Estimating the proximity between languages by their commonality in vocabulary structures

Yves Lepage[1], Julien Gosme[2], and Adrien Lardilleux[2]

[1]Waseda university, IPS Graduate school, Japan
[2]GREYC, University of Caen Basse-Normandie, France
Firstname.Lastname@{aoni.waseda.jp,info.unicaen.fr}
http://www.waseda.jp/ips/english/kyoin/01_12.html

Abstract. This article proposes a possible way of measuring proximity between languages: it consists in measuring the commonality of structures between the vocabularies of two languages. Experiments conducted on a multilingual lexicon of nine European languages acquired from the *Acquis communautaire* confirmed usual knowledge on the closeness or remoteness of these languages.

Key words: Proximity between languages, analogy between words, morphology.

1 Introduction

This article deals with the problem of closeness between languages. Since the Renaissance, a number of observations have been made that relate Latin to vernacular languages like Italian (Tolomei, Castelvetro, both sixteenth century). In the eighteenth century, Sanskrit has been recognized by several philologists as being related to other European languages and Old-Persian (van Boxhorn 1647, Cœurdoux 1760, Jones 1786). The idea of a common origin of all those languages led to the study of the phonetic laws that explain sound differences between present languages (Grimm 1822, Bopp 1833, Verner 1875), and to the manual reconstruction of a hypothetical Indo-European language (Schleicher 1868). All these works interpret closeness between languages as the clue for a historical relation between languages in terms of language derivation visualized as an evolutionary tree. This phylogenetic point of view, typical of Indo-European studies, has however been challenged by several linguists who rather explain language closeness in the Finno-ugric domain in terms of borrowings through language contact rather than inheritance, an approach sometimes called the areal influence point of view.

In order to look for similarities among different languages, the American linguist Swadesh [1] has proposed a list of 207 common and human-centered words that surely appear in the largest possible number of languages (see Table 1). Building on works by Greenberg on Eurasiatic (a work parallel to that on Nostratic by Dolgoplsky and his colleagues), a trial made by Ruhlen [2] at extending

pl	cs	ro	it	es	fr	en	da	de
ja	*já*	*eu*	*io*	*yo*	*je*	*I*	*jeg*	*ich*
ty	*ty*	*tu*	*tu*	*tú*	*tu*	*you*	*du*	*du*
on	*on*	*el*	*egli*	*él*	*il*	*he*	*han*	*er*
my	*my*	*noi*	*noi*	*nosotros*	*nous*	*we*	*vi*	*wir*
wy	*vy*	*voi*	*voi*	*vosotros*	*vous*	*you*	*I*	*ihr*
oni	*oni*	*ei*	*loro*	*ellos*	*ils*	*they*	*de*	*sie*
to	*tento*	*acesta*	*questo*	*este*	*ceci*	*this*	*denne*	*dieses*
tamto	*tamten*	*acela*	*quello*	*ese*	*cela*	*that*	*den*	*jenes*
tu	*tady*	*aici*	*qui*	*aquí*	*ici*	*here*	*her*	*hier*
tam	*tam*	*acolo*	*là*	*ahí*	*là*	*there*	*der*	*dort*
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 1: The beginning of the Swadesh lists for the nine European languages considered in our experiments. One word per entry only is given here.

this kind of comparison, for classification purposes, by looking for similarities in several languages from close regions at one time, led to a controversy over the method used. All these works are the results of considerable human effort by individuals.

Indeed, manual work has long been the standard in comparative linguistics and only few works in Natural language processing have tried to automatize the methods of comparative linguistics to help guess how words correspond [3], [4], or to help derive a phylogenetic classification of languages by application of statistical methods [5], [6], or even to reconstruct proto-languages [7].

2 Basics of the comparative method

The comparative method basically looks for similarities between words of similar meanings in different languages and deduces regular sound correspondences on that basis. For instance, it has long been established that Latin /s/ at the beginning of words corresponds to Ancient Greek /h/, because there exists a series of words of similar meanings in both languages exhibiting this contrast (see Table 2). The same kind of sound contrasts can of course be identified in living languages as Table 3 shows for German and Dutch and Table 4 for French, Italian and Portuguese.

