Algorithms for Language Reconstruction

Kondrak's 2002 thesis

Armin W. Buch

February 8, 2013

How to reconstruct a proto-language?

- Identification of cognates
- Alignment of cognates
- Discovery of sound correspondences
- Reconstruction of proto-forms
- Kondrak contributes unsupervised algorithms for the first three tasks

Alignment

- Alignment is usually calculated with a dynamic programming algorithm (Wagner-Fischer)
- It needs a distance metric

1.	$\forall a, b : d(a, b) \geq 0$	*nonnegative property*
2.	$\forall a, b : d(a, b) = 0 \Leftrightarrow a = b$	*zero property*
3.	$\forall a, b : d(a, b) = d(b, a)$	*symmetry*
4.	$\forall a, b, c : d(a, b) + d(b, c) \geq d(a, c)$	*triangle inequality*

Table 4.2: The metric axioms.

- Kondrak adapts extensions to the algorithm to phonetic data

Similarity vs. distance

- To a large extent, similarity measures and distance metrics can be exchanged
- The metric properties do not always make sense for phoneme distance (we will see examples)
- Linguistic intuitions are sometimes easier to express as similarities
- The alignment algorithm is easily adapted to similarities
  - Assign similarity scores instead of costs
  - Choose the maximum, not the minimum

Local alignment

- Let the usual alignment be called *global*
- *Local* alignment strips off prefixes and suffixes
- by having no indel costs at the beginning and at the end of words
- instead, it maximizes the similarity of similar substrings (possibly the root)

	‖	ā	p	a	k	o	‖	sīs
w	‖	ā	p	i	k	o	‖	nōha

Table 4.10: An example of local alignment.

Half-local alignment

- Words tend to change a lot at their right edge, while the left edge is quite stable
- *Half-local* alignment aligns globally on the left, and locally on the right

‖	-	ā	p	a	k	o	‖	sīs
‖	w	ā	p	i	k	o	‖	nōha

Table 4.13: An example of half-local alignment.

Gap penalties

- Gaps can be longer than just one segment
- e.g. by loss of an entire syllable
- In order to weigh this less than a series of deletions, gap costs can be calculated with a linear function
- initial gap cost + segment cost * number of deleted segments

Compression and expansion

- Many-to-one and one-to-many relations can be modeled as substitution plus deletion/insertion
- but this is not linguistically adequate
- and its cost/similarity would be judged differently
- As an example, consider En. 'fact' vs. Sp. 'hecho'

f	a	k	t
-	e	č	-

f	a	k	t
-	e	-	č

f	a	kt
-	e	č

Table 4.15: An example of cognate alignment that requires the operation of compression/expansion.

Transposition

- In phonology, transposition is rare
- Span. *cocodrilo*
- The most common instance is metathesis of adjacent segments
- Metathesis is highly irregular
- For practical purposes, it will be ignored here

Phoneme similarity

- The easiest measure of phoneme distance is identity

	a	i	y	n	p	r	s
a	0	1	1	1	1	1	1
i	1	0	1	1	1	1	1
y	1	1	0	1	1	1	1
n	1	1	1	0	1	1	1
p	1	1	1	1	0	1	1
r	1	1	1	1	1	0	1
s	1	1	1	1	1	1	0

Table 4.17: An elementary cost function.

Covington's measure

- Covington (1996) defines a phonetic distance measure
- gap penalty equals 10 base costs + 40 per segment

Penalty	Conditions
0	Exact match of consonants or glides (w, y)
5	Exact match of vowels (reflecting the fact that the aligner should prefer to match consonants rather than vowels if it must choose between the two)
10	Match of two vowels that differ only in length, or i and y, or u and w
30	Match of two dissimilar vowels
60	Match of two dissimilar consonants
100	Match of two segments with no discernible similarity
40	Skip preceded[2] by another skip in the same word (reflecting the fact that affixes tend to be contiguous)
50	Skip not preceded by another skip in the same word

Table 4.18: Covington's [1996] "evaluation metric".

Phoneme similarity 2

- Covington's measure has a low resolution

	a	i	y	n	p	r	s
a	5	30	100	100	100	100	100
i	30	5	10	100	100	100	100
y	100	10	0	60	60	60	60
n	100	100	60	0	60	60	60
p	100	100	60	60	0	60	60
r	100	100	60	60	60	0	60
s	100	100	60	60	60	60	0

Table 4.19: A partial distance matrix for Covington's distance function.

