Authors: Thomas Mayer[1], Christian Rohrdantz[2], Frans Plank[1], Miriam Butt[1], Daniel A. Keim[2]
University/Affiliation: Department of Linguistics[1], Department of Computer Science[2] (University of Konstanz)
Email addresses: {thomas.mayer,christian.rohrdantz}@uni-konstanz.de

A Quantitative Approach to the Contrast and Stability of Sounds

The phonology of a language (like other components of grammar) undergoes change in the course of time. Languages differ as to which phonemic contrasts are made and also which changes their phonological system undergoes. How intimately related these two aspects of a phonological system are remains an open question, though usually some relationship is expected. In this paper we look at two aspects of phonological systems and language change: 1) the relative stability of phonemes as presumed indirectly from a cross-linguistic study of cognates; 2) expressions of phonemic contrast with respect to a cross-linguistic study of word forms which differ in only one sound.

Relative stability of sounds

Ideally, we should be using diachronic data in order to directly investigate the historical stability of phonemes. However, due to the lack of suitable diachronic corpora for a cross-linguistically representative set of languages, we decided to experiment with assessing the stability of sounds indirectly by basing our work on synchronic data that is available for a wide range of languages. In particular, we decided to use the data collected as part of the ASJP (Automated Similarity Judgment Program) database (version 12, Wichmann et al., 2010), since the database includes data on a wide range of languages and in a phonetically transcribed form.

Generally, historical linguists have tacitly assumed consonants to be more reliable/stable than vowels in the search for cognates as the basis for reconstructing sound changes (Campbell, 2004; Wälchli, 2010). But can it be quantitatively defended that vowels are generally less stable than consonants? And is there a general stability cline in the sounds of the languages (either for individual families or universal)? In addressing these questions, we experimented with automatically comparing items in related languages. Since the vocabulary items in the Swadesh list are expected to be culturally neutral and stable over time, areal influence is kept to a minimum and diachronic conclusions are potentially justified. We further make a simplifying assumption that the same Swadesh item in related languages is a cognate. This is not true for all items (e.g., English *tree* and German *Baum* are not cognates, yet fill the corresponding slot in the Swadesh list), but across languages in our approach cases like this can be considered to be noise in the data.

Our experiments show that setting up genealogical relationships with synchronic data on Swadesh list items yields reasonably accurate results when comparing a restricted set of languages. In Figure 1, for example, an automatically created neighbor net based on the Levenshtein distance of corresponding Swadesh items groups languages in accordance with expert classifications (see also Brown et al., 2008 for similar results). So despite of the sparse data available for individual languages we assume that interesting conclusions can be drawn when comparing languages within language families.

In order to investigate the historical stability of sounds via automatic methods, we compared each Swadesh item in the ASJP database for all languages within a language family with its corresponding item. For each comparison we counted the substitutions that are required for each word pair with respect to its Levenshtein distance. This gives us an approximation of the sound changes that might have taken place. It is a only a rough approximation because the method assumes that one of the sounds must have been in the respective form of the proto-language from which the other sound diverted. The identified substitutions were then statistically analyzed as to their association strength. For this purpose we used the ϕ value (the normalized χ^2 value, see Manning and Schütze, 1999) in order to be able to compare language families with differing number of pairs. The direction of the sound change cannot be determined with synchronic data, therefore all substitutions must necessarily be considered to be sound correspondences rather than changes. Table 1 shows the top and bottom consonant correspondences for the Germanic and Romance language families, respectively. It is easy to find examples for the top sound correspondences in the languages of the respective families (e.g., English ***s****tone* [**s**təʊn] vs. German ***S****tein* [**ʃ**taɪn]).

If we grant that synchronic comparison can inform us about historical factors, then our data indicate that the top ranked sounds should be historically less stable than ones that are lower ranked. How these values can be related to the stability of vowels vs. consonants is currently a matter of on-going investigation.

Cross-linguistic study of sound contrast

Our initial results show that an automatic analysis of cognates across languages is successfully able to identify language relatedness and provides information about which sounds are most likely to be changed. One factor in sound change is the expression and/or preservation of phonemic contrast in a language. We therefore looked at word forms which differ in only one sound across languages in the ASJP database to see whether one could automatically identify patterns among sounds based on their distribution across maximally large contexts on the word level. The substitutions in these cross-linguistic minimal pairs have been counted and statistically analyzed with the ϕ value (see above).