The important point in the identification of sound contrasts is the regularity with which they occur. Only series of words allow for such identification and no contrast should be drawn from unique examples. In other words, structural oppositions between series of words allow to draw more reliable conclusions. We exploit this remark in the next section to specify a certain number of properties that an automatic method inspired by comparative linguistics should possess.

Latin	Ancient Greek	‘meaning’
semi	*hemi*	‘half’
sextem	*hexa*	‘six’
septem	*hepta*	‘seven’
serpens	*herpes*	‘a snake’
similis	*homolos*	‘similar’

Table 2: A series of words in Latin and Ancient Greek that have the same meaning: Latin /s/ corresponds to Ancient Greek /h/ at the beginning of a word.

3 Linguistic specifications

Avoiding direct sound similarities The amateur misinterpretation of the comparative method is to consider mere anecdotal similarities between words in different languages as meaningful.[1] The history of comparative linguistics itself exhibits some examples where words first considered as phonetic variations have been later reinterpreted as not connected: German *haben* was first considered as sharing the same root with Latin *habēre*, when it is now recognized that Lat. *capĕre* is indeed its corresponding form. The method used by Ruhlen, originally proposed by Greenberg and known as “massive comparison,” has been mostly criticized from this point of view, although linguists perfectly know that the evolution of sounds has to be studied thoroughly to explain in the end the differences in forms observed in different languages.

In order to discard any temptation into looking at mere similarities, an automatic method to measure proximity between languages that is not equipped with a linguist’s knowledge of sound evolution, should ideally not look at mere similarities between words across languages. The best way to implement such a method that avoids looking at the substance of words is to simply make it insensitive to encoding across languages.

Avoiding isolated loan words A robust method for measuring proximity between languages should also avoid to look at isolated loan words as they are a source of errors in the characterization of a language. If a word has been borrowed from a different language and for that reason still resembles the original word, this fact should be simply ignored, unless the borrowed word finds an adequate place in the structure of the borrowing language.

An automatic method inspired by the comparative method should thus ideally look for corresponding structures in the vocabularies of the languages considered rather than looking at individual words. It should thus concentrate on detecting regular series of aligned contrasts, *i.e.*, it should be able to detect regular series of corresponding sounds (or letters), whatever the sounds (or letters), as illustrated in Tables 2, 3 and 4.

[1] Ancient Greek $\gamma\epsilon\lambda\tilde{\alpha}\nu$ /gelan/ ’to laugh’ and Japanese /gela-gela to warau/ ’to laugh loudly’ do not make Ancient Greek and Japanese close in any way!

German	Dutch	‘meaning’
Haus	*huis*	‘house’
Schaum	*schuim*	‘foam’
braun	*bruin*	‘brown’
ausbreiten	*uitbreiden*	‘extend’
Weltraum	*wereldruim*	‘space’

Table 3: A series of words in German and Dutch that have the same meaning: German /au/ corresponds to Dutch /ui/.

French	Italian	Portuguese	‘meaning’
plan	*piano*	*prano*	‘a plane’
plat	*piatto*	*prato*	‘a plate’
plaie	*piaga*	*praga*	‘an injury’
plage	*spiaggia*	*praia*	‘a beach’
plaisir	*piacere*	*prazer*	‘to please’

Table 4: A series of words in French, Italian and Portuguese that have the same meaning: /pla/ in French corresponds to /pia/ in Italian and to /pra/ in Portuguese.

Measuring areal influence that counts In opposition to a purely phylogenetic goal, a method to measure closeness between languages should respect the degree by which the vocabulary structures of two languages correspond, as structures constitute the characteristics of a given language. Indeed, a productive structure in a language characterizes that language whatever its origin, be the structure inherited from history through the application of phonetic laws (French *-té* from Latin *-tas, -tatis*) or be it massively borrowed from a neighboring language with phonetic transposition (English -ty or German *-tät,* from French *-té* or Latin *-tas, -tatis*).

