SuGaLi

Susanne Fertmann
Guy Emerson
Liling Tan

Presentation Outline

- Task
- Difficulties
- Contribution

Task

Language Identification for Low-Resource Languages

Task

ᐃᓄᓗᒃᑖᑦ ᐃᓅᓕᓵᙳᖅᐳᑦ ᓇᖕᒥᓂᓇᖕᒥᓂᓯᒪᖅᑲᖅᖁᑎᒃ ᐊᔾᔨᖕᒥᓪᓗ ᐃᓕᑕᕆᔭᐅᔾᔪᑦᓯᐊᖅᑲᖅᖁᑎᓪᓗ ᐱᔪᖕᓇᖅᐅᑎᑕᐅᖅᑲᖅᖁᑎᒃ. ᐃᓱᒪᒃᓴᖅᓯᐅᕈᖕᓇᕐᑎᐊᕐᓂᕐᒥᒃ ᐃᓅᕐᑎᐊᕈᑎᒋᔭᕐᓗ ᐱᓕᖅᑐᖕᒐᐅᕐᑐᑦ, ᐊᓯᐊᙳᕐᑐᕐᓄᓪᓗ ᐃᓕᐅᕐᓂᕐᕕᕐᑲᕐᑲᑎᒌᕐᑕᕈᒃᓴᓇᐊᕐᑲᕋᓗᐊᖅᐳᑦ ᕐᑲᑕᖕᒍᑎᒌᖅᑲᑎᒌᕐᑐᑦ ᐊᓂᕐᓂᖅᓵᕐᓂ.

(ᑎᑎᖅᑲᖅᓯᒪᔪᖅ 1 ᑭᑐᑐᐃᓐᓇᕐᓄᑦ ᐱᔪᖕᓇᖅᐅᑎᑦ ᓯᓚᕐᔪᐊᕐᒥᐅᕐᑲᑎᒌᓐᓄᑦ ᓇᓗᓇᐃᖅᓯᐅᑎᕐ)

