Computational Linguistics for Low-Resource Languages

October 26, 2011

Alexis Palmer

UNIVERSITÄT
DES
SAARLANDES

CL for LRL

Questions of interest

- What is a low-resource language? (aka less-studied language, resource-poor language, minority language, less-privileged language, ...)
- What are the challenges posed by LRL, and what are the major approaches to addressing these challenges?

CL for LRL

Questions of interest

- What is a low-resource language? (aka less-studied language, resource-poor language, minority language, less-privileged language, ...)
- What are the challenges posed by LRL, and what are the major approaches to addressing these challenges?

Some major themes

- Role of labeled/annotated data
- Role of expert/linguistic knowledge (anno & beyond)
- Single language vs. “universal” solutions
- Resource creation: does it always make sense? how can it be done most efficiently?

And another question...

Why do we care?

- ✦ practical reasons
- ✦ theoretical reasons

Course requirements & organization

- ✦ reading & participation: read papers prior to relevant meeting, discuss
- ✦ presentation: 30-45 minute presentation of selected paper(s), discussion after
- ✦ additional: 1 lg. resource case study, 2 critical reviews (1-2 pages each)
- ✦ term paper: original research or in-depth survey and analysis (15-20 pages)
- ✦ optional: guest post(s) on Cyberling blog

Language endangerment

Language loss

- Current estimated rate of language death: one every 2 weeks (Crystal 2000)
- Half of world's languages extinct by end this century
- UNESCO Endangered Languages Programme (under auspices of Section on Intangible Cultural Heritage)
- UN General Assembly: 2008 was International Year of Languages

UNESCO endangerment status

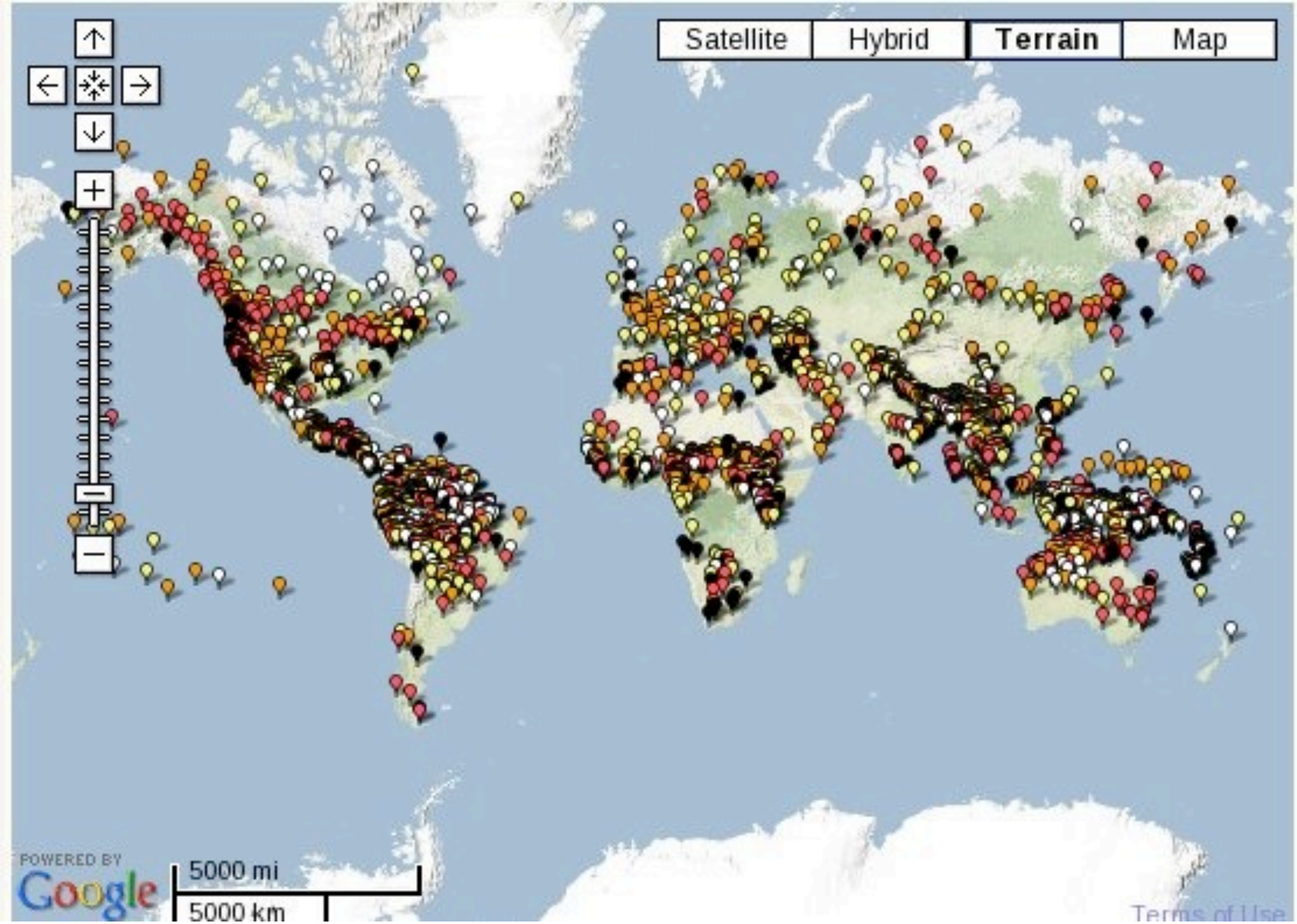
- six levels: safe, unsafe (or vulnerable), definitively endangered, severely endangered, critically endangered
- criteria go beyond number of speakers