In consequence, in our opinion, a measure of closeness between languages should not only measure phylogenetic kinship, but also the degree of similarity induced by areal influence or language contact as the degree of similarity between the vocabulary structures of two languages equally characterizes both of these languages.

Measuring the similarity of vocabulary structures We thus propose to concentrate on the amount of structures shared by two different languages. To this end, the method should be ignorant of accidental borrowings, but should consistently count systematic borrowings. In this sense, the massive presence of French words (a quarter to a half in written English texts) that constitute a system in that language (*e.g.,* nouns in -ty as opposed to nouns in -ness) should be identified by the method, but anecdotal borrowings of words from, say, Japanese, like *sushi,* *geisha*, etc. that do not enter in any consistent series should not be accounted for.

4 Formal specifications

4.1 Recent works on vocabulary structure

Recently a certain number of studies in Natural Language Processing have exploited the structure of vocabularies for different purposes or delivered some insights into it: [8] shows how the morphological structure of words in five languages, French, English, Turkish, Finnish, and Arabic, can be abstracted thanks to analogy. [9] shows how word forms relate to their meaning through analogical relations and how they can be placed in graphs that exploit regular oppositions like:

connector : to connect :: editor : to edit.

This ability for words to find a place in such formal and semantic structures has been exploited to coin terminological equivalents in the medical domain [10] or to translate unknown words to feed a machine translation system [11], [12].

In linguistics, some recent studies in morphology also aim at rendering an account of the organization of the vocabulary of a language by trying to make it emerge automatically through word segmentation into stems and affixes [13].[2] On the contrary, the proponents of Whole-word morphology refuse to cut down words into pieces: they consider that the positions of words in lattices structured by analogy give a view on the vocabulary that is as rich as the standard view while it avoids the necessity to solve some undecidable problems of segmentation [14].

4.2 Analogy in morphology

All the above-mentioned studies rely on analogy between words. Analogies can be seen either on the semantic level: 'traffic : street :: water : riverbed' [15] or on the formal level as a relationship between any kind of character strings:

$$aaaabbbb : aabb :: aaabbb : ab.$$

[16] proposed a formalization of analogies between strings of characters in terms of factors, *i.e.,* through adequate decomposition of strings in terms of permuting substrings, an idea that amounts to cutting words into presumed stems and affixes. As our goal is to exploit the structure of the vocabularies of languages without a necessity to decompose words into parts, we shall prefer the formalization proposed in [17] and adhere to the view of Whole-word morphology that the structure of a vocabulary can be captured without breaking words into pieces. The chosen formalization will also avoid some spurious analogies, as the

[2] These ideas go back to Z. Harris himself.

definition in [16] is claimed to be a generalization of that in [17], the latter being thus more restrictive than the former.

According to this formalization, a 4-tuple of strings, A, B, C, and D, forms an analogy only if:

$$\begin{cases} |A|_x - |B|_x = |C|_x - |D|_x, \forall x \\ \quad d(A, B) = d(C, D) \end{cases}$$

where $|A|_x$ is the number of occurrences of character x in string A. d is the edit distance that involves only insertion and deletion with equal weights.[3] As B and C may be exchanged in any analogy, the two constraints above have also to be verified for A, C, B, and D in that order, so that $d(A, C) = d(B, D)$ has also to be verified.[4] With this definition,

abundant : abundance :: present : presence

constitues an analogy as one verifies $d(A, B) = d(C, D) = 3$, and $d(A, C) = d(B, D) = 11$, and the constraint on the number of occurrences holds for each character. We illustrate it for ‘e’ only:

$$\begin{array}{ccccccc} |\text{abundant}|_\text{e} & - & |\text{abundance}|_\text{e} & = & |\text{present}|_\text{e} & - & |\text{presence}|_\text{e} \\ 0 & - & 1 & = & 2 & - & 3 \end{array}$$

This definition implies an important property: analogy is insensitive to encoding. Any one-to-one correspondence between alphabets will leave any analogy invariant. For instance, *‘bcvoebou* : *bcvoebodf* :: *qsftfou* : *qsftfodf ’* holds for exactly the same reasons as the reasons for which the analogy ‘abundant : abundance :: present : presence’ holds, as the former one has been derived from the latter one by application of Caesar’s cipher, *i.e.,* replacing each letter with the following letter in the alphabet.