Some results are shown in Figures 2 and 3. Figure 2 shows that vowels form a group that can be differentiated very clearly from consonants. This is to be expected, since vowels should mainly be contrasting with one another. However, the results in Figure 3

Figure 1: Neighbor net of all Germanic languages in the ASJP database on the basis of their Levenshtein distance (created with SplitsTree, cf. Huson and Bryant, 2006).

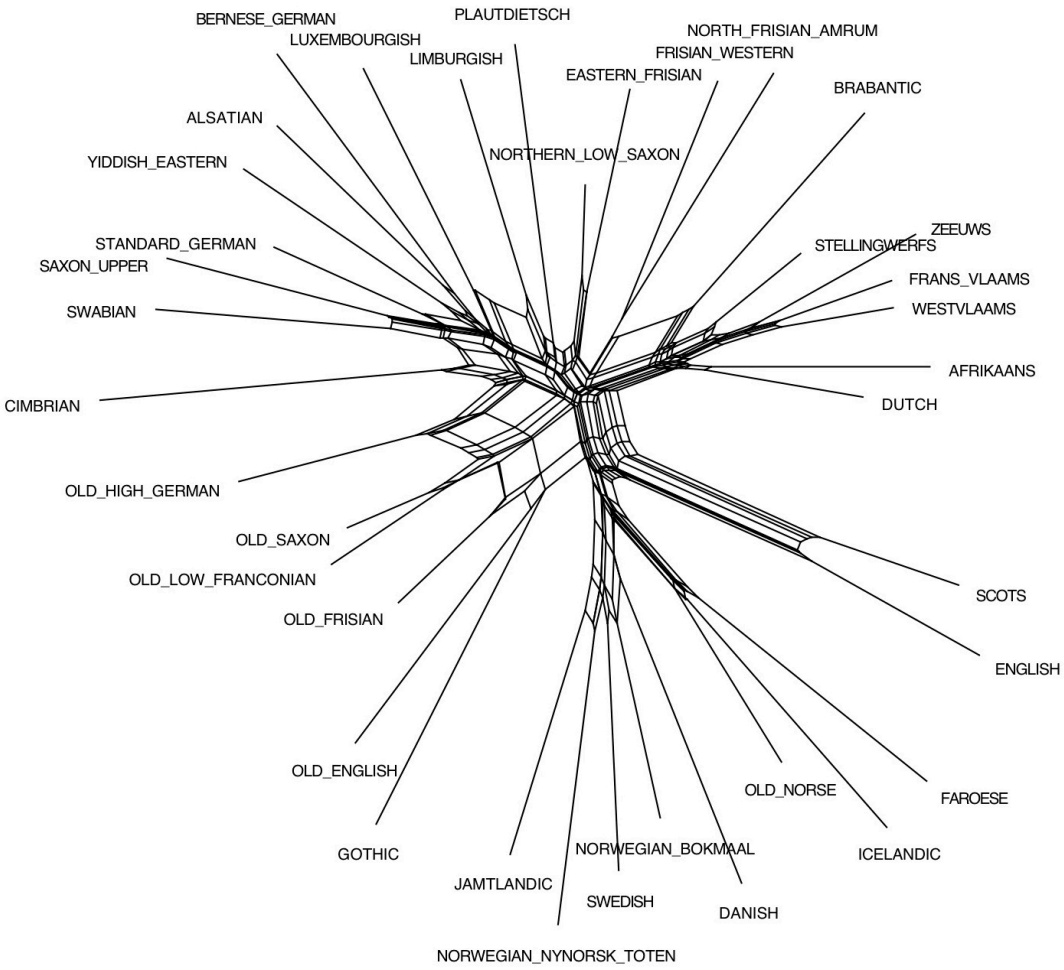


Table 1: Top and bottom sound correspondences for Germanic (left) and Romance (right) languages in the ASJP database. If symbols in the ASJP orthography represent more than one sound, all corresponding IPA symbols are listed. The sound correspondences have been sorted according to their signed ϕ values.

