# Technical Correspondence

## Automatic Clustering of Languages

Vladimir Batagelj*†
University of Ljubljana

Tomaž Pisanski†
University of Ljubljana

Damijana Keržič‡
Jožef Stefan Institute

*Automatic clustering of languages seems to be one possible application that arose during our study of mathematical methods for computing dissimilarities between strings. The results of this experiment are discussed.*

### 1. Introduction

The purpose of this paper is to show that current mathematics and computer science can offer expertise to various "soft" sciences, e.g., linguistics. Sixty-five languages are automatically grouped into clusters according to the analysis of sixteen common words. The authors regard the results presented in this paper merely as an example of a possible application of cluster analysis to linguistics. The results should not be regarded as conclusive but rather as suggestions to linguists that similar projects can be carried out on a much greater scale, hopefully yielding similar results and better understanding of language families.

This is by no means the first application of mathematical methods to this problem; see for instance Kruskal, Dyen, and Black (1971) and Sujoldžić et al. (1987).

### 2. Problem and Data

It is more or less clear that some words are similar in certain languages and dissimilar in other languages. Obviously two languages are *similar* if most words are similar. Therefore the most general problem is to determine for each pair of languages how similar or how dissimilar they are. Is Spanish closer to Latin than English to Danish? In general, perhaps such quantitative questions do not always make sense. But suppose we decide to make an experiment. Suppose we decide to measure dissimilarity between two languages by defining it in a strict mathematical manner. From the linguistic viewpoint this may be quite absurd. Nevertheless we have defined certain ways to measure dissimilarity between two words and used this to measure dissimilarity between two languages. There are several ways one can define such a dissimilarity. In this paper we will show some examples. The choice of the dissimilarity will of course influence the outcome. It is interesting that changing the choice of the dissimilarity does not affect the outcome too drastically. It is for the linguists to tell whether this can be interpreted by saying that the results are stable, i.e., "almost independent" of the choice of dissimilarity functions and make sense for the languages.

* Supported in part by the Research Council of Slovenia.
† Department of Mathematics, University of Ljubljana, Ljubljana, Slovenia.
‡ Department of Digital Communications, Jožef Stefan Institute, Ljubljana, Slovenia.

1. Let $u$ be a word in a language $L_1$ and let $v$ be its translation into another language $L_2$. Let $d(u,v)$ be a **dissimilarity measure** or simply **dissimilarity** between the two words as it is described below. Hence $d(u,v)$ is a nonnegative integer. In order to make things simpler we assume that both languages are written in the same alphabet. Let us give some examples for dissimilarity $d(u,v)$.

   (a) Assume that $d_1(u,v)$ is the minimum number of the letters that have to be **inserted** or **deleted** in order to change $u$ into $v$. For example:

   $$\begin{aligned} u &= belly \\ v &= bauch. \end{aligned}$$

   Obviously in order to transform $u$ into $v$ we have to delete the letters *"elly"* and insert the letters *"auch."* Hence $d_1(belly, bauch) = 8$.

   (b) The second possibility is the smallest number of **substitutions**, **deletions**, and **insertions** to change $u$ into $v$.
   In our example:

   $$u = belly,\ v = bauch,\ d_2(belly, bauch) = 4.$$

   We have to substitute the letters *"elly"* with letters *"auch"* and this is the shortest way to change $u$ into $v$.

   Both $d_1(u,v)$ and $d_2(u,v)$ are called the **Levenshtein distance** (Kruskal 1983).

   (c) We can measure dissimilarity between two words also with the length of their shortest common supersequence (LSCS). Any "word" (string) $z$ is a **supersequence** of a word $u$ if it can be obtained from $u$ by inserting letters into it.

   For example:
   if $u = belly$, $v = bauch$, then some possibilities for their shortest common supersequence are *"bellyauch,"* *"bealulcyh,"* *"belauchly,"* ... They all contain 9 characters. Therefore, $d_3(belly, bauch) = 9$.

   There are other possibilities for defining dissimilarity $d(u,v)$ that have been used in data analysis; see for instance Kashyap and Oommen (1983).

2. In our study we have used only written languages and dialects. We used transliterations into standard Latin (English) alphabet. The data were provided from a variety of sources such as native speakers and dictionaries. However, transliterations were not checked. The translations were not given by experts; hence it is quite likely that there are several inconsistencies present both in translations and in transliterations. Obviously the choice of a particular method of transliteration and translation may influence the outcome.

