PRESENTING THE ‘SOUND COMPARISONS’ WEBSITE

Paul Heggarty & Jakob Runge

Dept. of Linguistics
Max Planck Institute for Evolutionary Anthropology, Leipzig

heggarty@eva.mpg.de   jakob_runge@eva.mpg.de
1. **BACKGROUND**

• Research context:
  – *Measuring divergence in phonetics.*
  – Between *related* languages, dialects and accents.

• Major effort of:
  – Data collection: recordings — c. 120 words in c. 350 language varieties.
  – Data analysis: detailed phonetic transcription.

• Aim of *websites*: make those data and analyses available and useful to:
  – Scientific community of *linguists*, as a training and research resource.
  – *Native-speakers* of (esp. endangered) language varieties covered, for raising awareness, understanding, prestige, revitalisation (?).
1.1 Research Context

• **Input data** for a technique for quantifying divergence in phonetics (as precisely as possible).

• Determined data-set: phonetic sample → list of **cognates** (not meanings).

• Applications: in dialectology, historical linguistics, sociolinguistics.


1.2 **Divergence Measures: Single Cognate**

![Divergence Measure Table](image-url)
1.3 **DIVERGENCE MEASURES: ENTIRE REFERENCE LIST**

![Image of Excel table showing phonetic similarity ratings to accents of English]
1.4 **Visualisations: Divergence of English Through Space & Time**

- **Old English**
- **Middle English**
- **Modern Traditional Scotland**
- **Historical Scots**

= Middle and E.M. English Periods
1.5 **Phonetic Divergence Between Dialects in Germanic**
1.6 **Project Origins and Development**

- **2004-2007**
  - funding: Arts and Humanities Research Council, UK
  - Linguistics, University of Sheffield:
    *Quantitative Methods in Comparative Linguistics*
  - Linguistics, University of Edinburgh:
    *Sound Comparisons: Dialect and Language Comparison and Classification by Phonetic Similarity*

- **2006-2009**
  - funding: Leverhulme Trust, UK
  - *McDonald Institute for Archaeological Research, University of Cambridge:*
    *Languages and Origins in Europe*

- **2011-2015**
  - funding: Max-Planck-Gesellschaft
  - Linguistics, Max Planck Institute for Evolutionary Anthropology, Leipzig.
1.7 **Data and Earlier Websites**

- Language recordings collected since 2000, continuing whenever possible.
  
  - *Sounds of the Andean Languages*
    > www.quechua.org.uk/sounds
  
  - *Accents of English from Around the World*
    > www.soundcomparisons.com
  
  - Regional dialects and languages of *Germanic* (+ Romance, Balto-Slavic)
    > www.languagesandpeoples.com

- *Any language family* can easily be added, now that system is set up...
1.8 People

- Website: originally by Heggarty, but now completely recreated by Jakob Runge (Uni Leipzig).

- Phonetic transcriptions:
  - Andes: Heggarty, Scott Sadowsky (UFRO, Chile).

- Data collection:
  - English dialects: Warren Maguire.
  - All others: Paul Heggarty.

- Initial funding/direction: April McMahon (Aberystwyth).

- Hundreds of native-speakers!
1.9 **AIMS OF \textit{W}EB\textit{S}ITE**

- **Make use** of databases of recordings and phonetic transcriptions.
- Fundamental purpose: \textit{compare} pronunciations of ‘\textit{same}’ cognates.
- To serve two user groups together.
  - **Speakers** of language varieties concerned, i.e. general public.
    - Esp. for endangered language varieties.
    - Regional languages/‘dialects’ of main European families.
    - Indigenous languages of Peru, Bolivia, Ecuador: Quechua and Aymara.
    - Part of wider website to support literacy, through understanding and uptake of proposed standard orthography problems.
  - **Linguistics researchers.**
    - Make database valuable and searchable for their ends.
2. FEATURES

• User-friendly, accessible, no specialised linguistic knowledge needed:
  – Sound files.
  – Maps.

• Powerful research tool:
  – IPA transcriptions.
  – ‘Linguistically informed’ functions.

• User can tailor site to specific interests:
  – Select languages / words / sounds.
  – Select any combination of these.
2.1 **Languages Axis**

- Map view, includes:
  - Zooming on selected regions.
  - Only showing selected languages of interest.

- Add transcriptions of any known *historical varieties* of a language.

- Add (hypothesised) phonetics for a family’s *proto-language*.
  