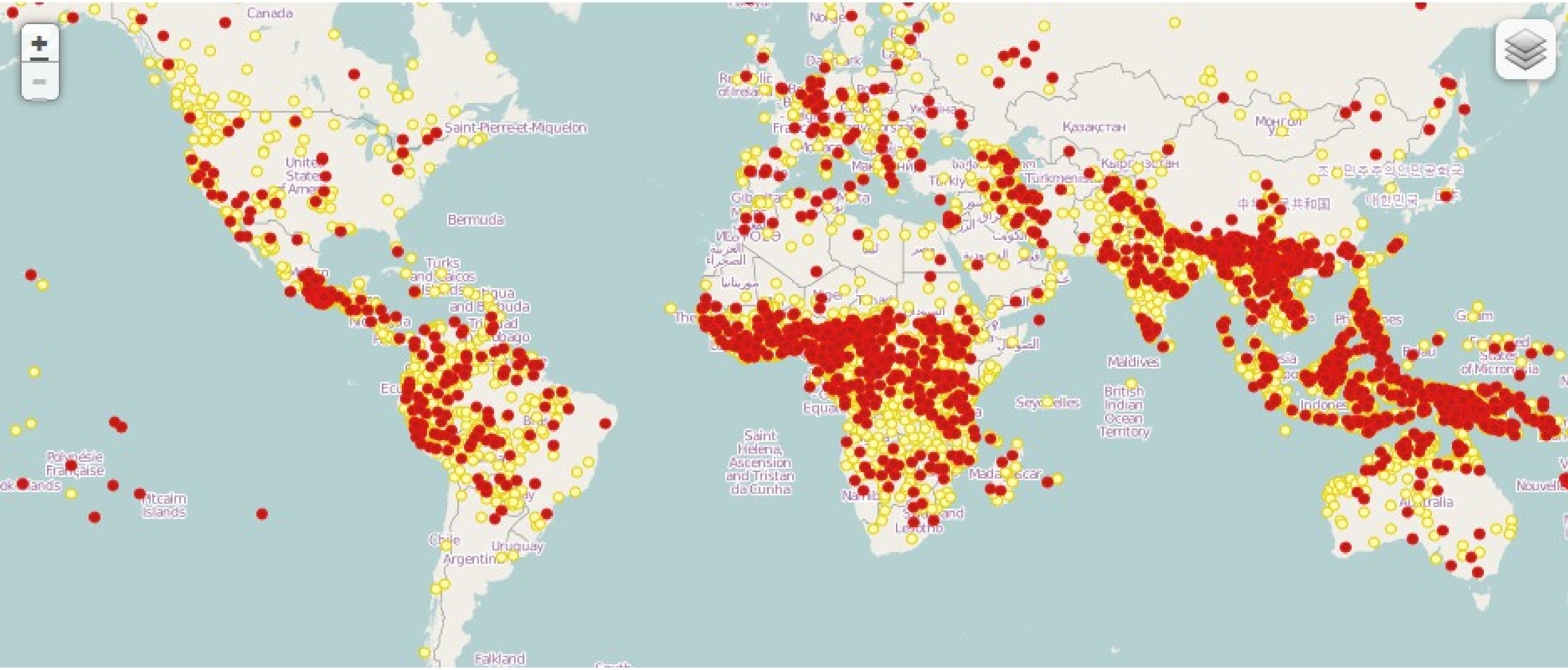
Task

Language Identification for
Low-Resource Languages
(text only)