   The letters that do not appear in the Latin alphabet were changed into similar letters of the Latin alphabet. For example: in the Slovenian alphabet there are three nonstandard letters *č, š, ž*. We have chosen to omit diacritical marks: *c, s,* and *z*. A possible alternative would be to use *ch, sh, zh*. Also we omit diacritical marks in other languages. For instance: *ä, à, á* are represented as *a*.

| | | 1. | 2. | ... | n. |
|---|---|---|---|---|---|
| *Language* | $L_1$ | $w_{11}$ | $w_{12}$ | ... | $w_{1n}$ |
| *Language* | $L_2$ | $w_{21}$ | $w_{22}$ | ... | $w_{2n}$ |
| | | ... | ... | ... | ... |
| *Language* | $L_m$ | $w_{m1}$ | $w_{m2}$ | ... | $w_{mn}$ |

**Figure 1**
Data array.

3. We have chosen 16 English words. Actually, we have started with data in Hartigan's *Clustering Algorithms*, page 243. Later we used *The Concise Dictionary of 26 Languages in Simultaneous Translation* to expand the data. Over 30 people all over the world have given corrections and data for lesser known languages and dialects. The resulting data are given in Appendix A.

   Only linguists should carefully select the words that would be used in the "real" project. We hope that they will contact us in order to carry out the "big" project. For some well-studied sets of words the reader should consult Kruskal, Dyen, and Black (1971) and Sujoldžić et al. (1987).

4. The computer program for computing dissimilarity measure uses the data about the languages in the large array shown in Figure 1.

   There are $m$ languages and $n$ words in each language. We have selected $m = 65$ languages and $n = 16$ words.

   Note that Appendix A gives essentially this array for our experiment. For instance $L_1 = Albanian$, $w_{12} = keq$.

5. Once we select a dissimilarity measure $d(u, v)$ between two words, the next step is to define the dissimilarity $D(L_i, L_j)$ between two languages. There are many possibilities. We decided to take the sum of dissimilarity measures of words. Mathematically, it is defined as:

$$D(L_i, L_j) = d(w_{i1}, w_{j1}) + d(w_{i2}, w_{j2}) + \ldots + d(w_{in}, w_{jn}).$$

   We would like to point out that this is studied by data analysis; the reader is referred to Hartigan (1971) for further discussion and background.

6. The next step is to select an appropriate **clustering method**. There are many different methods available (Hartigan 1971). We wanted to have the results expressed in the form of a binary tree (see Aho, Hopcroft, and Ullman 1974 for the discussion of binary trees) or more precisely in the form of a **dendrogram**; see for instance Anderberg (1973) and Gordon (1981).

   We selected Ward's method, which tends to give realistic results. This method is discussed in Anderberg (1973) and Gordon (1981).

## 3. Results and Comments

The results are presented in Appendix B in the form of three dendrograms. Each of them corresponds to a specified dissimilarity measure. The three results are not identical; however, they are quite similar.

If we cut the dendrogram horizontally at any height we obtain a partition of the set of the languages into a certain number of parts that we call **clusters**. The dendrogram tells us how many clusters are suitable for data that we analyze. The number of clusters we obtain from the cut at the largest "jump" of two neighboring levels of the union.

Looking at our three dendrograms we can easily notice that our data form five clusters:

- Slavic
- Germanic
- Romance
- Indic
- all others.

We can also notice that first the Slavic branch is formed. Next the Germanic and the Romanic languages form their groups (clusters) nearly at the same point. At the end the Indic languages are branching off the others. The remaining languages do not form any other evident cluster. See Figure 2.

The five clusters that are formed are very stable. Any pair of languages classified in one of our clusters in the first dendrogram are also in the same class in the other two dendrograms. Notice that in some clusters languages also form subclusters. For example look at the Germanic languages in any dendrogram where two parts are very pronounced: the Scandinavian languages and the German-related languages and dialects. It is interesting that the simplest dissimilarity measure $d_1$ (i.e., the number of insertions and deletions) gives the best separation of languages.

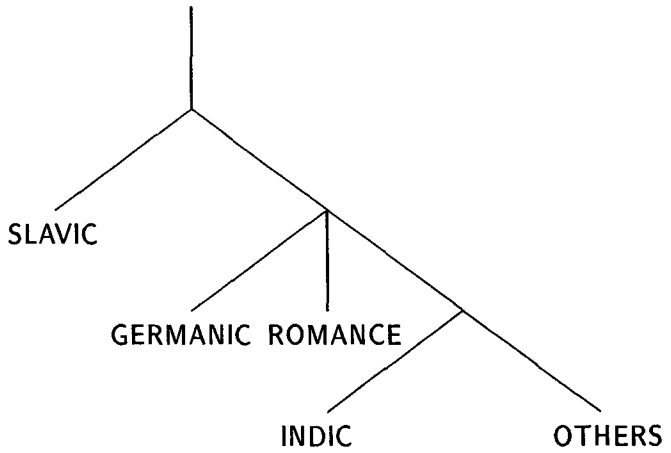



**Figure 2**
Family tree of languages.

We can mention that clusters we found with cluster analysis are very close to the language families established in linguistics (Kruskal, Dyen, and Black 1971).

Obviously one could ask the following questions or problems that can only be answered by a large-scale project.