4.3 A measure of similarity between vocabulary structures

From the above ideas that the structure of the vocabulary of a language is captured by all analogies that can be formed between its elements, *i.e.,* words, without necessarily trying to cut down words into components, it is easy to derive a natural measure of the similarity between the vocabularies of two different languages. This measure is:

> the size of vocabulary structure common to two languages; that is, the proportion of the structure of the vocabulary of one language that can be transposed in the second language through translation.

[3] Slightly different from the Levenshtein distance that has substitution as an additional edit operation.

[4] Trivially, $|A|_x - |B|_x = |C|_x - |D|_x \Leftrightarrow |A|_x - |C|_x = |B|_x - |D|_x$.

One can naturally compute this quantity as a Dice coefficient, by taking the number of analogies in common in both vocabularies divided by the sum of the numbers of analogies in each of the vocabularies of the two languages, $\mathcal{L}_1$ and $\mathcal{L}_2$:

$$\frac{2 \quad \times \quad \#\text{ of analogies in common through translation}}{\#\text{ of analogies in } \mathcal{L}_1 \quad + \quad \#\text{ of analogies in } \mathcal{L}_2}$$

	Polish	Danish	meaning
A	*oddziału*	*filialens*	‘of a subsidiary’
B	*oddziałów*	*filialer*	‘of the subsidiaries’
C	*wynalazku*	*opfindelsens*	‘of an invention’
D	*wynalazków*	*opfindelser*	‘of the inventions’
	Polish	French	meaning
A	*farmaceutyczne (pl.)*	*pharmaceutiques*	‘pharmaceutical’
B	*farmaceutycznej (f.sg.gen.)*	*pharmaceutique*	‘pharmaceutical’
C	*wspólne (pl.)*	*communes*	‘common’
D	*wspólnej (f.sg.gen.)*	*commune (f.)*	‘common’
A	*przedstawiciela (gen.)*	*représentant*	‘representative (n.)’
B	*przedstawicieli*	*représentants*	‘representatives (n.)’
C	*wierzyciela (gen.)*	*créancier*	‘creditor’
D	*wierzycieli*	*créanciers*	‘creditors’
	Polish	Spanish	meaning
A	*dostosowanie*	*adaptación*	‘adaptation’
B	*dostosowania*	*adaptaciones*	‘adaptations’
C	*wyłaczenie*	*exención*	‘unplugging (sg)’
D	*wyłaczenia*	*exenciones*	‘unplugging (pl.)’
A	*desynfekujacy*	*desinfettanto*	‘disinfectant’
B	*desynfekujace*	*desinfettanti*	‘disinfectants’
C	*wojskowy*	*militario*	‘military (sg.)’
D	*woyskowe*	*militarii*	‘military (pl.)’
	Polish	Italian	meaning
A	*komitet*	*comiteto*	‘comity’
B	*komitety*	*comiteti*	‘comities’
C	*port*	*porto*	‘harbor’
D	*porty*	*porti*	‘harbors’

Table 5: Series of words in different languages, output in our experiments, that have the same meaning and share the same analogical structure, *i.e.*, in each language, A is to B as C is to D. The structure is in correspondence, but the words are not necessarily etymologically related.

Table 5 shows examples of analogies in common through translation between two languages. The measure defined above meets the requirements mentioned earlier.

Firstly, any language is maximally close to itself according to this measure, as the proportion of analogies found in common with itself is 1.

Secondly, the measure is insensitive to encoding as required by the rationale in Section 3. Consequently, any analogy in a language will remain an analogy under

language	code	family
Polish	pl	Slavic language
Czech	cs	Slavic language
Romanian	ro	Romance language + Slavic influence
Italian	it	Romance language
Spanish	es	Romance language
French	fr	Romance language
English	en	Germanic language + Romance influence
Danish	da	Germanic language
German	de	Germanic language

Table 6: Languages used in our experiments.

any one-to-one mapping between alphabets, yielding a measure of 1 between two transcriptions of the same language.[5] In this way, any language having undergone a general shift in phonemes (or letters), will remain fundamentally the same for the proposed measure.

