25th ICM: NUI Galway, 24–26 June 2011

Language ascription by use of character libraries

A Work in Progress

Rob Lee

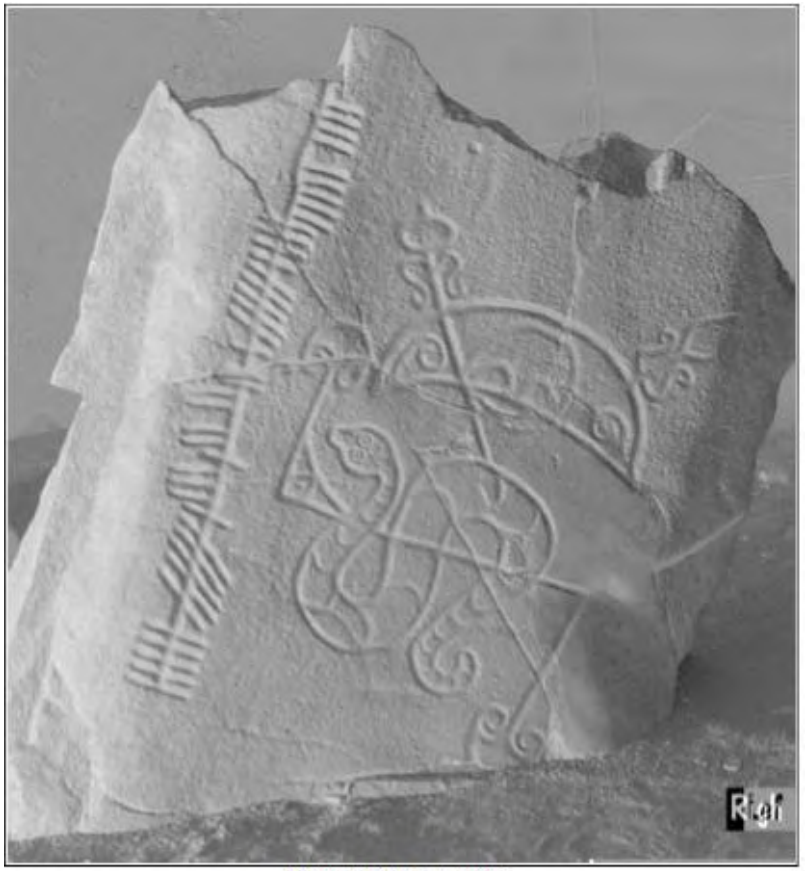
&

Phil Jonathan

The Problem

- Inscriptions on stone and portable objects from the Early Medieval period
  - For some it is hard to identify:
    - Language
    - Meaning
    - Hidden meaning

Pictish Ogam



Photograph by R. J. Henery

- E.g. - Brandsbutt ‘irataddoarens’ or ‘cqeraollavari’
- Language type
  - Celtic?
  - Nordic?
  - Other?
  - Mixture?

Irish Ogam

Image - Claire Library

- E.g. - Ennis amber bead

‘atucmlu’

- Meaning
  - Name?
  - Magical?

Early Scandinavian Runes

Image - Arild Hauge

- E.g. - Skonager
- ‘liRaiwui ildaituha’
- Language?
- Meaning?
- Random?

Welsh Latin

- E.g Llanerfyl

‘hic (in) tvmvloia …’

- Hidden Meaning?
  - Numerology?
  - Letter imagery?
  - C. Thomas

Image - RCAHMW

The Question

- Can statistics be used to:
  - Differentiate between language types?
  - Ascribe language type(s) to an unknown inscription?
  - Flag inscriptions with hidden meaning?
  - Flag inscriptions that are ‘random’ jottings

The Proposal

- Similar to process used by linguists and historians to qualitatively identify units or words within an inscription from a language lexicon
- For each language build a library of texts, each text of fixed size (e.g. 7000 letters using the 26 letter alphabet)
- Define all the letter groupings from 2-12 letters (N-grams) to be found in each text to give a lexicon of N-grams for that text
- Compare inscription N-grams with the lexicons found in the different language libraries for a match
  - the process will be quantitative rather than qualitative