  *e.g.* → Compare all modern reflexes of Proto-Germanic word-initial */t*/.
2.2 **Words Axis: Search/Filter Functions**

- By *spelling*, i.e. graphemes (or sequences)...
  - In any language variety that does have a standard orthography.

- By *sounds*, i.e. symbols in *IPA*...
  - In transcription of any language variety in database.
  - Including *IPA* diacritics, e.g. vowel length [ː].

- In both, ‘advanced search’ features:
  - Results *filtered* in real-time as search string is typed.
  - ‘Regular expressions’ to search for contexts, e.g. $S = \text{word-final} [ʃ]$, etc.

- Add family- or language-specific data to search by:
  
  *e.g.* Wells’ *(1982: 127-67)* “lexical sets” for English dialectology.

  *e.g.* Use upper case for: C, V, archiphonemes N, R, etc.
2.3 Combined Selections: Words and Languages

• Compare on one screen multiple selected words *and* languages.

  *e.g.* Numerals 1 to 10 — In all languages.
  *e.g.* All words for body parts — In all Scandinavian varieties.
  *e.g.* All words with `<r>` in English spelling — In all English varieties.
  *e.g.* All words that contain `[t]` in RP — In all English varieties.
  *e.g.* All words that had Proto-Germanic `[k]` — In all Continental Germanic.
2.4 **Website User Language: Multilingual Support**

- **'Outreach':** promote awareness and understanding of regional languages. → Make site available in such languages themselves.

- **Collaborative:** enter translations of site language remotely online. (Password protected.)
3. **Web Policies**

- Free, collaborative (site language translations), open to new families!
- Ensure that website functions in all browsers.
- Sound files available in two formats: .mp3 and .ogg.
- No static webpage at all: all pages *generated in real time*, ‘on the fly’.
- ‘Links’ and ‘addresses’ are just queries to underlying database.
- Words and languages *selected appear in address line*, so can be typed in to search/filter all pages previously visited.
- Linked data (‘semantic web’).
3.1 **Some Technical Data...**

- **Total size of programme:**
  - Only 6 MB of code (+ images + sound files).
  - 6129 lines of PHP.
  - 1038 lines of Javascript.
  - 506 lines of SQL.

- **Which technologies?**
  - PHP to generate the website on demand.
  - MySQL as database backend for PHP script.
  - Javascript for more powerful features and speed.

- **Any technical questions?**
  → Ask Jakob Runge.
3.2 ** Links to Other Resources on Languages Covered **

- Link to entries on same language varieties in:

- **Problems:**
  - In different site languages, names of languages to link to are different.
  - Use *ISO language codes* wherever possible.
  - Solution thanks to Lexvo and Sebastian Nordhoff (MPI-EVA).

- **Dialects/accents** very sporadically and inconsistently present, no *ISO* codes.
  - Some proposals available, otherwise need to create *ad hoc* links.
4. FUTURE PLANS

4.1 EXTEND EXISTING DATABASES

• Structure now established, no further programming needed.

• Can now extend coverage to:
  – More site languages.
  – More data languages within the families already covered.
  – More families / regions.
4.2 ADDING NEW FAMILIES

• For each new family, data required:
  – List of languages (by classification?/by region?), lat/long co-ordinates.
  – List of ‘pan-family’ cognates (or meanings) for that family.
  – Sound recordings.
  – Phonetic transcriptions (in Unicode fonts).
4.3 A NEW WEBSITE FOR THE INTERCONTINENTAL DICTIONARY SERIES

• Re-launch the *Intercontinental Dictionary Series*.
  – Begun by Mary Ritchie Key, 1960s.
  – Now managed by Linguistics Dept, MPI-EVA, Leipzig.
    [http://lingweb.eva.mpg.de/ids](http://lingweb.eva.mpg.de/ids)

• Also essentially comparative, but in **lexis**: list of meanings, not cognates.

• A much bigger list: **1450 meanings**, structured in semantic categories.

• As also used for: World Loanword Database: [http://wold.livingsources.org](http://wold.livingsources.org).

• A couple of hundred minority/endangered languages worldwide.

• Transcriptions to be updated to **IPA**.

• No original sound recordings, but now add recordings where possible.
4.4 **Feedback, Co-Operation?**

- Any feedback, suggestions on features?
- Interest in using our structure to showcase *your* data?
- Please let us know...
REFERENCES