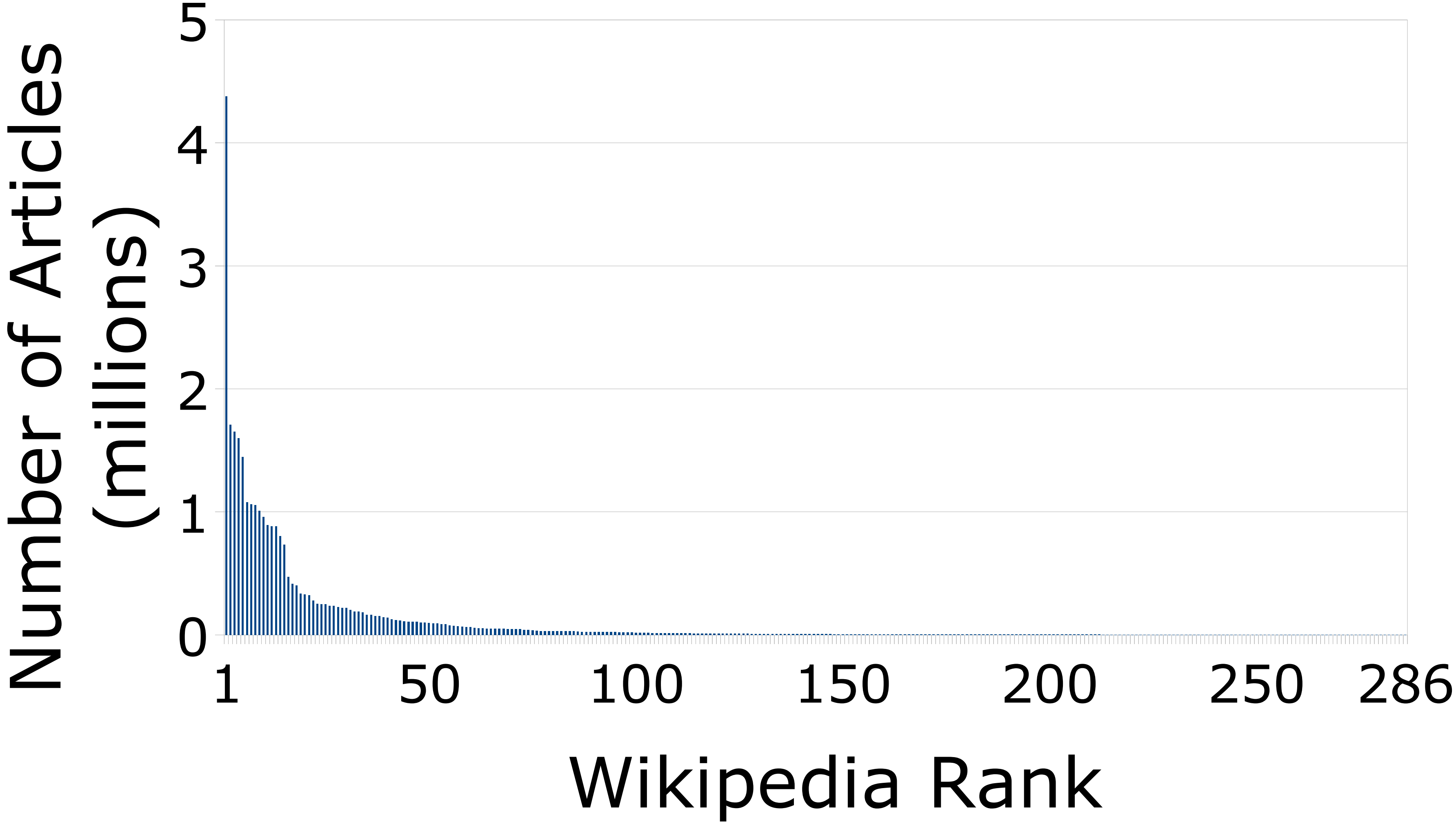
Task

Nou tou imen nou'n ne dan laliberte ek legalite, dan nou dignite ek nou bann drwa. Nou tou nou annan kapasite pou rezonnen, e fodre nou azir anver lezot avek en lespri fraternel.