Example Library Text - UDHR

- UniversalDeclarationofHumanRightsPreambleWhereasrecognitionoftheinherentdignityandoftheequalandinalienablerightsofallmembersofthehumanfamilyisthefoundationoffreedomjusticeandpeaceintheworldWhereasdisregardandcontemptforhumanrightshaveresultedinbarbarousactswhichhaveoutragedtheconscienceofmankindandtheadventofaworldinwhichhumanbeingsshallenjoyfreedomofspeechandbeliefandfreedomfromfearandwanthasbeenproclaimedasthehighestaspirationofthecommonpeopleWhereasitisessentialifmanisnottobecompelledtohaverecourseasalastresorttorebellionagainsttyrannyandoppressionthathumanrightsshouldbeprotectedbytheruleoflawWhereasitisessentialtopromotethedevelopmentoffriendlyrelationsbetweennationsWhereasthepeoplesoftheUnitedNationshaveintheCharterreaffirmedtheirfaithinfundamentalhumanrightsinthedignityandworthofthehumanpersonandintheequalrightsofmenandwomenandhavedeterminedtopromotesocialprogressandbetterstandar
- And so on for another 8000 characters!
  - Nothing but letters c.f. inscriptions
- Many other texts used to generate the libraries - Modern English, Finnish, Polish, Welsh, Irish, Norse, Latin, Icelandic
- Libraries for older languages also generated - Middle English, Middle Irish, Old English, Old Irish, Old Norse

N-grams

- N-grams are letter groups found in texts

e.g. ‘thecatsat’

2-grams = th,he,ec,ca,at,ts,sa,at,

3-grams = the, hec, eca, cat, ats, tsa, sat

Etc

Texts are split into their component N-grams (N=2-12) and stored as a lexicon.

Inscriptions are split into their component N-grams and compared against the lexicons

Does the process differentiate languages?

- YES!
- Okay - How?

How does it differentiate languages?

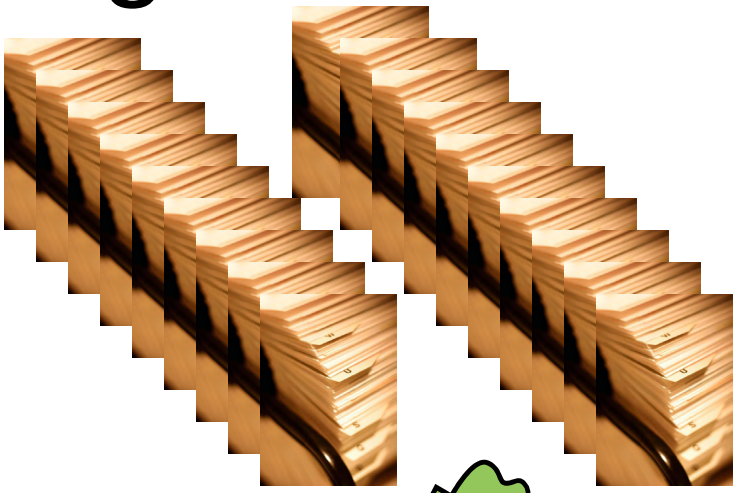
- Take the UDHR for 140+ languages

How does it differentiate languages?

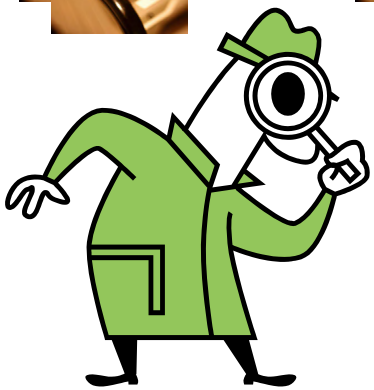
- Take the UDHR for 140+ languages
- Create N-gram lexicon using the first 7000 letters of UDHR for each language

How does it differentiate languages?