Thirdly, such a measure renders an account of the commonality in structures between two languages by taking into account the structural sub-systems that may have been borrowed by a language from another one.

5 Experiments and results

5.1 Languages and purpose of the experiments

We tested the proposed measure of proximity between languages on nine European languages for which the family and the historical links are well established (see Table 6). Let us repeat that the measure is not designed to derive a phylogenetic tree from the figures obtained. Rather, what is expected is really a measure of closeness between languages that will reflect either a common ancestral origin or structurally consistent borrowings between the two languages. In this respect, the proximity between English and French should be spotted by the measure, the former having borrowed a good part of its vocabulary, and hence a good part of the structure of its vocabulary, from the Old French Anglo-Norman dialect.

5.2 Experiments with Swadesh lists

The first experiment we performed was intentionally a negative one: we applied the proposed method to the 207 word long Swadesh lists of the nine selected

[5] This ensures that Turkish or Mongolian or Malay will be recognized as the same language and will get a near score of 1 when processed as two different languages in their two different respective transcriptions: Arabic or Latin, Mongolian or Cyrillic, Jawi or Latin. A perfect score of 1 may not be reached because of some subtleties in transcription rules.

pl	cs	ro	it	es	fr	en	da	de
agencji	agentury	agenţiei	agenzia	agencia	agence	agency	agenturets	agentur
austrii	rakousku	austria	austria	austria	autriche	austria	østrig	österreich
asystenci	pomocní	auxiliar	ausiliario	auxiliares	auxiliaires	assistants	medhjaelpere	hilfskräfte
lat	let	ani	anni	años	ans	years	år	jahren
kanadą	kanadou	canada	canada	canadá	canada	canada	canada	kanada
ewg	ehs	cee	nel	cee	cee	eec	eoef	ewg
centralnego	centrální	centrale	centrale	central	centrale	central	centralbank	zentralbank
rady	prostředků	consiliului	commissione	comisión	commission	commission	kommissionens	kommission
jurysdykcja	příslušnost	competenţa	competenza	competencia	compétence	jurisdiction	kompetence	zuständigkeit
podsumowanie	závěr	concluzii	conclusione	conclusión	conclusion	conclusion	konklusion	ergebnis
kontyngenty	kvóty	contingentele	contingenti	contingentes	contingents	quotas	kontingenter	kontingente
uchyla	zrušuje	abrogă	abrogato	deroga	abrogé	repealed	udgår	aufgehoben
uchylenie	zrušení	abrogare	abrogazione	derogaciones	abrogatoires	repeal	ophævelse	aufhebung
rozpowszechnianie	rozšiřování	dezvăluirea	diffusione	difusión	diffusion	dissemination	formidling	verbreitung
dni	kdy	zile	giorni	días	jours	days	dage	tage
ementaler	ementál	emmental	emmental	emmental	emmental	emmentaler	emmental	emmentaler
gwarancje	jistota	garanţii	cauzione	garantías	garanties	guarantees	garantier	garantien
generoso	generoso	generoso	generoso	generoso	generoso	generoso	generoso	generoso
grupa	skupina	grupa	gruppo	grupo	groupe	group	gruppe	gruppe
szpitalach	nemocnicích	spitale	ospedali	hospitales	hôpitaux	hospitals	hospitaler	krankenhäusern
mrl	mrl	lmr	lmr	ingestión	lmr	mrl	mrl	mrl
mleka	mléka	lapte	latte	leche	lait	milk	mælk	milch
lolium	lolium	lolium	lolium	lolium	lolium	lolium	lolium	lolium
środki	prostředky	mijloace	mezzi	medios	moyens	means	midler	mittel
grzywny	pokuty	amenzi	ammende	multas	amendes	fines	bøder	geldbußen
murfatlar	murfatlar	murfatlar	murfatlar	murfatlar	murfatlar	murfatlar	murfatlar	murfatlar
nafo	nafo	nafo	nafo	nafo	nafo	nafo	nafo	nafo
lub	nebo	sau	uno	o	ou	or	eller	eines
października	října	libera	ottobre	octubre	octobre	october	oktober	oktober
oferowaną	nabízené	oferită	offerto	ofrecida	offerte	offered	tilbudte	angebotsmenge
pomoru	moru	pestei	peste	peste	peste	fever	svinepest	schweinepest
przeglądów	provedená	revizuirilor	riesami	revisiones	révisions	reviews	vurderinger	durchführt
spec	lolium	spec	spec	spec	spec	spec	spec	spec
skreślony	zrušuje	elimină	soppresso	suprime	supprimé	deleted	udgår	gestrichen
skreśla	zrušuje	elimină	soppresso	suprimido	supprimé	deleted	ophaeves	entfällt
toksykologia	toxikologie	toxicologie	tossicologia	toxicología	toxicologie	toxicology	toksikologi	toxikologie
traktatu	smlouvy	hotărând	trattato	tratado	traité	treaty	traktatens	vertrag
obowiązywania	platnosti	durata	validità	validez	validité	validity	forordningens	geltungsdauer