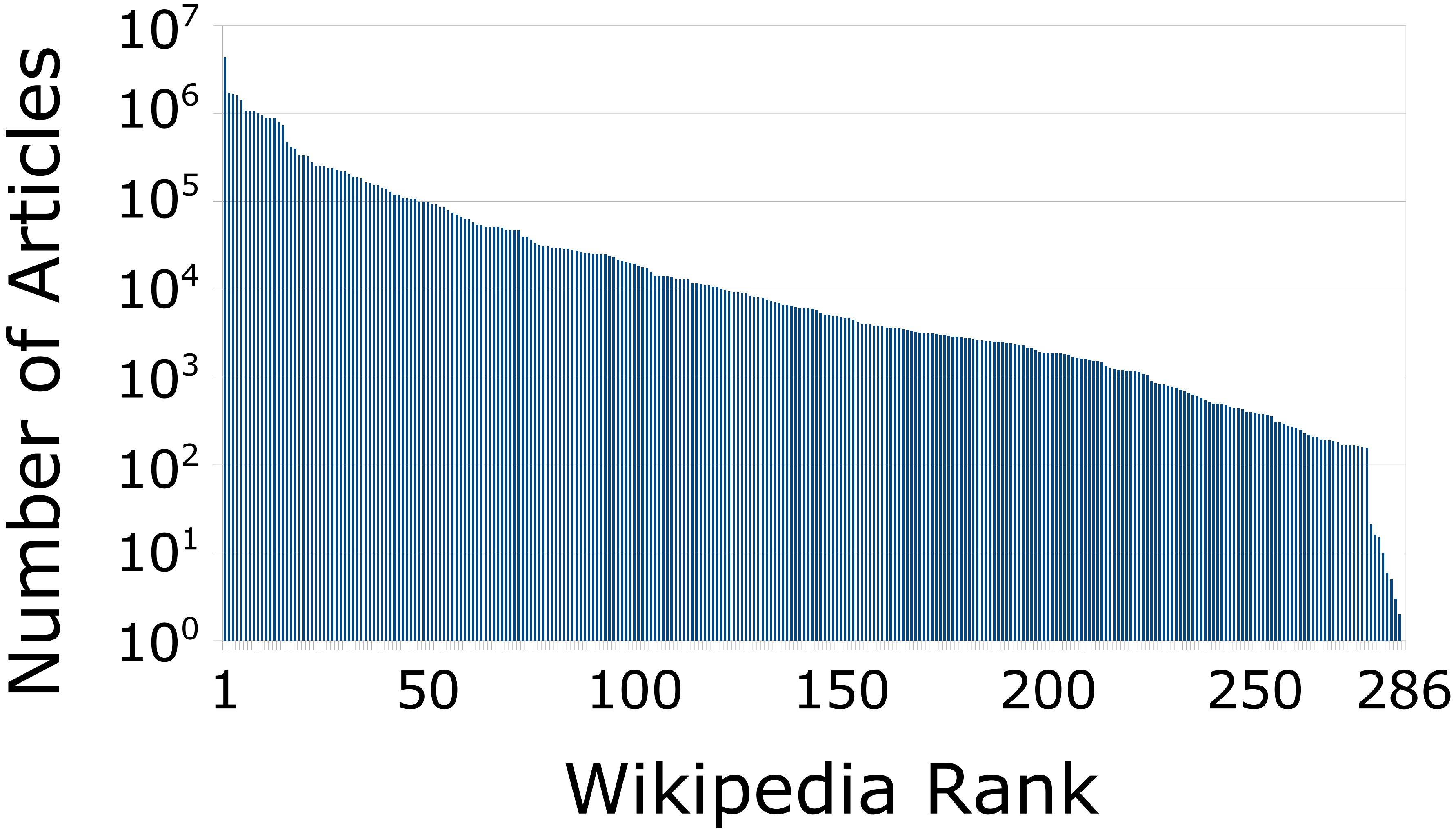
Difficulty: Scale



Difficulty: Balance



Difficulty: Balance



Difficulty: Encoding

ადამიანის უფლებათა
საყოველთაო დეკლარაცია

flfvbfybc eakt<fsf
cf>jdtksfj ltrkfhfwbf

Difficulty: Access

©

Difficulty: Format

```
&lt;!---------------------------------
SALUDU
-----------------------------------&gt;
{{Portada_saludu}}
&lt;!----------------------
CATEGORÍES
------------------------&gt;
{{Categoríes}}
&lt;!----------------------
A 2 columnes primeru cola izquierda
------------------------&gt;
&lt;div style=&quot;display:block;width:99%;float:left&quot;&gt;
&lt;div style=&quot;width:48%;display:block;float:left;&quot;&gt;
&lt;!----------------------
ARTÍCULU DESTACÁU
------------------------&gt;
{{Portada_artículu
|Artículu=Barcelona
|testu='''Barcelona''' ye una [[ciudá]] capital de [[Cataluña]] y de la [[provincia de Barcelona]]. La ciudá acueye les sedes de les instituciones d'autogobiernu más importantes de Cataluña: la Generalitat de Cataluña y el Parllamentu de Cataluña. Por ser capital del Condáu de Barcelona, recibe'l nomatu de ''Ciudá Condal'' (''Ciutat Condal'' en [[catalán]]).
```

Difficulty: Variation

> “A language is a dialect with an army”