Covington's measure 2

- it is not a metric
  - zero property violated with a:i
  - Preference for matching identical C over matching id. V cannot be expressed in a metric
  - triangle inequality violated with a:i:y
  - cf. labio-velars (double marked, close to both); also cf. j/dʃ
- "just a stand-in for a more sophisticated, perhaps feature-based, system"
- Kondrak reports a good correlation between these trial-and-error costs and feature based Hamming distance, when the latter is an average over all sounds in the category

feature name	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	r	s	t	u	v	w	x	y	z
[tense]	+	−	−	−	+	−	−	−	+	−	−	−	−	−	+	−	−	−	−	+	−	+	−	+	−
[spread glottis]	−	−	−	−	−	−	−	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
[voice]	+	+	−	+	+	−	+	−	+	+	−	+	+	+	+	−	+	−	−	+	+	+	−	+	+
[back]	+	−	−	−	−	−	+	+	−	−	+	−	−	−	+	−	−	−	−	+	−	+	+	−	−
[coronal]	−	−	+	+	−	−	−	−	−	+	−	+	−	+	−	−	+	+	+	−	−	−	−	−	+
[continuant]	+	−	−	−	+	+	−	+	+	−	−	−	−	−	+	−	+	+	−	+	+	+	+	+	+
[high]	−	−	+	−	−	−	+	−	+	+	+	−	−	−	−	−	−	−	−	+	−	+	+	+	−
[strident]	−	−	+	−	−	+	−	−	−	+	−	−	−	−	−	−	−	+	−	−	+	−	−	−	+
[round]	−	−	−	−	−	−	−	−	−	−	−	−	−	−	+	−	−	−	−	+	−	+	−	−	−
[syllabic]	+	−	−	−	+	−	−	−	+	−	−	−	−	−	+	−	−	−	−	+	−	−	−	−	−
[obstruent]	−	+	+	+	−	+	+	+	−	+	+	−	−	−	−	+	−	+	+	−	+	−	+	−	+
[nasal]	−	−	−	−	−	−	−	−	−	−	−	−	+	+	−	−	−	−	−	−	−	−	−	−	−
[consonantal]	−	+	+	+	−	+	+	+	−	+	+	+	+	+	−	+	+	+	+	−	+	−	+	−	+
[low]	+	−	−	−	−	−	−	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
[anterior]	−	+	+	+	−	+	−	−	−	+	−	+	+	+	−	+	+	+	+	−	+	−	−	−	+
[distributed]	+	+	+	+	+	−	+	+	+	−	+	−	+	−	+	+	−	+	+	+	−	+	+	+	+
[delayed release]	−	−	+	−	−	−	−	−	−	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−

Table 4.20: Feature vectors adopted from Hartman [1981].

Phoneme similarity 3

	a	i	y	n	p	r	s
a	0	3	4	10	9	8	10
i	3	0	1	9	8	7	9
y	4	1	0	8	7	6	8
n	10	9	8	0	5	2	6
p	9	8	7	5	0	5	3
r	8	7	6	2	5	0	4
s	10	9	8	6	3	4	0

Table 4.21: A partial distance matrix based on binary features.

Problems with binary features

- Binary features are interpreted within a language
- they do not always reflect confusability / possible historical change:
- /j/ $\rightarrow$ /dʃ/ is likely, but the two are very dissimilar

Multi-valued features

- e.g. with values within [0,1]
- possibly also weighted features (place > manner of articulation)
- efforts at the time (Nerbonne & Heringa 1997) found worse alignments with better weightings
- still, beneficial weightings might be derived automatically
- possibly today with more hand-annotated cognate data

Feature name	Phonological term	Numerical value
Place	[bilabial]	1.0
	[labiodental]	0.95
	[dental]	0.9
	[alveolar]	0.85
	[retroflex]	0.8
	[palato-alveolar]	0.75
	[palatal]	0.7
	[velar]	0.6
	[uvular]	0.5
	[pharyngeal]	0.3
	[glottal]	0.1
Manner	[stop]	1.0
	[affricate]	0.9
	[fricative]	0.8
	[approximant]	0.6
	[high vowel]	0.4
	[mid vowel]	0.2
	[low vowel]	0.0
High	[high]	1.0

Syllabic	5	Place	40
Voice	10	Nasal	10
Lateral	10	Aspirated	5
High	5	Back	5
Manner	50	Retroflex	10
Long	1	Round	5

Table 4.27: Features used in ALINE and their salience settings.

Phoneme similarity 4

	a	i	y	n	p	r	s
a	15	8	2	−50	−56	−28	−40
i	8	15	10	−26	−32	−4	−16
y	2	10	15	−21	−27	1	−11
n	−50	−26	−21	35	9	−7	5
p	−56	−32	−27	9	35	−13	19
r	−28	−4	1	−7	−13	35	3
s	−40	−16	−11	5	19	3	35

Table 4.29: A partial similarity matrix based on multivalued features with diversified salience values.