1. In our case all treated words have equal weight. The similarity measure between two languages can also be defined in such a way that different weights (based on linguistic theory) are given to the words and/or transformations.
2. How much does the choice of words influence the final tree structure? In our analysis English belongs to the Germanic cluster, when we know that it also has a strong Romance component.
3. Obviously a larger number of words would give a more accurate picture. The question is: how much and in what way do the results vary if we increase the number of words?
4. How much would the results differ if we study spoken language instead of written language? We can consider for example some phonetic properties of written letters or strings of letters.
5. Any choice of transliteration introduces a "systematic error" in the results. One way of eliminating such an error would be to test for patterns and then not to penalize patterns that occur often. For example: if we find that *"tch"* $\longrightarrow$ *"zh"* very often then we would not count it every time it occurs but only once.

Of course for such precise analysis one needs much better knowledge of the linguistic field than we have as laypersons.

**References**

Aho, A. V.; Hopcroft, J. E.; and Ullman, J. D. (1974). *The Design and Analysis of Computer Algorithms*. Addison Wesley.

Anderberg, M. R. (1973). *Cluster Analysis for Applications*. Academic Press.

Gordon, A. D. (1981). *Classification*. Chapman and Hall.

Hartigan, J. A. (1971). *Clustering Algorithms*. John Wiley.

Kashyap, R. L., and Oommen, B. J. (1983). "A common basis for similarity measures involving two strings." *Intern. J. Computer Math.*, 13: 17–40.

Kruskal, Joseph B. (1983). "An overview of sequence comparison: Time warps, string edits, and macromolecules." *SIAM Review*, 25(2): 201–237.

Kruskal, Joseph B.; Dyen, Isidore; and Black, Paul. (1971). "Some results from the vocabulary method of reconstructing languages trees." In Lexico-Statistics in Genetic Linguistics, Proceedings of the Yale Conference, Yale University.

Sujoldžić, A.; Šimunović; Finka B.; Bennett L. A.; Angel J. L.; Roberts D. F.; and Rudan P. (1987). "Linguistic microdifferentation on the Island of Korčula." *Anthropol. Ling.*, 28: 405–432.

*The Concise Dictionary of 26 Languages in Simultaneous Translation*, compiled by P. M. Bergman. A Signet Book from New America Library.

## Appendix A. Sixteen Words in Sixty-Five Languages

| | 1. | 2. | 3. | 4. |
|---|---|---|---|---|
| ALBANIAN | gjithcka | keq | bark | galm |
| AR. TUNISIAN[1] | ilkul | xiab | kirsh | akhal |
| BAH. MALAYSIA[2] | semua | jahat | perut | hitam |
| BENGALI | sob | kharap | pet | kalo |
| BERBER | akith | diri | aaboudh | averkan |
| BULGARIAN | vseki | los | korem | ceren |
| BYELORUSSIAN | use | kepski | brukha | chrni |
| CATALAN | tot | dolent | panxa | negre |
| CH. CANTONESE[3] | chyun | waai | tou | hak |
| CH. MANDARIN[4] | dou | bu hao | du zi | hei |
| CROATIAN | sve | los | trbuh | crn |
| CROAT. CAKAVSKI[5] | se | los | trbuh | crn |
| CROAT. KAJKAVSKI[6] | sve | los | trebuh | crn |
| CZECH | vsechno | spatny | bricho | cerny |
| DANISH | all | slet | bug | sort |
| DUTCH | geheel | slecht | buik | zwart |
| ENGLISH | all | bad | belly | black |
| ESPERANTO | cio | malbona | ventro | nigra |
| FINNISH | kaikki | huono | vatsa | musta |
| FRENCH | tout | mauvais | ventre | noir |
| GERMAN | alle | schlecht | bauch | schwarz |
| GER. BAVARIAN[7] | ail-zam | schlecht | wampn | schwoaz |
| GER. SWISS D. 1[8] | aui | schlaecht | buch | schwarz |
| GER. SWISS D. 2[9] | alles | schlaecht | buch | schwarz |
| GREEK NEW | olos | kakos | kilya | mavros |
| GREEK OLD | holos | kakos | koilia | mavros |
| HEBREW | kol | ra | beten | shachor |
| HINDI | sab | kharab | pet | kala |
| HUNGARIAN | minden | rossz | has | fekete |
| INDONESIAN | semua | buruk | perut | hitam |
| ITALIAN | tutto | male | ventre | nero |
| IT. N. LOMBARDY[10] | tu:t | catiiv | pansa | negher |
| IT. VENETII D.[11] | tut | brut | panza | caif |