Evaluating language endangerment

Criteria to consider (UNESCO 2003)

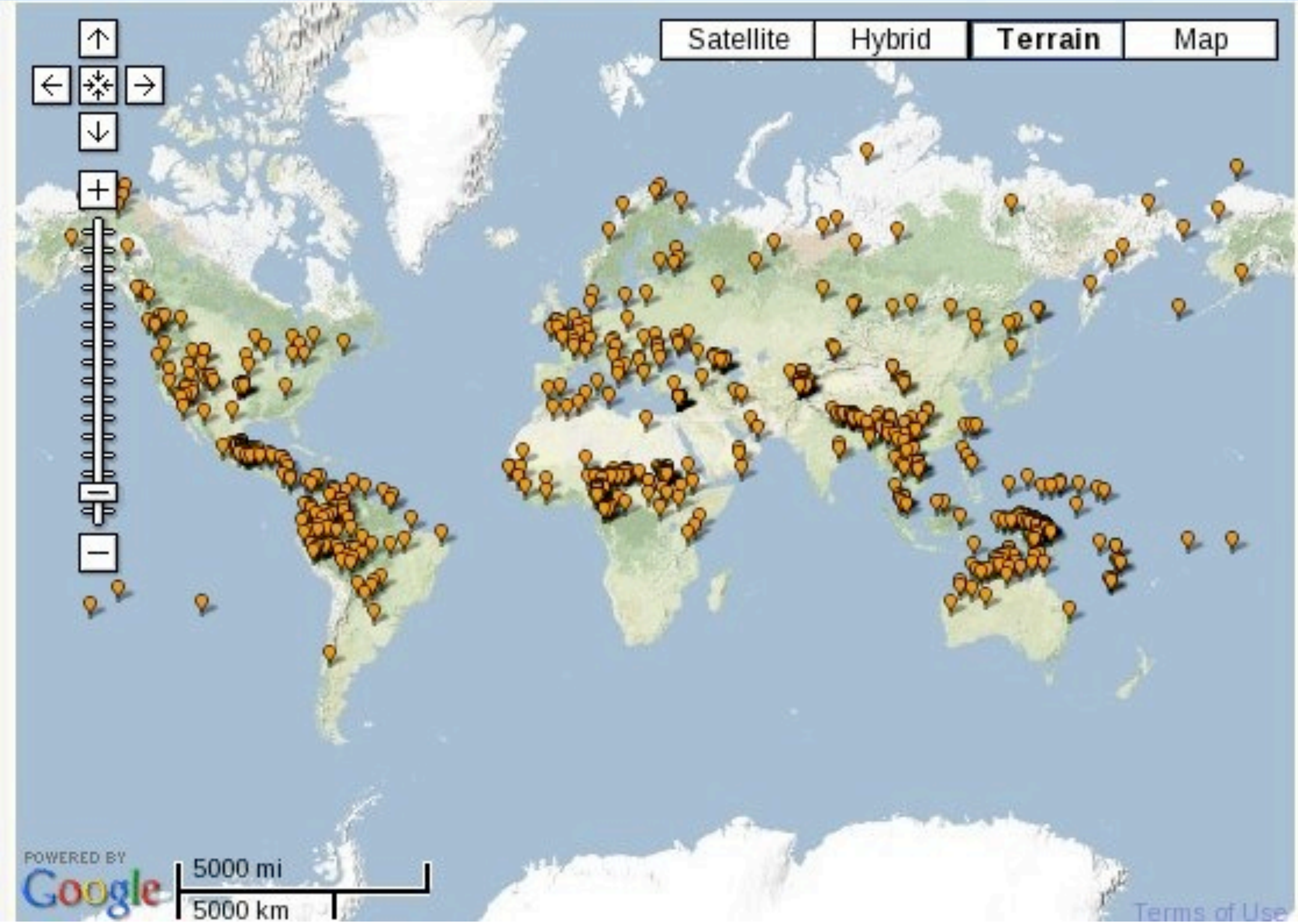
- Intergenerational language transmission
- Absolute number of speakers
- Proportion of speakers within the total population
- Trends in existing language domains
- Response to new domains and media
- Materials for language education and literacy
- Governmental and institutional attitudes and policies, including official status and use
- Community members' attitudes toward their own language
- Amount and quality of documentation