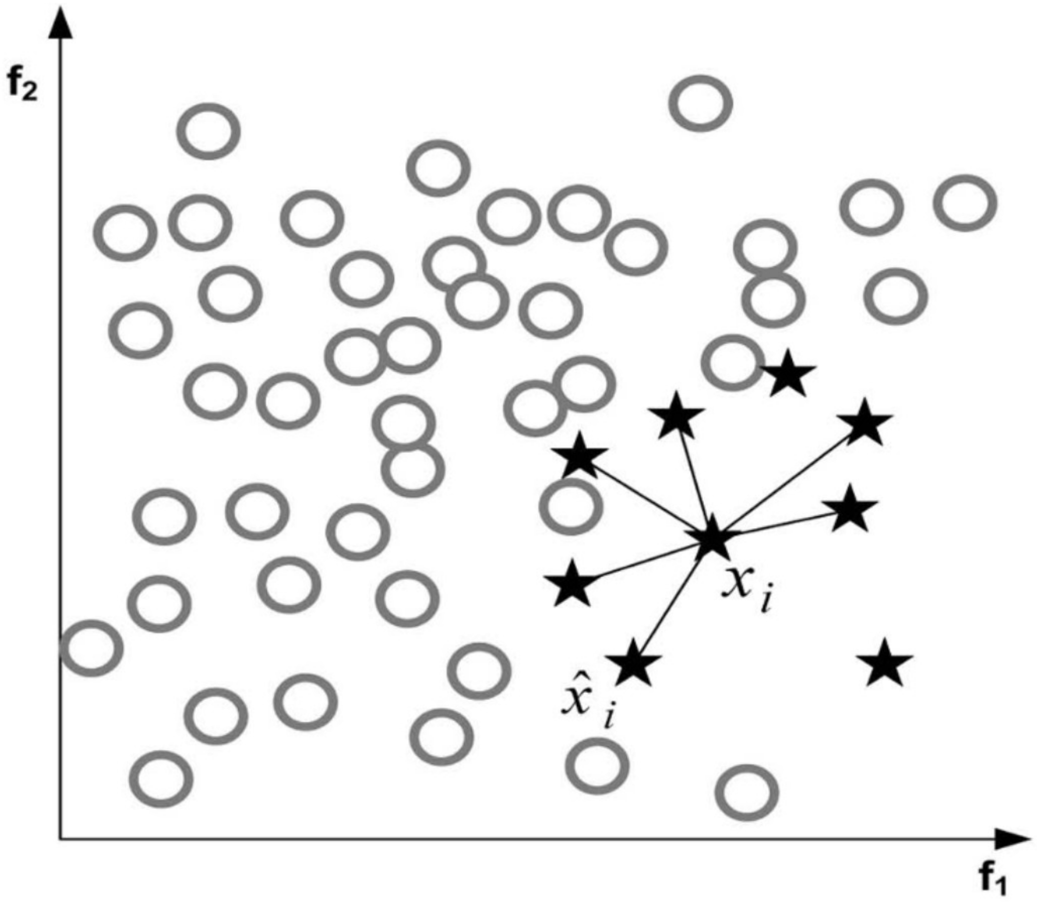
Difficulties

- Scale
- Balance
- Variation
- Encoding
- Access
- Format

Contribution: Scale

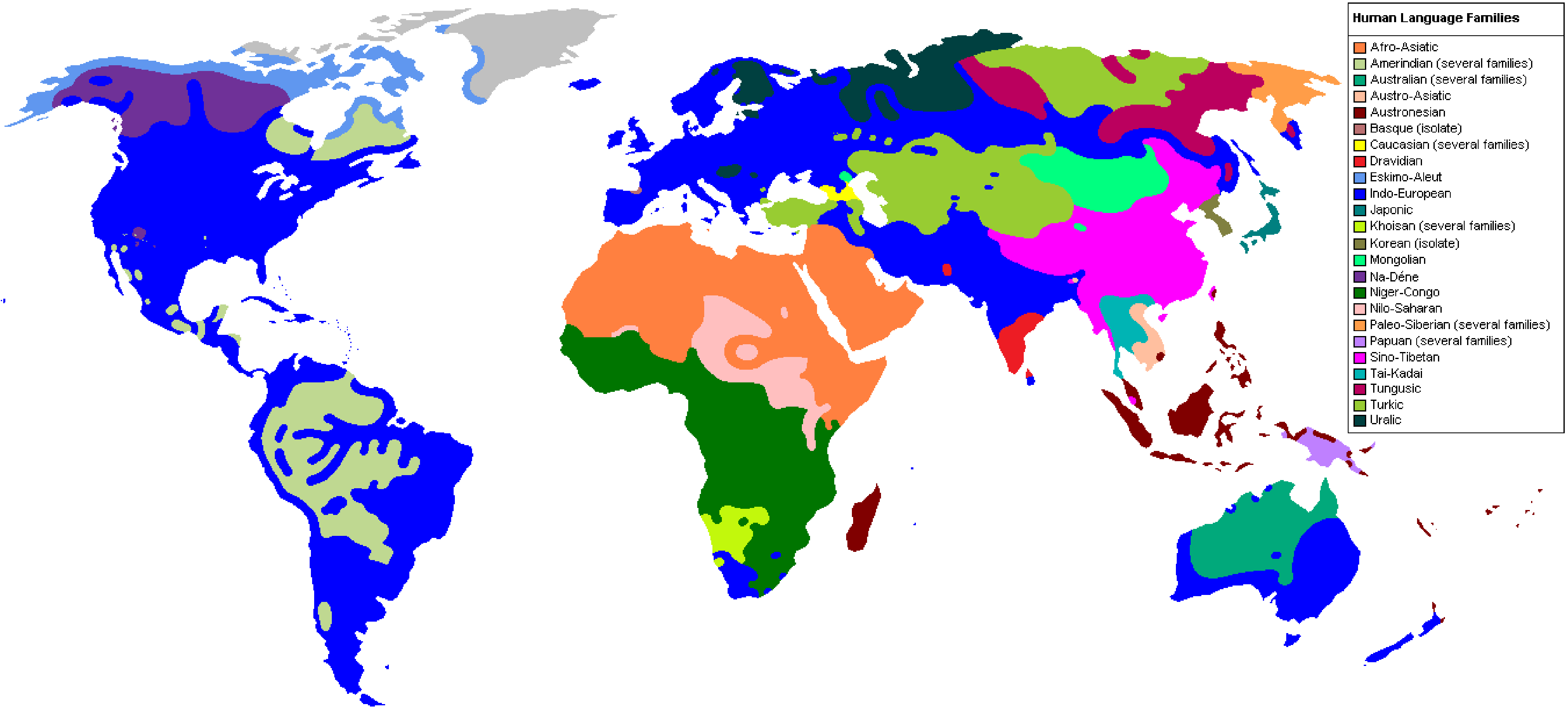
- Uni. Dec. of Human Rights
- Omniglot
- ODIN
- An Crúbadán
- Wikipedia

Contribution: Balance



(He and Garcia, 2009)

Contribution: Variation



(Wikimedia Commons)

Contribution: Variation

Danish:

Alle mennesker er født frie og lige i værdighed og rettigheder.