1 ARABIC TUNISIAN
2 BAHASA MALAYSIA
3 CHINESE CANTONESE
4 CHINESE MANDARIN
5 CROATIAN CAKAVSKI - Dialect of Croat
6 CROATIAN KAJKAVSKI - Dialect of Croat
7 GERMAN BAVARIAN
8 GERMAN SWISS DIALECT - Bernese Oberland
9 GERMAN SWISS DIALECT - Northeastern Switzerland
10 ITALIAN NORTHERN LOMBARDY
11 ITALIAN VENETII DIALECT - distinct from Venetians

| | | | | |
|---|---|---|---|---|
| IRISH | vile | olc | bolg | dubh |
| JAPANESE | zenbu | warui | hara | kuroi |
| KANNADA | yella | ketta | hoatti | kahri |
| LATIN | totus | malus | venter | niger |
| LATVIAN | visi | slikts | veders | melns |
| LITHUANIAN | vise | blogas | pilvas | jaudas |
| MACEDONIAN | site | los | stomak | crn |
| MALAYALAM | ellam | cheetta | vayaru | karuppu |
| MALTESE | kollox | trazin | zaqq | iswed |
| MAORI | katoa | kino | hoopara | hiwahiwa |
| MARAATHI | sarva | waeet | poat | kaale |
| NORWEGIAN | alle | daarlig | mage | svart |
| ORIYA | sabu | kharap | peta | kala |
| PANJABI | sab | bura | pet | kala |
| PERSIAN | hame | bad | shekam | siah |
| POLISH | wszystko | zly | brzuch | czarny |
| PORTUGUESE | todo | mau | barriga | negro |
| RAJASTHANI | sab | kharab | pet | kalo |
| ROMANIAN | tot | rau | burta | negru |
| RUSSIAN | vse | plokhoi | brjukho | cjornji |
| SANSKRIT | sara | bura | paat | kala |
| SERBIAN | sve | los | trbuh | crn |
| SLOVAK | vsetko | zly | brucho | cierny |
| SLOVENIAN | vse | slab | trebuh | crn |
| SPANISH | todo | mal | vientre | negro |
| SWAHILI | ote | baya | tumbo | karipia |
| SWEDISH | alla | daolig | mage | svart |
| TAMIL | ellaam | keduthy | vayiru | karuppu |
| TELUGU | antha | chedda | kadupu | nalla |
| TURKISH | butun | fena | karin | kara |
| UKRAINIAN | vse | pohane | zhevit | chorne |
| WELSH C | pawb | drwg | bola | du |

| | 5. | 6. | 7. | 8. |
|---|---|---|---|---|
| ALBANIAN | asht | dite | vdes | pi |
| AR. TUNISIAN | adhum | yuum | met | ushrub |
| BAHASA MALAYSIA | tulang | hari | mati | minum |
| BENGALI | harh | din | mora | khaoa |
| BERBER | ighass | as | amath | sew |
| BULGARIAN | kost | den | umiram | pi |
| BYELORUSSIAN | kostka | dzen' | pamertsi | pits' |
| CATALAN | os | dia | morir | beure |
| CH. CANTONESE | gwat | yat | sei | yam |
| CH. MANDARIN | si | tian | si | he |
| CROATIAN | kost | dan | umrijeti | piti |

| | | | | |
|---|---|---|---|---|
| CROAT. CAKAVSKI | kost | dan | umret | pit |
| CROAT. KAJKAVSKI | kost | dan | umreti | piti |
| CZECH | kost | den | umrit | piti |
| DANISH | ben | dag | at doe | at drikke |
| DUTCH | bot | dag | sterven | drinken |
| ENGLISH | bone | day | to die | to drink |
| ESPERANTO | osto | tago | morti | trinki |
| FINNISH | luu | paiva | varjata | juoda |
| FRENCH | os | jour | mourir | boire |
| GERMAN | knochen | tag | sterben | trinken |
| GER. BAVARIAN | gnocha | dag | schteam | saufn |
| GER. SWISS D. 1 | chnoche | tag | staerbe | trinke |
| GER. SWISS D. 2 | chnoche | dag | staerbe | drinke |
| GREEK NEW | kokalo | mera | petheno | pino |
| GREEK OLD | kokkalos | hemera | thneskein | pinein |
| HEBREW | etsem | yom | lamut | lishtot |
| HINDI | haddi | din | marna | pina |
| HUNGARIAN | csont | nap | hal | iszik |
| INDONESIAN | tulang | hari | mati | minum |
| ITALIAN | osso | giorno | morire | bere |
| IT. N. LOMBARDY | oss | di' | muri' | bever |
| IT. VENETII D. | os | di | morir | bever |
| IRISH | chaimh | la | doluidh | olaim |
| JAPANESE | hone | hi | shinu | nomu |
| KANNADA | yalabu | dina | satta | kudi |
| LATIN | os | dies | mori | bibere |
| LATVIAN | kauls | diena | nomirt | dzert |
| LITHUANIAN | kaulas | dena | numire | gerti |
| MACEDONIAN | koska | den | umira | pie |
| MALAYALAM | ellu | divasam | marikkuka | kudikkuka |
| MALTESE | gtradma | gurnata | miet | xorob |
| MAORI | iwi | maeuao | hemo | inu |
| MARAATHI | haad | diwas | marney | piney |
| NORWEGIAN | ben | dag | aa doe | aa drikke |
| ORIYA | hada | dina | mariba | pieeba |
| PANJABI | hadi | din | marna | pina |
| PERSIAN | ostokhan | ruz | mordan | nushidan |
| POLISH | kosc | dzien | umrzec | pic |
| PORTUGUESE | osso | dia | morrer | beber |
| RAJASTHANI | haddi | din | marno | peeno |
| ROMANIAN | os | zi | a muri | a bea |
| RUSSIAN | kost | den' | umirat | pit |
| SANSKRIT | haddi | din | marna | peena |
| SERBIAN | kost | dan | umret | piti |
| SLOVAK | kost | den | zomriet | pit |
| SLOVENIAN | kost | dan | umreti | piti |
| SPANISH | hueso | dia | morir | beber |
| SWAHILI | mfupa | siku | kufov | nywa |