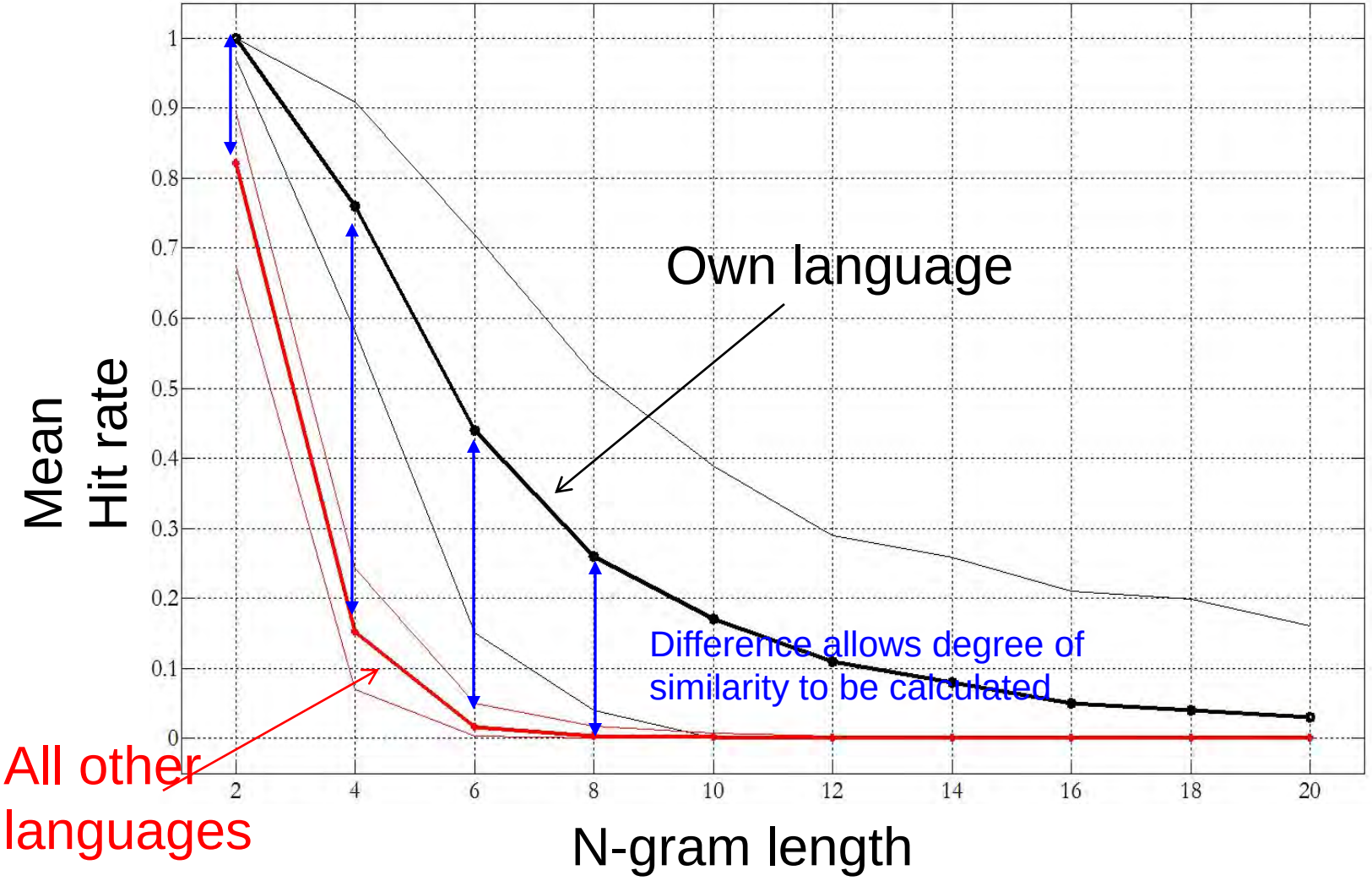
- Take the N-gram lexicon for each language