rank	corresponding sounds	ϕ value
1.	[ʃ].[s]	0.4511388
2.	[t].[d]	0.4202489
3.	[t].[ts,dz]	0.3814132
4.	[v].[w]	0.3765362
5.	[b,β].[p,ɸ]	0.3727747
6.	[s].[z]	0.3529081
7.	[ɯ,u].[ɤ,ʌ,ɑ,ɒ,o,ɔ]	0.2538024
8.	[k].[x,ɣ]	0.2249017
⋮	⋮	⋮
660.	[t].[i,ɪ,y,ʏ]	-0.04844407
661.	[s].[ɯ,u]	-0.04988262
662.	[s].[ɤ,ʌ,ɑ,ɒ,o,ɔ]	-0.05132239
663.	[t].[a,æ,ɛ,œ]	-0.05507378
664.	[s].[ɐ]	-0.05950131
665.	[t].[ɤ,ʌ,ɑ,ɒ,o,ɔ]	-0.06332589
666.	[t].[ɐ]	-0.06932040

rank	corresponding sounds	ϕ value
1.	[v].[b,β]	0.4192410
2.	[ɯ,u].[ɤ,ʌ,ɑ,ɒ,o,ɔ]	0.4096724
3.	[ɨ,ɘ,ə,ɜ,ʉ,ɵ,ɞ].[ɐ]	0.2465891
4.	[n].[ɲ]	0.2305777
5.	[ʃ].[s]	0.2179642
6.	[L,ɭ,ʎ̝].[h,ɦ]	0.1868974
7.	[k].[ʒ]	0.1449062
8.	[g].[x,ɣ]	0.1354777
⋮	⋮	⋮
555.	[e,ø].[.g]	-0.04514492
556.	[ɯ,u].[a,æ,ɛ,œ]	-0.04949300
557.	[i,ɪ,y,ʏ].[ɤ,ʌ,ɑ,ɒ,o,ɔ]	-0.05676500
558.	[ɤ,ʌ,ɑ,ɒ,o,ɔ].[a,æ,ɛ,œ]	-0.05838840
559.	[ɯ,u].[ɐ]	-0.06042425
560.	[e,ø].[ɯ,u]	-0.06334957
561.	[ɤ,ʌ,ɑ,ɒ,o,ɔ].[ɐ]	-0.06435476

are unexpected. Figure 3 focuses on just the consonant patterns and removes the main effect of the vowels. Once this main effect is removed, a clear pattern with respect to the consonants emerges. The consonants appear to fall into two major groups. This division is unexpected as we do not see it following from distinctions established so far in the phonological literature.

The results presented here are part of a larger on-going effort to introduce methods from visual analytics (Thomas and Cook 2005; Keim et al. 2008) into quantitative linguistic analyses. In this paper, we show that the automatic examination of sound patterns across languages can be used to further our understanding of sound change (phoneme stability) and phonemic distinctions. In particular, the results in Figure 3 have brought to light a new linguistic pattern which can now be explored further in terms of a fruitful interaction between theoretical and quantitative approaches.

Table 2: ASJP orthography (cf. Brown et al., 2008)

ASJP symbol	IPA symbol(s)
i	[i,ɪ,y,ʏ]
e	[e,ø]
E	[a,æ,ɛ,œ]
3	[ɨ,ɘ,ə,ɜ,ʉ,ɵ,ɞ]
a	[ɐ]
u	[ɯ,u]
o	[ɤ,ʌ,ɑ,ɒ,o,ɔ]
p	[p,ɸ]
b	[b,β]
m	[m]
f	[f]
v	[v]
8	[θ,ð]
4	[n̪]
t	[t]
d	[d]
s	[s]
z	[z]
c	[ts,dz]
n	[n]

ASJP symbol	IPA symbol(s)
S	[ʃ]
Z	[ʒ]
C	[ʧ]
j	[ʤ]
T	[c,ɟ]
5	[ɲ]
k	[k]
g	[g]
x	[x,ɣ]
N	[ŋ]
q	[q]
G	[G]
X	[χ,ʁ,ħ,ʕ]
7	[ʔ]
h	[h,ɦ]
l	[l]
L	[L,ɭ,ʎ]
w	[w]
y	[j]
r	[r,R,etc.]

Figure 2: Sound correspondences within minimal pairs across languages. Vowels can be clearly differentiated from consonants. Rows and columns have been sorted automatically according to the similarity of the sounds. For the symbols see Table 2.