| | | | | |
|---|---|---|---|---|
| SWEDISH | ben | dag | att doe | att dricka |
| TAMIL | elumbu | naal | irappu | kuditthal |
| TELUGU | yamuka | thinam | chavu | thagu |
| TURKISH | kemik | gun | olmek | icmek |
| UKRAINIAN | kistka | den' | vmerte | pihte |
| WELSH C | asgwrn | dydd | marw | yfed |

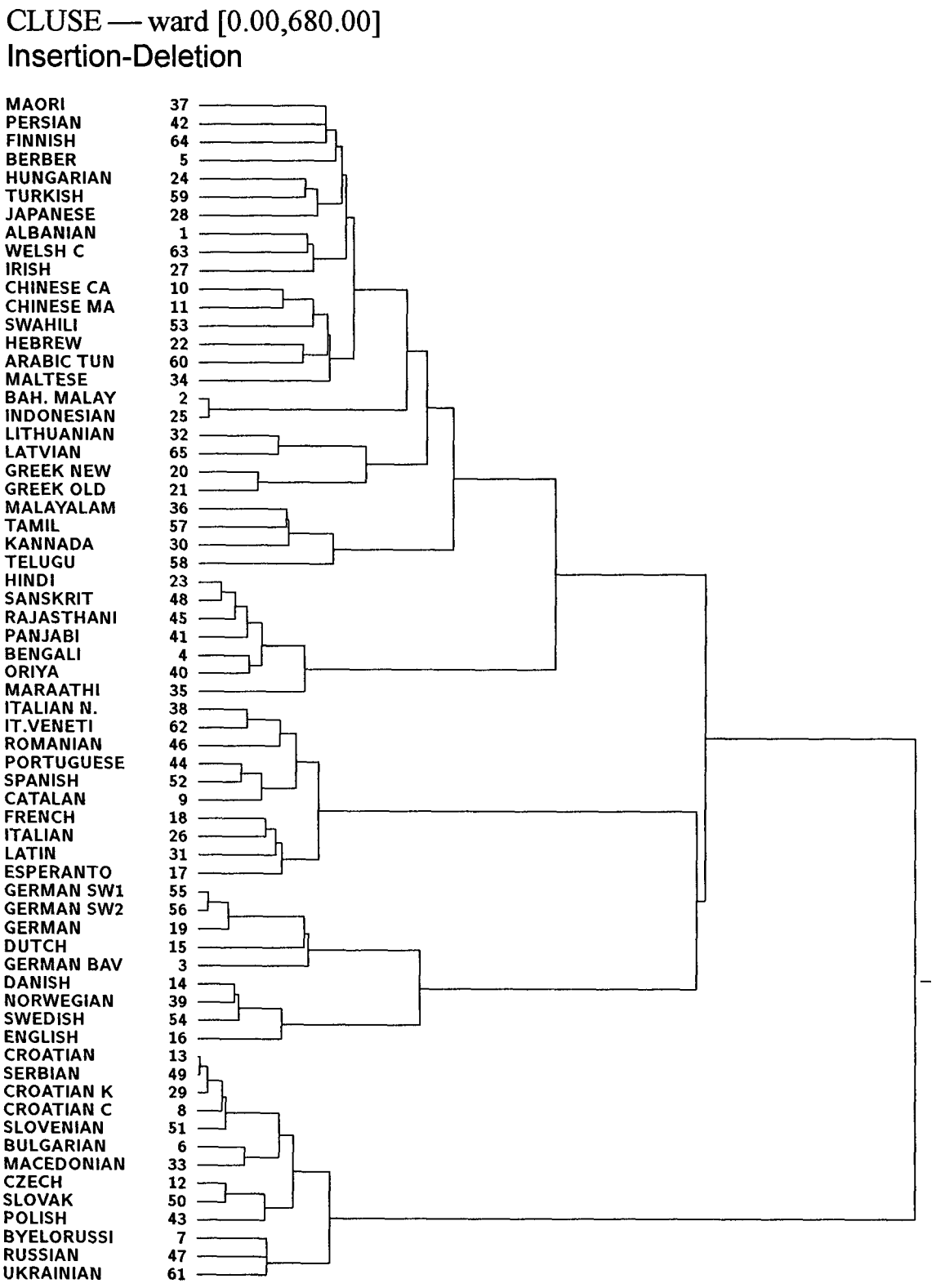
| | 9. | 10. | 11. | 12. |
|---|---|---|---|---|
| ALBANIAN | vesh | ha | ve | sy |
| AR. TUNISIAN | wdhin | akul | adhum | ain |
| BAH. MALAYSIA | telinga | makan | telur | mata |
| BENGALI | kan | khaoa | dim | chokh |
| BERBER | amazough | atch | thamalalt | thit |
| BULGARIAN | uho | jaim | jaice | oko |
| BYELORUSSIAN | vukha | estsi | yaika | voka |
| CATALAN | orella | menjar | ou | ull |
| CH. CANTONESE | yi | sik | dan | ngan |
| CH. MANDARIN | sheng | chi | dan | yen jin |
| CROATIAN | uho | jesti | jaje | oko |
| CROAT. CAKAVSKI | uho | jist | jaje | oko |
| CROAT. KAJKAVSKI | vuho | jesti | joje | oko |
| CZECH | ucho | jisti | vejce | oko |
| DANISH | ore | at spise | aeg | oje |
| DUTCH | oor | eten | ei | oog |