Norwegian (Bokmål):

Alle mennesker er født frie og med samme menneskeverd og menneskerettigheter.

Norwegian (Nynorsk):

Alle menneske er fødde til fridom og med same menneskeverd og menneskerettar.

Swedish:

Alla människor är födda fria och lika i värdighet och rättigheter.

Contribution: Variation

Zulu:

Bonke abantu bazalwa bekhululekile belingana ngesithunzi nangamalungelo.

Xhosa:

Bonke abantu bazalwa bekhululekile belingana ngesidima nangokweemfanelo.

Ndebele:

Abantu bonke bazalwa bekhululekile njalo belingana kumalungelo abo.

Southern Sotho:

Batho bohle ba tswetswe ba lokolohile mme ba lekana ka botho le ditokelo.

Plan

- Consider more languages
- Identify language families

Comments on Slides

4 To give a flavour of the kind of task we would anticipate, imagine you find this text but don't know what language it's written in. In fact, it is the first article of the Universal Declaration of Human Rights, in Inuktitut. They use a syllabic writing system, where symbols are rotated to change the vowel.

6 The previous example could be indentified from the script alone (assuming we don't distinguish between varieties of Inuktitut), but here is another language, written in the Latin alphabet without diacritics. Again, it is the first article of the Universal Declaration of Human Rights, but in Seychelles Creole.

7 Most previous attempts at language identification systems focus on a handful of high-resource languages. Extending this to all the of the world's languages would be a challenging task.

8-9 The amount of data available for different languages varies hugely - for illustration, the size of the English Wikipedia dwarfs all other Wikipedias. Plotting on a logarithmic scale makes the graph easier to read (even if it becomes less striking). Some previous work has performed language identification for 60 or 70 languages, where we would still have tens of thousands of Wikipedia articles. Moving up to 250 languages leaves us with only a few hundred articles per language.

10 There are many competing encoding standards, and even if we try to convert everything to unicode, we can still run into problems. For instance, although the Georgian alphabet has its own unicode symbols, some people instead use Ascii and a special font - searching this Latin string on Google will give you results that look like the Georgian string. (It means "universal declaration of human rights".)

11 Access to many resources is restricted.

12 We only want text, which means we need to remove all metadata and formatting. The example is from the Aragonese Wikipedia - we are only interested in the text at the bottom, and not in the Wikipedia markup.

13 Even if we can get access to data, which uses the encoding we expect, and the text is easy to extract, we still need to decide what we're trying to identify. For low-resource languages, it can be particularly difficult to draw a line between "language" and "dialect".

14 The three problems on the righthand side are issues that we will have to deal with, but it's far beyond the scope of our project to solve them in any meaningful way. However, the three problems on the lefthand side are issues where we believe we can try something that hasn't been done before.

15 We've identified five data sources which are publicly available, and which cover at least a few hundred languages.

16 To deal with imbalanced data, we can try applying techniques from the machine learning literature.

17-19 In cases when identifying a specific language variety is difficult, we can try to identify a language family instead. Examples are given from Scandinavian and Bantu languages (in both cases, the first sentences of the Universal Declaration of Human Rights). As well as demonstrating that closely related languages can be hard to distinguish, we can see some interesting differences in orthography: for example Swedish uses <ö> where Danish and Norwegian use <ø>; Southern Sotho writes <ba> as a separate word, while Zulu, Xhosa and Ndebele write it as a prefix (Bantuists refer to these as "disjunctive" and "conjunctive" orthographies).

20 Our first step will be to train a naive Bayes classifier using word-lists and n-grams. We will then start experimenting with feature engineering and more sophisticated algorithms.

References

7 I think this image is from an old version of the UNESCO Atlas of the World's Languages in Danger. The current version is: http://www.unesco.org/culture/languages-atlas/

8-9 Data from Wikimedia Commons (http://meta.wikimedia.org/wiki/List_of_Wikipedias). Graphs are my own.

12 http://dumps.wikimedia.org/

13 This adage is often attributed to Max Weinreich, who heard from an audience member. It is presumably much older.

16 Haibo He and Edwardo A. Garcia. "Learning from imbalanced data." Knowledge and Data Engineering, IEEE Transactions on 21.9 (2009): 1263-1284.

17 http://en.wikipedia.org/wiki/File:Primary_Human_Language_Families_Map.png

4,6,18,19 http://www.omniglot.com/udhr/