Commonality across vocabulary structures as an estimate of the proximity between languages

Yves Lepage, Adrien Lardilleux, Julien Gosme

Commonality across vocabulary structures as an estimate of the proximity between languages

Yves Lepage, Adrien Lardilleux and Julien Gosme

GREYC, university of Caen, BP 5186, 14032 Caen Cedex, France
firstname.lastname@info.unicaen.fr

Abstract

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

Keywords: Proximity between languages, comparative linguistics, analogy.

1. Introduction

This paper 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, Coeurdoux 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 (Swadesh, 1952) 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 hypothesis by Dolgoplsky and his colleagues), a trial made by Ruhlen (Ruhlen, 1994) at extending 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.

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 (Covington, 1996), (Kondrak, 2003), or to help derive a phylogenetic classification of languages by application of statistical methods (Gray and Atkinson, 2003), (Rexová et al., 2005), or even to reconstruct proto-languages (Lowe and Mazaudon, 1994).

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	nos-otros	nous	we	vi	wir
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

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.

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.

Latin	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 Greek /h/ at the beginning of a word.

The important point in such 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.

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/.

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. 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) are, as in Tables 2 and 3.

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 to deliver some insights into it: (Claveau and L'Homme, 2005) try to show how word forms relate to their meaning 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 (Langlais et al., 2008) or to translate unknown words to feed a machine translation system (Langlais and Patry, 2006), (Denoual, 2007).

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 (Goldsmith, 2007).[1] On the contrary, (Neuvel and Singh, 2001) in the presentation of their 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.

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

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' (Turney, 2008) or on the formal level as a relationship between any kind of character strings: *'aaaabbbb : aabb :: aaabbb : ab.'*

(Stroppa and Yvon, 2005) 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 (Lepage, 2004) 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 definition in (Stroppa and Yvon, 2005) is claimed to be a generalization of that in (Lepage, 2004), 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:

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

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.[2] 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.[3] 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.

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 $\mathcal{L}_2$ considered:

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

Table 4 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. According to the definition given above, analogy is insensitive to encoding. Consequently, any analogy in a language will remain an analogy under any one-to-one mapping between alphabets, yielding a measure of 1 between two transcriptions of the same language.[4] 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 5). 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.

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

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

[4] 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.

	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	Spanish	'meaning'
A	*dostosowanie*	*adaptación*	'adaptation'
B	*dostosowania*	*adaptaciones*	'adaptations'
C	*wyłaczenie*	*exención*	'unplugging (sg)'
D	*wyłaczenia*	*exenciones*	'unplugging (pl.)'

Table 4: 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.

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 5: Languages used in our experiments.

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 European languages.[5] 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*.[6] These lines were aligned on the sub-sentential level in one pass using the multilingual sub-sentential aligner `anymalign`.[7] 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	6,988	fr	19,169
cs	5,748	en	13,608
ro	13,515	da	7,321
it	12,952	de	6,678
es	11,623		

Table 6 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.

	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 6: 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.

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 in computing the Dice coefficient of the number of analogies between words, 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.

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

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

[7] `http://users.info.unicaen.fr/~alardill/anymalign/`

pl	cs	ro	it	es	fr	en	da	de
źródła	zdroj	surse	fonte	fuente	sources	source	kilde	quelle
wszystkie	všechny	toate	tutte	todas	toutes	all	alle	aus
asystenci	pomocní	auxiliar	ausiliario	auxiliares	auxiliaires	assistants	medhjaelpere	hilfskräfte
budżecie	rozpočtu	bugetul	bilancio	presupuesto	budget	budget	budget	gesamthaushaltsplan
tiret	odrážky	liniuţă	trattino	guión	tiret	indent	led	gedankenstrich
dalej	jen	continuare	seguito	denominado	après	hereinafter	benævnt	genannt
gutunek	druh	specia	specie	especie	espèce	species	art	art
uchyla	zrušuje	abrogă	abrogato	derogado	abrogé	repealed	udgår	gestrichen
i	a	şi	alle	y	et	and	og	und
pigmenty	pigmentů	pigmenţi	pigmenti	pigmentos	pigments	pigments	pigmenter	pigmente
czerwca	června	consiliului	giugno	junio	juin	june	juni	juni
zbóż	obiloviny	cerealelor	cereali	cereales	céréales	cereals	korn	getreide
hiszpańskim	španělštině	spaniolă	spagnola	española	espagnole	spanish	spansk	spanisch

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

References

Claveau, Vincent and Marie-Claude L'Homme, 2005. Terminology by analogy-based machine learning. In *Proceedings of the 7th International Conference on Terminology and Knowledge Engineering, TKE 2005*. Copenhagen (Denmark).

Covington, Michael A., 1996. An algorithm to align words for historical comparison. *Computational Linguistics*, 22(4):481–496.

Denoual, Etienne, 2007. Analogical translation of unknown words in a statistical machine translation framework. In *Proceedings of Machine Translation Summit XI*. Copenhagen.

Goldsmith, John, 2007. Morphological analogy: Only a beginning.

Gray, Russell D. and Quentin D. Atkinson, 2003. Language-tree divergence times support the anatolian theory of indo-european origin. *Nature*, 426:435–439.

Kondrak, Grzegorz, 2003. Identifying complex sound correspondences in bilingual wordlists. In *CICLing*.

Langlais, Philippe and Alexandre Patry, 2006. 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)*.

Langlais, Philippe, François Yvon, and Pierre Zweigenbaum, 2008. *Analogical Translation of Medical Words in Different Languages*, volume 5221/2008 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg: Springer Berlin / Heidelberg, pages 284–295.

Lepage, Yves, 2004. Analogy and formal languages. *Electronic notes in theoretical computer science*, 47:180–191.

Lowe, John B. and Martine T. Mazaudon, 1994. The reconstruction engine: A computer implementation of the comparative method. *Computational Linguistics*, 20:381–417.

Neuvel, Sylvain and Rajendra Singh, 2001. Vive la différence ! what morphology is about. *Folia Linguistica*, 35(3-4):313–320.

Rexová, Kateřina, Daniel Frynta, and Jan Zrzavý, 2005. Cladistic analysis of languages: Indo-european classification based on lexicostatistical data. *Cladistics*, 19(2):120–127.

Ruhlen, Merritt, 1994. *The Origin of Language: tracing the evolution of the mother tongue*. New York: John Wiley & Sons.

Stroppa, Nicolas and François Yvon, 2005. An analogical learner for morphological analysis. In *Proceedings of the 9th Conference on Computational Natural Language Learning (CoNLL 2005)*. Ann Arbor, MI.

Swadesh, M., 1952. Lexico-statistic dating of prehistory ethnic contacts, with special reference to north american indians and eskimos. *Proc. Am. Philos. Soc.*, 95:453–462.

Turney, Peter, 2008. 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.