| ENGLISH | ear | to eat | egg | eye |
| ESPERANTO | orelo | mangi | ovo | okulo |
| FINNISH | korva | syoda | muna | silma |
| FRENCH | oreille | manger | oeuf | oeil |
| GERMAN | ohr | essen | ei | auge |
| GER. BAVARIAN | oa-waschln | essn | oar | augn |
| GER. SWISS D. 1 | ohr | aesse | ei | oug |
| GER. SWISS D. 2 | ohr | aesse | ei | oug |
| GREEK NEW | afti | troo | avgho | mati |
| GREEK OLD | us | trogein | oon | blemma |
| HEBREW | ozen | leechol | beytsah | a'yin |
| HINDI | kan | khana | anda | ankh |
| HUNGARIAN | ful | eszik | tojas | szem |
| INDONESIAN | telinga | makan | telur | mata |
| ITALIAN | orecchio | mangiare | uovo | occhio |
| IT. N. LOMBARDY | urecia | pacha' | o:v | o:ch |
| IT. VENETII D. | recia | magnar | ovo | ocio |
| IRISH | cluas | ithim | ubh | suil |
| JAPANESE | mimi | taberu | tamago | me |
| KANNADA | kivi | tinnu | tatti | kannu |
| LATIN | auris | edere | ovum | oculus |
| LATVIAN | ausis | est | ola | acis |
| LITHUANIAN | auses | valgit | kiesinis | akys |
| MACEDONIAN | uvo | jade | jajce | oko |

