languages using probabilistic models of sound change. *Proceedings of the National Academy of Sciences*, **36**(2):141–150.

Brown, C. H., E. Holman, and S. Wichmann (2013). Sound correspondences in the world's languages. *Language*, **89**(1):4–29.

Dellert, J. (2015). Compiling the Uralic dataset for NorthEuraLex, a lexicostatistical database of Northern Eurasia. Proceedings of the First International Workshop on Computational Linguistics for Uralic Languages. January 16, Tromsø, Norway.

Dyen, I., J. B. Kruskal, and P. Black (1992). An Indoeuropean classification: A lexicostatistical experiment. *Transactions of the American Philosophical Society*, **82**(5):1–132.

Hruschka, D. J., S. Branford, E. D. Smitch, J. Wilkins, A. Meade, M. Pagel, and T. Bhattachary (2015). Detecting regular sound changes in linguistics as events of concerted evolution. *Current Biology*, **25**(1):1–9.

List, J.-M. (2014). *Sequence Comparison in Historical Linguistics*. Düsseldorf University Press, Düsseldorf.

List, J.-M. (2016). Beyond cognacy: historical relations between words and their implication for phylogenetic reconstruction. *Journal of Language Evolution*, **1**(1):119–136. Doi: 10.1093/jole/lzw006.

Mennecier, P., J. Nerbonne, E. Heyer, and F. Manni (2016). A Central Asian language survey: Collecting data, measuring relatedness and detecting loans. *Language Dynamics and Change*, **6**(1). In press.

Swadesh, M. (1955). Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics*, **21**:121–137.

Swadesh, M. (1971). *The Origin and Diversification of Language*. Aldine, Chicago.

Wichmann, S. and E. W. Holman (2013). Languages with longer words have more lexical change. In L. Borin and A. Saxena, eds., *Approaches to Measuring Linguistic Differences*, pp. 249–284. Mouton de Gruyter, Berlin.