Kondrak's ALINE algorithm

- similarities, not distances
- best alignments within a threshold ϵ
- local alignments; this replaces gap functions
- indels, substitution, expansion, compression
- transpositions are rare and too irregular
- multivalued features

```
algorithm Alignment
input: phonetic strings x and y
output: alignment of x and y
define S(i, j) = −∞ when i < 0orj < 0

for i := 0 to |x| do
    S(i, 0) := 0
for j := 0 to |y| do
    S(0, j) := 0
for i := 1 to |x| do
    for j := 1 to |y| do
        S(i, j) := max(
            S(i − 1, j) + σ_skip(x_i),
            S(i, j − 1) + σ_skip(y_j),
            S(i − 1, j − 1) + σ_sub(x_i, y_j),
            S(i − 1, j − 2) + σ_exp(x_i, y_{j−1}y_j),
            S(i − 2, j − 1) + σ_exp(x_{i−1}x_i, y_j),
            0)

T := (1 − ϵ) · max_{i,j} S(i, j)

for i ← 1 to |x| do
    for j ← 1 to |y| do
        if S(i, j) > T then
            Retrieve(i, j, 0)
```

$$\sigma_{skip}(p) = C_{skip}$$

$$\sigma_{sub}(p,q) = C_{sub} - \delta(p,q) - V(p) - V(q)$$

$$\sigma_{exp}(p,q_1q_2) = C_{exp} - \delta(p,q_1) - \delta(p,q_2) - V(p) - max(V(q_1),V(q_2))$$

where

$$V(p) = \begin{cases} 0 & \text{if } p \text{ is a consonant} \\ C_{vwl} & \text{otherwise} \end{cases}$$

$$\delta(p,q) = \sum_{f \in R} \text{diff}(p,q,f) \times \text{salience}(f)$$

where

$$R = \begin{cases} R_C & \text{if } p \text{ or } q \text{ is a consonant} \\ R_V & \text{otherwise} \end{cases}$$

Table 4.26: Scoring functions.

Annotations to ALINE

- diff(p,q,f) returns the difference between p and q for feature f
- Vowel features: syllabic, nasal, retroflex, high, back, round, long
- Consonant features: syllabic, manner, voice, nasal, retroflex, lateral, aspirated, place & double (= secondary place)
- Double leads to violation of triangle inequality, because the closest is taken

Evaluation

- 82 words (from Covington 1996), manually coded for cognacy
- Spanish–French, English–German, English–Latin, Fox–Menomini, and some solitary examples
- This was the best data available
- And still it may contain errors, and it has too many too easy pairs
- Furthermore, it's used for development and for evaluation
- ALINE outperforms Covington's method, but still has errors

	Covington's alignments	*ALINE's alignments*
three:trēs	θ r i y	‖ θ r iy ‖
	t r ē s	‖ t r ē ‖ s
blow:flāre	b l - - o w	‖ b l o ‖ w
	f l ā r e -	‖ f l ā ‖ re
full:plēnus	f - - - u l	‖ f u l ‖
	p l ē n u s	‖ p - l ‖ ēnus
fish:piscis	f - - - i š	‖ f i š ‖
	p i s k i s	‖ p i s ‖ kis
I:ego	- - a y	‖ ay ‖
	e g o -	‖ e ‖ go
tooth:dentis	- - - t u w θ	‖ t uw θ ‖
	d e n t i - s	den ‖ t i s ‖

Table 4.33: Examples of alignments of English and Latin cognates.

Armin Buch

Results

- ALINE achieves 95% accuracy compared to Kondrak's manual alignments
- it outperforms earlier approaches
- 'tooth' cannot be correctly aligned without referring to regular sound changes

‖	t	uw	-	θ	‖	
‖	d	e	n	t	‖	is

Table 4.34: The correct alignment of *tooth:dentis*.

Not everything is a cognate

Spanish	English	Classification
sal	*salt*	genetic cognates
suéter	*sweater*	direct borrowing
ambición	*ambition*	borrowing from a third language
mucho	*much*	chance similarity
carpeta 'folder'	*carpet*	"false friends"
cuclillo	*cuckoo*	onomatopoeic words
mamá	*mommy*	nursery words

Table 5.1: Examples of similar words in Spanish and English.