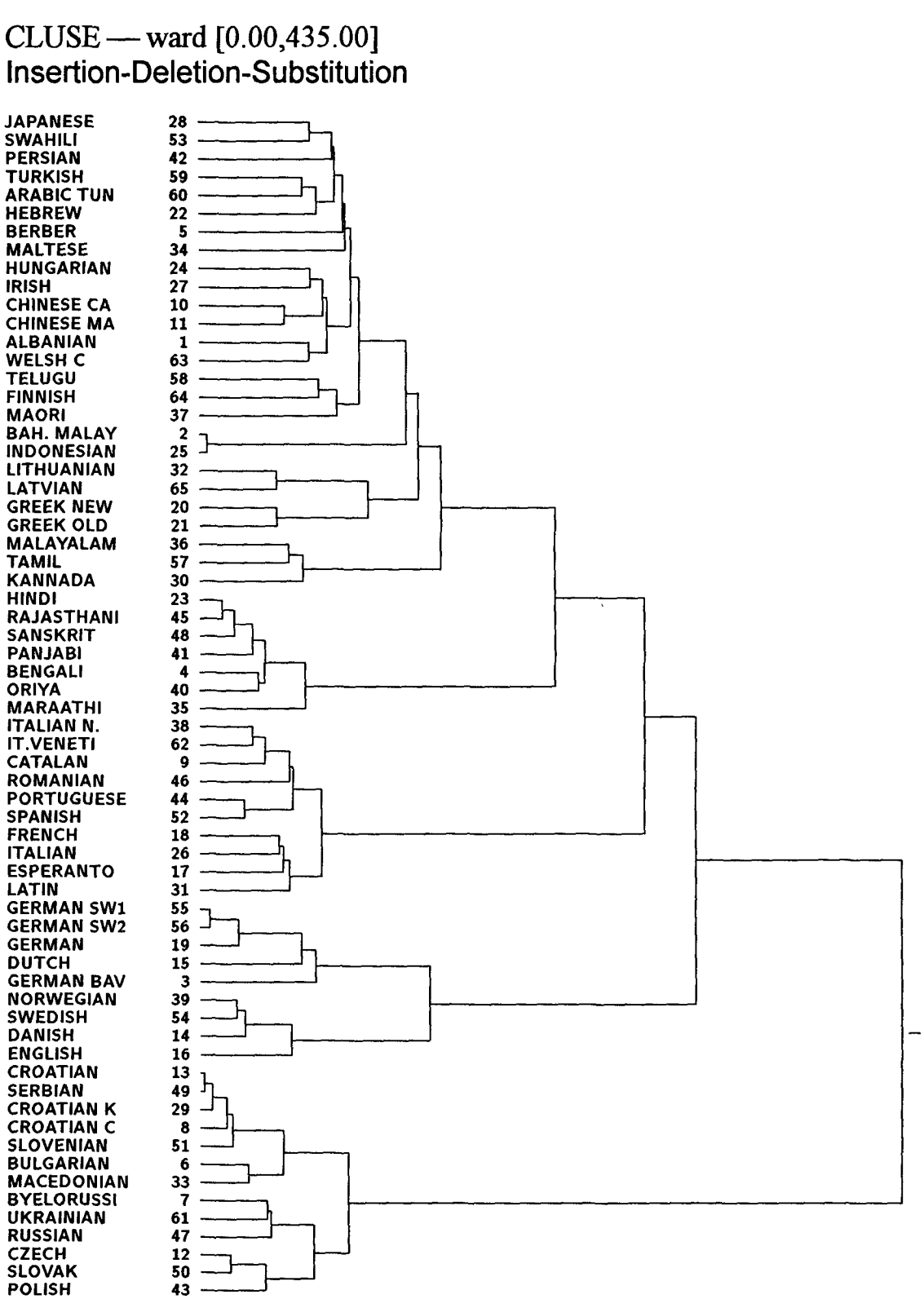
| | | | | |
|---|---|---|---|---|
| MALAYALAM | chhevy | thinnuka | mutta | kannu |
| MALTESE | widna | kiel | bajda | gtrajn |
| MAORI | pokoraringa | haupa | heeki | kaikamo |
| MARAATHI | kaan | khaney | undey | dohlaa |
| NORWEGIAN | oere | aa spise | egg | oeye |
| ORIYA | kana | khaiba | anda | akhee |
| PANJABI | kan | khana | anda | akh |
| PERSIAN | gush | khordan | tokhm | chashm |
| POLISH | ucho | jesc | jajko | oko |
| PORTUGUESE | orelha | comer | ovo | olho |
| RAJASTHANI | kon | khano | ando | onkh |
| ROMANIAN | orechie | a minca | ou | ochi |
| RUSSIAN | ukho | jest | jajtso | glaz |
| SANSKRIT | kaan | khana | anda | aankh |
| SERBIAN | uho | jesti | jaje | oko |
| SLOVAK | ucho | jest | vajce | oko |
| SLOVENIAN | uho | jesti | jajce | oko |
| SPANISH | oreja | comer | huevo | ojo |
| SWAHILI | sikio | la | yai | jicho |
| SWEDISH | oera | att aeta | aegg | oega |
| TAMIL | kaathu | saapiduthal | muttai | kann |
| TELUGU | chevi | thinadam | kuddu | kallu |
| TURKISH | kulak | yemek | yumurta | goz |
| UKRAINIAN | ukho | yiste | jajtse | oko |
| WELSH C | clust | bwyta | wy | llygad |