Table 7: A sample of the multilingual lexicon of 3,833 entries extracted from the *Acquis communautaire*.

European languages.[6] It is obvious at first sight that Swadesh lists do not exhibit the kind of analogical structures our method looks for. The result obtained confirms this: on all languages, only four analogies were found (one in English: all : ash :: to pull : to push) with no single analogy common to any two different languages through translation.

This clearly makes the point that our method does not rely on similarities that can be established directly between the elements of the vocabularies of two languages. We argued that this is indeed desirable for the method to be able to still recognize as identical, languages that would have undergone some general phonetic shift.

5.3 Experiments with a multilingual lexicon extracted from the *Acquis communautaire*

In a second experiment, we use a multilingual lexicon obtained from a multilingual corpus made of 86,005 lines taken from the *Acquis communautaire*.[7] These lines were aligned on the sub-sentential level in one pass using the multi-

[6] Source: `http://en.wiktionary.org/`

[7] `http://langtech.jrc.it/JRC-Acquis.html`

	pl	cs	ro	it	es	fr	en	da	de
pl	.	**103**	37	26	27	36	48	40	44
cs	**103**	.	31	21	30	34	48	36	43
ro	37	31	.	36	**47**	**47**	34	26	31
it	26	21	36	.	123	**142**	79	29	30
es	27	30	**47**	123	.	**270**	136	38	43
fr	36	34	**47**	**142**	**270**	.	**222**	48	56
en	48	48	34	79	136	**222**	.	53	56
da	40	36	26	29	38	48	53	.	**67**
de	44	43	31	30	43	56	56	**67**	.

Table 8: Proximity between nine European languages obtained by measuring commonality of vocabulary structures. The values are computed according to the formula given in Section 4.3 multiplied by 10^3 for higher readibility. For each language, the highest score on the corresponding line is typeset in boldface and is then reported by symmetry on the corresponding column. The same is done for the weakest scores with the gray color.

lingual sub-sentential aligner `anymalign`[8] with options `-n 1 -N 1` to get word alignments only. This resulted in 7,462 word alignments. From these, we deleted all alignments consisting of numbers or the like, which gave a final multilingual lexicon of 3,833 entries for each different language. A sample is shown in Table 7.

The number of analogies obtained with the previous 3,833 words in each language is listed below.

pl	cs	ro	it	es	fr	en	da	de
12,523	11,089	22,155	16,554	14,479	22,756	20,183	13,069	12,333

Table 8 summarizes the measures of proximity obtained by counting the number of analogies in common across vocabularies through translation, as defined in Section 4.3 These measures reflect the usual knowledge about the proximity of these nine European languages. In particular, the mutual high scores exhibited by Polish and Czech on one side, Romanian, Italian, Spanish and French on another one, and German and Danish on a third one, reflect the three main language families represented by these languages. In addition, according to these measures, English is closer to Romance languages than to the Germanic language family because of the overwhelming attested influence of Anglo-Norman on the structure of its vocabulary.