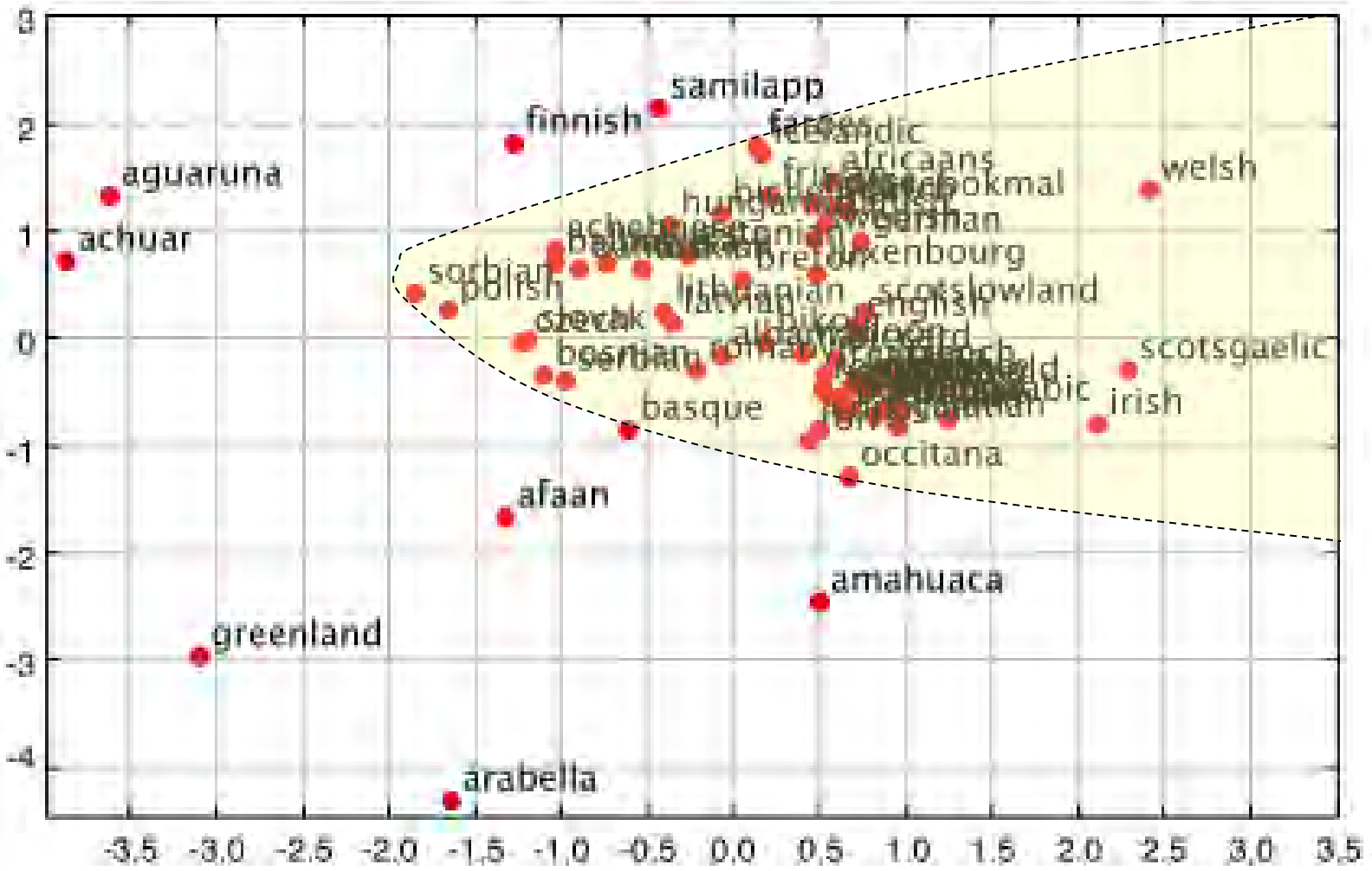
- Compare with a smaller (1000 letter) N-gram lexicon from remainder of UDHR