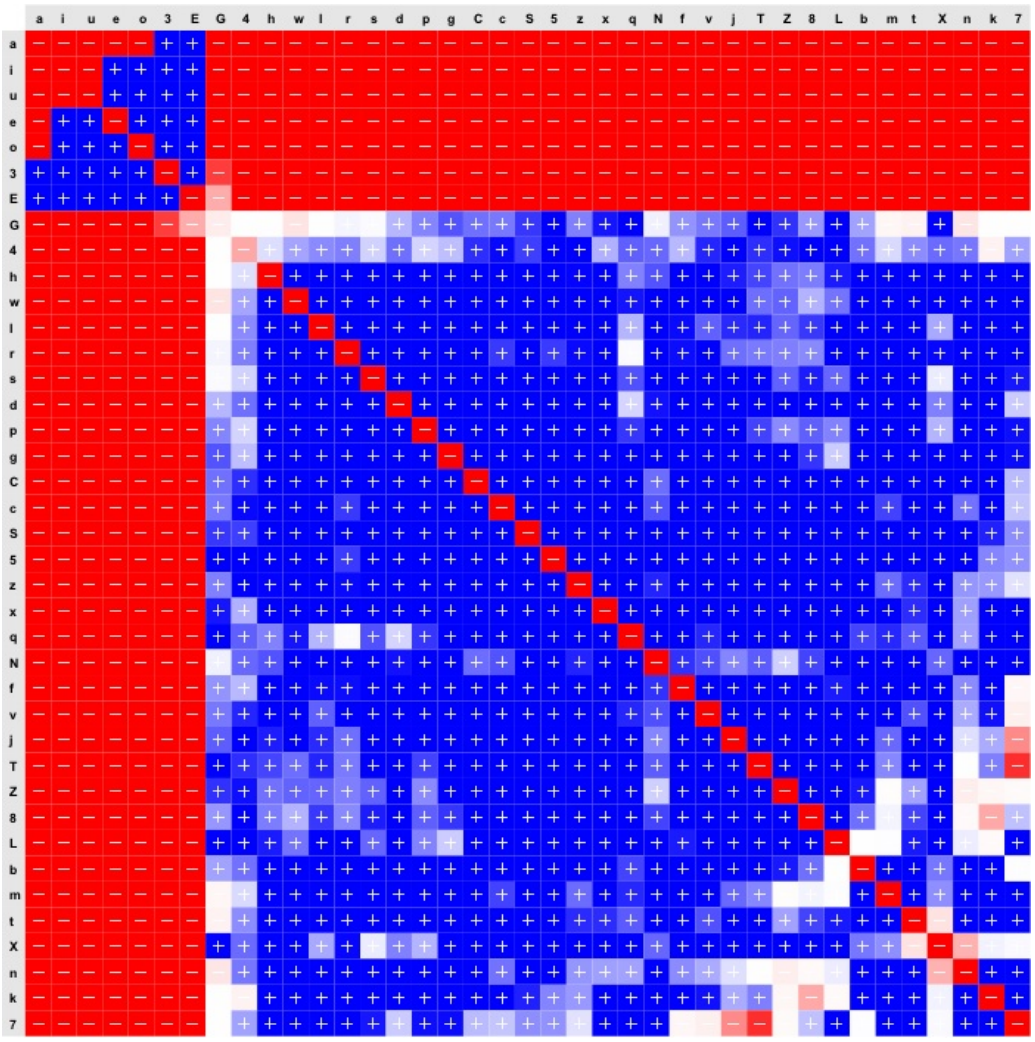
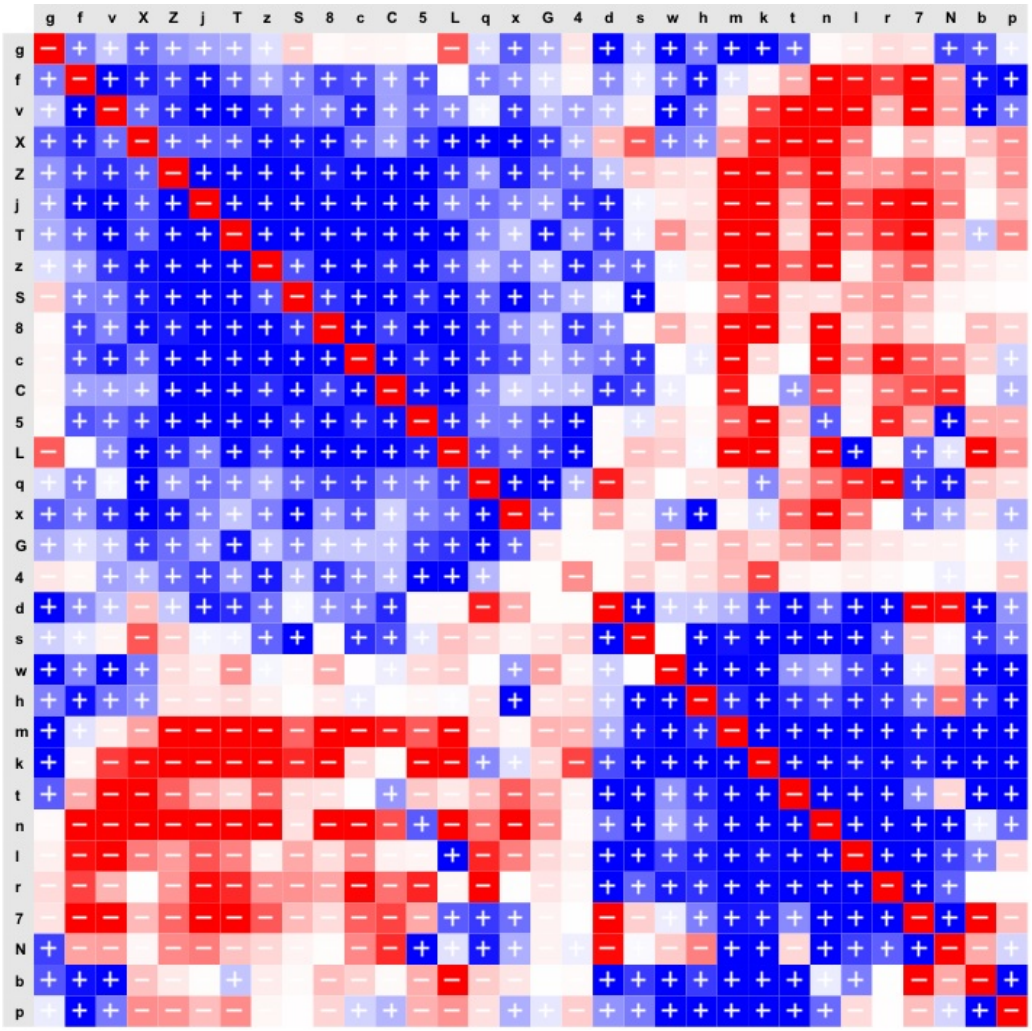