6 Conclusion

We have proposed a method to measure the proximity between languages that relies on the structure of the vocabularies of the languages considered. It consists

[8] `http://users.info.unicaen.fr/∼alardill/anymalign/`

in computing the Dice coefficient of the number of analogies (between words) that are common, through translation, to two languages.

We applied this measure to a multilingual lexicon of nine European languages automatically extracted from the *Acquis communautaire,* and computed a proximity matrix for these nine languages. This matrix is in general conformity with the knowledge about the relative proximity of these nine languages.

The main problem encountered in our experiments is the availability of data. Without large enough multilingual lexicon of the same size in each language, it will remain difficult to try to solve some of the haunting problems in language closeness by automatic means, *e.g.*:

- Is Basque really related to Caucasian languages?
- Do Korean and Japanese share some vocabulary structure and in which area of their vocabulary?
- Is the bold hypothesis that the Japanese vocabulary concerning rice growing relates to Dravidian languages, valid?

However, with the increase in free resources like wiktionaries, or the increase in the number of translated materials available on the Internet from which parallel lexicons can be extracted by applying alignment tools, we remain optimistic in the possibility of conducting new experiments on a larger number of languages.

References

1. Swadesh, M.: Lexico-statistic dating of prehistory ethnic contacts, with special reference to north american indians and eskimos. Proc. Am. Philos. Soc. **95** (1952) 453–462.
2. Ruhlen, M.: The Origin of Language: tracing the evolution of the mother tongue. John Wiley & Sons, New York (1994).
3. Covington, M.A.: An algorithm to align words for historical comparison. Computational Linguistics **22** (1996) 481–496.
4. Kondrak, G.: Identifying complex sound correspondences in bilingual wordlists. In: CICLing. (2003) 432–443.
5. Gray, R.D., Atkinson, Q.D.: Language-tree divergence times support the Anatolian theory of Indo-European origin. Nature **426** (2003) 435–439.
6. Rexová, K., Frynta, D., Zrzavý, J.: Cladistic analysis of languages: Indo-european classification based on lexicostatistical data. Cladistics **19** (2005) 120–127.
7. Lowe, J.B., Mazaudon, M.T.: The reconstruction engine: A computer implementation of the comparative method. Computational Linguistics **20** (1994) 381–417.
8. Lavallée, J.F., Langlais, P.: Unsupervised morphology acquisition by formal analogy. In: Lecture Notes in Computer Science. (2010) 8 pages.
9. Claveau, V., L'Homme, M.C.: Terminology by analogy-based machine learning. In: Proceedings of the 7th International Conference on Terminology and Knowledge Engineering, TKE 2005, Copenhagen (Denmark) (2005).
10. Langlais, P., Yvon, F., Zweigenbaum, P. In: Analogical Translation of Medical Words in Different Languages. Volume 5221/2008 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg, Springer Berlin / Heidelberg (2008) 284–295.

11. Langlais, P., Patry, A.: Translating unknown words by analogical learning. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL). (2007) 877–886.
12. Denoual, E.: Analogical translation of unknown words in a statistical machine translation framework. In: Proceedings of Machine Translation Summit XI, Copenhagen (2007).
13. Goldsmith, J.: Morphological analogy: Only a beginning. unpublished (2007) [Available online] http://hum.uchicago.edu/ jagoldsm/Papers/analogy.pdf.
14. Neuvel, S., Singh, R.: Vive la différence ! what morphology is about. Folia Linguistica **35** (2001) 313–320.
15. Turney, P.: A uniform approach to analogies, synonyms, antonyms, and associations. In: Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), Manchester, UK, Coling 2008 Organizing Committee (2008) 905–912.
16. Stroppa, N., Yvon, F.: An analogical learner for morphological analysis. In: Proceedings of the 9th Conference on Computational Natural Language Learning (CoNLL 2005), Ann Arbor, MI (2005) 120–127.
17. Lepage, Y.: Analogy and formal languages. Electronic notes in theoretical computer science **47** (2004) 180–191.