Cognate: a working definition

- For the present purposes, everything with similar meaning *and* form is a cognate
- Useful for unsupervised methods, including Greenberg's mass lexical comparison
- Better, and still to be established: automatically finding sound correspondences, and defining cognates accordingly
- Best data available on a large scale: transcribed word lists with glosses

Example word list 1

āniskōhōčikan	string of beads tied end to end
asikan	sock, stocking
kamāmakos	butterfly
kostāčīwin	terror, fear
misiyēw	large partridge, hen, fowl
namēhpin	wild ginger
napakihtak	board
tēhtēw	green toad
wayakēskw	bark

Table 5.2: An excerpt from a Cree vocabulary list [Hewson, 1999].

Example word list 2

āšikan	dock, bridge
anaka'ēkkw	bark
kipaskosikan	medicine to induce clotting
kottāčīwin	fear, alarm
mēmīkwan'	butterfly
misissē	turkey
namēpin	sucker
napakissakw	plank
tēntē	very big toad

Table 5.3: An excerpt from an Ojibwa vocabulary list [Hewson, 1999].

Kondrak's program COGIT

- An algorithm to identify cognates
- It needs to evaluate phonetic similarity (via ALINE) and semantic similarity
- Phonetic similarity is normalized by dividing by the self-similarity of the more self-similar word[1]
- Semantic similarity via WordNet
- Identity of glosses is in general not enough

[1] As I understand it

Problems in establishing semantic similarity

- Spelling errors / variants
- Inflection
- Modifiers: determiners, adjectives, compounds, complements, adjuncts
- synonymy (‘tomb’, ‘grave’)
- Semantic changes (‘fowl’, ‘turkey’; ‘broth’, ‘grease’)

Addressing these problems

- Spelling correction (even if manually)
- Removal of stop words (‘a kind of’, . . .)
- Extraction of keywords (syntactic heads heuristically found after POS-tagging)
- Lemmatization
- Employing WordNet

WordNet relations

Type	Name	Example	Inverse of
hypernymy	IS-A	*bird* → *animal*	hyponymy
hyponymy	SUBSUMES	*bird* → *robin*	hypernymy
meronymy	PART-OF	*beak* → *bird*	holonymy
holonymy	HAS-A	*tree* → *branch*	meronymy
antonymy	COMPLEMENT-OF	*leader* ↔ *follower*	*itself*

Table 5.4: The main lexical relations between nouns in WordNet.

Semantic shift

- generalization & specialization (‘deer’, ‘Tier’)
- melioration (Ancient Greek ‘guna’ “woman”, ‘queen’)
- pejoration (‘Frau’; ‘Weib’)
- metaphor (‘star’)
- metonymy (attribute for whole): ‘crown’
- synechdoche (pars pro toto)

⇒ some of them happen along WordNet’s semantic relations

Weighing semantic similarity

Rank	Similarity level	Score
1	gloss identity	1.00
2	gloss synonymy	0.70
3	keyword identity	0.50
4	gloss hypernymy	0.50
5	keyword synonymy	0.35
6	keyword hypernymy	0.25
7	gloss meronymy	0.10
8	keyword meronymy	0.05
9	none detected	0.00

Table 5.8: Semantic similarity levels.

- WordNet paths longer than 1 are considered useless

Example calculation

- COGIT's similarity score is a weighted sum of the phonetic and semantic similarity
- The weight is empirically set to 80% phonology, 20% semantics
- if it exceeds a threshold, record the pair as a cognate candidate
- Example: Cree *wahkwa* 'a lump of roe', Ojibwa *wakk* 'fish eggs'
  - remove determiner
  - identify keywords (lump, roe; fish, eggs)
  - lemmatize (egg)
  - hypernymy (roe IS-A egg) beats meronymy (roe PART-OF fish): 0.25
  - phonetic score 0.4167
  - overall score 0.3834

Evaluation

- evaluated on a set of dictionaries of North American languages, with its own inconsistencies
- weighting experimentally set to 80–20, so semantics isn't a strong indicator
- no threshold set: it is a trade-off between recall and precision
- precision levels reported as an average over 0%, 10%, . . . 100% recall thresholds
- better than older methods

The role of semantics

- gloss identity holds for 62.7% of all cognates (no special method needed at all)
- keyword identity holds for 12%
- others insignificant
- 19.3% are not connected via their glosses at all (by this method)
- No word sense disambiguation in the process → false positives via WordNet
- imperfect keyword extraction
- missing entries in WordNet

Identity vs. correspondence

- English ‘have’ is not cognate with Latin ‘habere’, but with ‘capire’
- by regular sound changes (Grimm’s Law, . . .)
- Is automatic identification of correspondences possible?
- Is it possible on data un-annotated for actual cognacy?
- That is, are correspondences stable enough to be visible under noise?