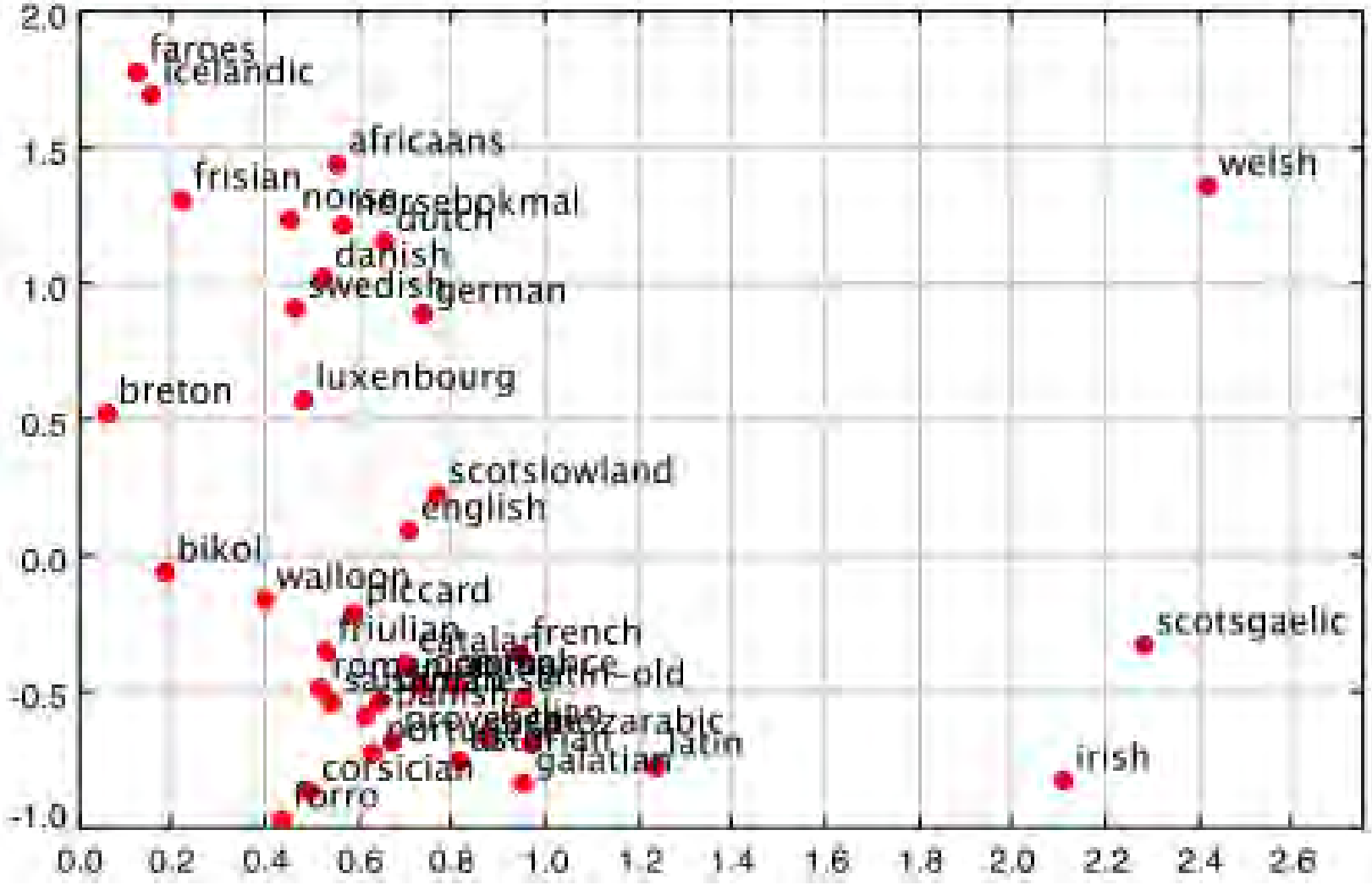
N-gram hit rate against own language & other languages, 7000 letter library using UDHR



Similarity of Languages based on UDHR



Similarity of Languages based on UDHR



What Next?

- Build large libraries of text (7000 letters) in a smaller number of languages of interest (English, Finnish, Polish, Welsh, Irish, Norse, Latin, Icelandic, Middle English, Middle Irish, Old English, Old Irish, Old Norse) & establish statistical validity.
- Repeat with larger texts (e.g. 30,000 letters)
- Trial with Scandinavian inscriptions - can we differentiate between readable and non-readable inscriptions?

- Invite input for areas or topics from you, the listener …