| | 13. | 14. | 15. | 16. |
|---|---|---|---|---|
| ALBANIAN | ate | peshk | pese | kembe |
| AR. TUNISIAN | baba | semica | xamsa | sak |
| BAH. MALAYSIA | ayah | ikan | lima | kaki |
| BENGALI | baba | mach | panch | pa |
| BERBER | vava | ahithiw | khamsa | akajar |
| BULGARIAN | otec | riba | pet | noga |
| BYELORUSSIAN | bats'ka | ryba | pyats | naga |
| CATALAN | pare | peix | cinc | peu |
| CH. CANTONESE | ba | yu | ng | geuk |
| CH. MANDARIN | fu qin | yu | wu | jiao |
| CROATIAN | otac | riba | pet | stopalo |
| CROAT. CAKAVSKI | otac | riba | pet | taban |
| CROAT. KAJKAVSKI | oca | riba | pet | stopalo |
| CZECH | otec | ryba | pet | noha |
| DANISH | fader | fisk | fem | fod |
| DUTCH | vader | vuur | vijf | voet |
| ENGLISH | father | fish | five | foot |
| ESPERANTO | patro | fiso | kvin | piedo |
| FINNISH | isa | kala | viisi | jalka |

| | | | | |
|---|---|---|---|---|
| FRENCH | pere | poisson | cinq | pied |
| GERMAN | vater | fisch | fuenf | fuss |
| GER. BAVARIAN | fadda | fiesch | fimfe | fuass |
| GER. SWISS D. 1 | fatter | fisch | fuef | fuess |
| GER. SWISS D. 2 | fatter | fisch | fuef | fuess |
| GREEK NEW | pateras | psari | pende | podhi |
| GREEK OLD | pater | opsarion | pente | pus |
| HEBREW | aba | dag | chamesh | regel |
| HINDI | bap | machli | panch | paer |
| HUNGARIAN | atya | hal | ot | lab |
| INDONESIAN | ayah | ikan | lima | kaki |
| ITALIAN | padre | pesce | cinque | piede |
| IT. N. LOMBARDY | pader | pe's | chinq | pe |
| IT. VENETII D. | pare | pes | zinque | pie |
| IRISH | athair | iasc | cuigear | cos |
| JAPANESE | chichi | sakana | go | ashi |
| KANNADA | appa | meena | aidu | paad |
| LATIN | pater | piscis | quinque | pes |
| LATVIAN | tevs | zivis | pieci | kaja |
| LITHUANIAN | tevas | zuves | penke | koja |
| MACEDONIAN | tatko | riba | pet | stapalo |
| MALAYALAM | acchan | meen | anju | kallu |
| MALTESE | missier | trut | transa | sieq |
| MAORI | paapara | ika | rima | wae |
| MARAATHI | wa-dil | maasaa | paach | paaool |
| NORWEGIAN | far | fisk | fem | fot |
| ORIYA | bapa | machchha | pancha | pada |
| PANJABI | bapa | ikan | lima | kaki |
| PERSIAN | pedar | mahi | panz | pa |
| POLISH | ojciec | ryba | piec | stopa |
| PORTUGUESE | pai | peixe | cinco | pe |
| RAJASTHANI | baap | machli | ponch | pug |
| ROMANIAN | tata | peste | cinci | picior |
| RUSSIAN | otjec | riba | pjat | noga |
| SANSKRIT | baap | machli | paanch | pea'r |
| SERBIAN | otac | riba | pet | stopalo |
| SLOVAK | otec | ryba | pet | noha |
| SLOVENIAN | oce | riba | pet | noga |
| SPANISH | padre | pez | cinco | pie |
| SWAHILI | baba | samaki | tano | mguu |
| SWEDISH | fader | fisk | fem | fot |
| TAMIL | appaa | meen | ainthu | kaal |
| TELUGU | nanna | chapa | ayithu | kalu |
| TURKISH | baba | balik | bes | ayak |
| UKRAINIAN | bat'ko | rihba | pyat | noha |
| WELSH C | tad | pisgodyn | pump | troed |

## Appendix B. Clustering Results




CLUSE — ward [0.00,435.00]
Insertion-Deletion-Substitution
JAPANESE 28
SWAHILI 53
PERSIAN 42
TURKISH 59
ARABIC TUN 60
HEBREW 22
BERBER 5
MALTESE 34
HUNGARIAN 24
IRISH 27
CHINESE CA 10
CHINESE MA 11
ALBANIAN 1
WELSH C 63
TELUGU 58
FINNISH 64
MAORI 37
BAH. MALAY 2
INDONESIAN 25
LITHUANIAN 32
LATVIAN 65
GREEK NEW 20
GREEK OLD 21
MALAYALAM 36
TAMIL 57
KANNADA 30
HINDI 23
RAJASTHANI 45
SANSKRIT 48
PANJABI 41
BENGALI 4
ORIYA 40
MARAATHI 35
ITALIAN N. 38
IT.VENETI 62
CATALAN 9
ROMANIAN 46
PORTUGUESE 44
SPANISH 52
FRENCH 18
ITALIAN 26
ESPERANTO 17
LATIN 31
GERMAN SW1 55
GERMAN SW2 56
GERMAN 19
DUTCH 15
GERMAN BAV 3
NORWEGIAN 39
SWEDISH 54
DANISH 14
ENGLISH 16
CROATIAN 13
SERBIAN 49
CROATIAN K 29
CROATIAN C 8
SLOVENIAN 51
BULGARIAN 6
MACEDONIAN 33
BYELORUSSI 7
UKRAINIAN 61
RUSSIAN 47
CZECH 12
SLOVAK 50
POLISH 43

CLUSE — ward [0.00,420.00]
LSCS - Length of their Shortest Common Supersequence