English	Latin		English	Latin	
t ε n	**d** e k e	'ten'	**t** ū	**d** u o	'two'
ī **t**	e **d**	'eat'	**t** ū **θ**	**d** e n **t**	'tooth'
n ε s **t**	**n** i **d**	'nest'	**n** ī	g e **n**	'knee'
n ε **f** j ū	**n** e **p** o t	'nephew'	**f** u **t**	**p** e **d**	'foot'
f ō **m**	s **p** u **m**	'foam'	w u **l** **f**	**l** u **p**	'wolf'
θ **r** ī	**t** **r** e	'three'	**r** ū **t**	**r** a **d** i k	'root'
s ɪ **t**	**s** e **d**	'sit'	**h** a **r** **t**	**k** o **r** **d**	'heart'
h ɔ **r** **n**	**k** o **r** **n**	'horn'	**b** **r** ə **ð** ə **r**	**f** **r** a **t** **r**	'brother'

Table 6.1: Examples of English–Latin cognates exhibiting correspondences.

Phoneme vs. word alignment

- Segment alignment is well-known from syntax
- Kondrak relies on Melamed's (2000) algorithm
- first, initialize correspondence likelihoods using co-occurrence counts (G^2 statistics, which I will not try to explain here)
- greedily link words 1-to-1, highest scores first
- re-estimate likelihoods and repeat (serves to prune accidental or indirect co-occurrences)
- extended for contiguous sequences being treated as one segment (many-to-one, one-to-many, many-to-many)

Kondrak's CORDI algorithm

- no crossing links expected, so the greedy aligner is replaced with a variant of the standard aligner
- half-local (don't consider word endings)
- threshold on links: Don't match everything even if you could
- negative weight on indels
- positive weight on each link

Evaluation 1

- 112 English-Latin cognate pairs
- Now, tooth:dent can be aligned correctly
- y:w is claimed to result from the diphtong [ay]

	cooc	*links*	*score*	valid
r:r	28	28	193.1	yes
n:n	23	23	158.6	yes
l:l	20	20	138.0	yes
s:s	17	17	117.3	yes
m:m	15	15	103.5	yes
f:p	13	13	89.7	yes†
t:d	11	11	75.9	yes†
k:g	8	8	55.1	yes†
y:w	6	6	41.4	no
b:f	6	6	41.4	yes†
h:k	5	5	34.5	yes†
θ:t	4	4	27.6	yes†

Table 6.2: English–Latin correspondences discovered by Method D in pure cognate data. The correspondences marked with a † are predicted by Grimm's Law.

Noise

- pure cognate data is hard to get
- 200 words (English/Latin), out of which only 29% are cognates
- highly robust

	cooc	*links*	*score*	valid
r:r	26	24	158.7	yes
n:n	24	23	154.2	yes
t:d	18	18	122.4	yes
k:k	12	11	72.5	yes
s:s	11	10	65.7	yes
f:p	9	9	61.2	yes
m:m	10	9	58.9	yes
d:t	10	8	49.8	no
l:l	14	9	49.7	yes
h:k	7	7	47.6	yes

Table 6.4: English–Latin correspondences discovered by CORDI in noisy synonym data.

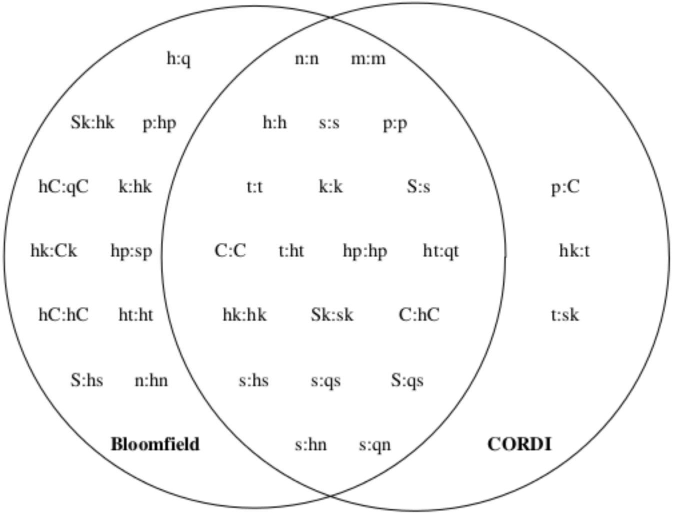

Figure 6.2: The Fox–Menomini consonantal correspondences determined by a linguist

Outlook

- A phoneme-by-phoneme correspondence likelihood table derived from actual (cognate) data wasn't available at the time
- Automatic reconstruction of proto-forms is still a hot topic