Figure 3: Sound correspondences within minimal pairs across languages (consonants only). The consonants seem to fall into two major groups (left top and bottom right corner). Rows and columns have been sorted automatically according to the similarity of the sounds. For the symbols see Table 2.



References

Brown, Cecil H., Eric W. Holman, Søren Wichmann, and Viveka Velupillai. 2008. Automated classification of the world's languages: a description of the method and preliminary results. *STUF Language Typology and Universals* 61:285–308.

Campbell, Lyle. 2004. *Historical Linguistics*. Edinburgh: Edinburgh University Press.

Huson, Daniel H., and David Bryant. 2006. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* 23:254–267.

Keim, Daniel A., Florian Mansmann, Joern Schneidewind, Jim Thomas, and Hartmut Ziegler. 2008. Visual Analytics: Scope and Challenges. In *Visual Data Mining: Theory, Techniques and Tools for Visual Analytics*, Lecture Notes in Computer Science, 76–91. Springer.

Manning, Christopher D., and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Massachusetts Institute of Technology.

Thomas, James J., and Kristin A. Cook. 2005. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. National Visualization and Analytics Ctr.

Wälchli, Bernhard. 2010. The consonant template in synchrony and diachrony. *Baltic Linguistics* 1:137–166.

Wichmann, Søren, André Müller, Viveka Velupillai, Cecil H. Brown, Eric W. Holman, Pamela Brown, Matthias Urban, Sebastian Sauppe, Oleg Belyaev, Zarina Molochieva, Annkathrin Wett, Dik Bakker, Johann-Mattis List, Dmitry Egorov, Robert Mailhammer, and Helen Geyer. 2010. The ASJP Database (version 12). URL: `http://email.eva.mpg.de/ wichmann/ASJPHomePage.htm`.