Globally, 2488 languages in danger



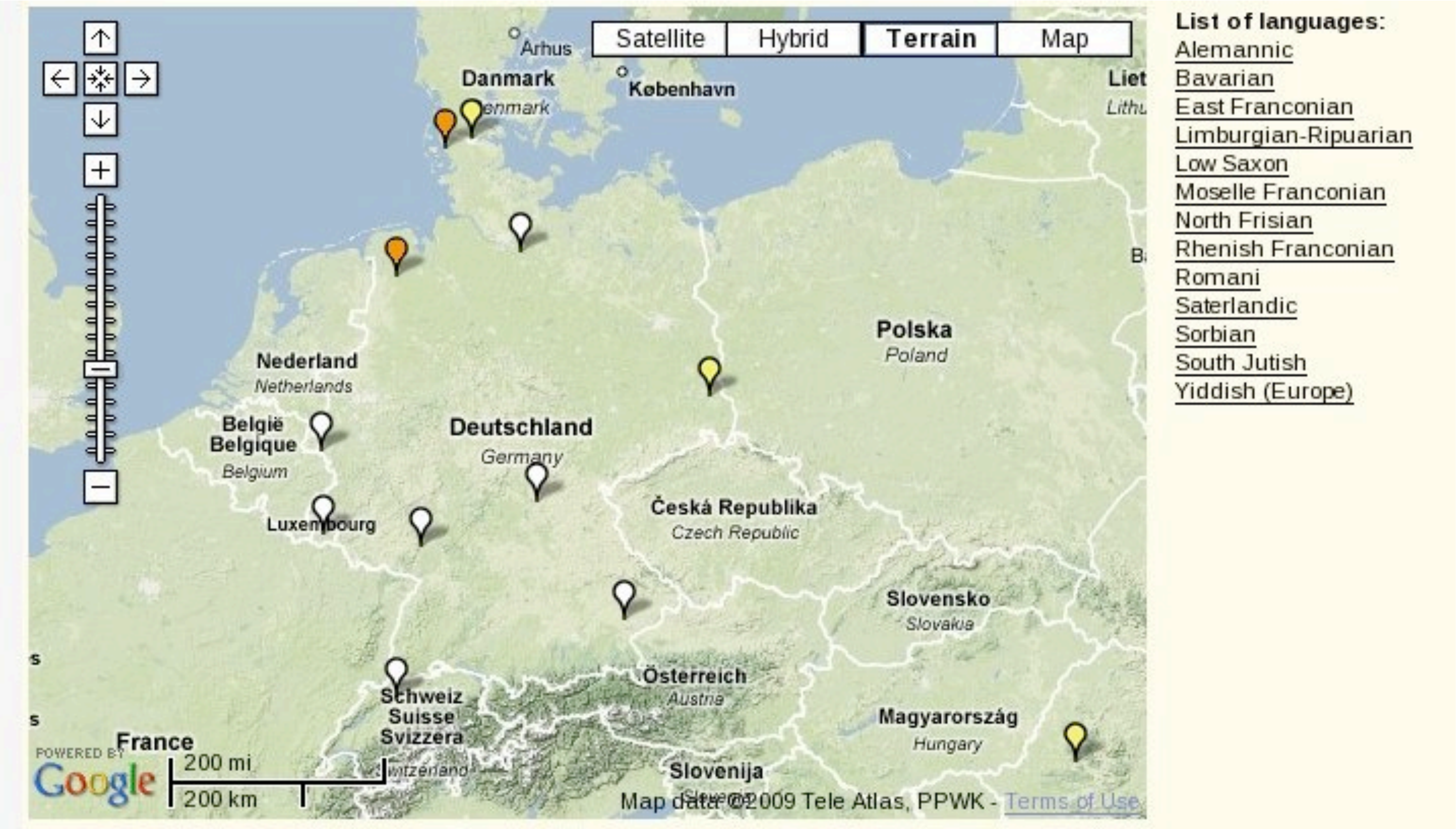
source: UNESCO Interactive Atlas of the World's Languages in Danger, 2009 edition

528 ‘severely endangered’ languages



source: UNESCO Interactive Atlas of the World’s Languages in Danger, 2009 edition

Germany: 13 endangered languages



List of languages:

Alemannic
Bavarian
East Franconian
Limburgian-Ripuarian
Low Saxon
Moselle Franconian
North Frisian
Rhenish Franconian
Romani
Saterlandic
Sorbian
South Jutish
Yiddish (Europe)

source: UNESCO Interactive Atlas of the World's Languages in Danger, 2009 edition

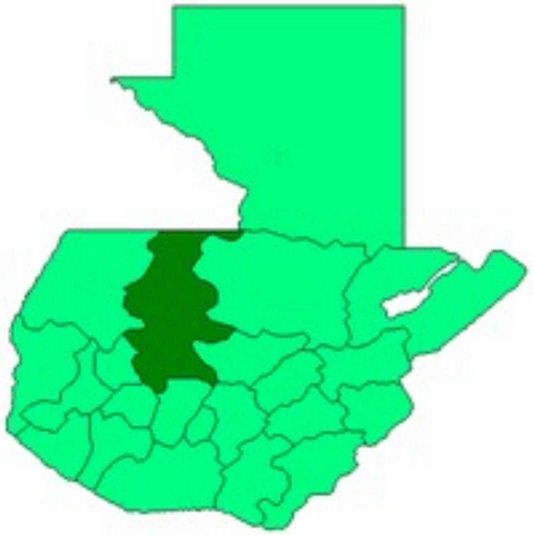
Documenting endangered languages

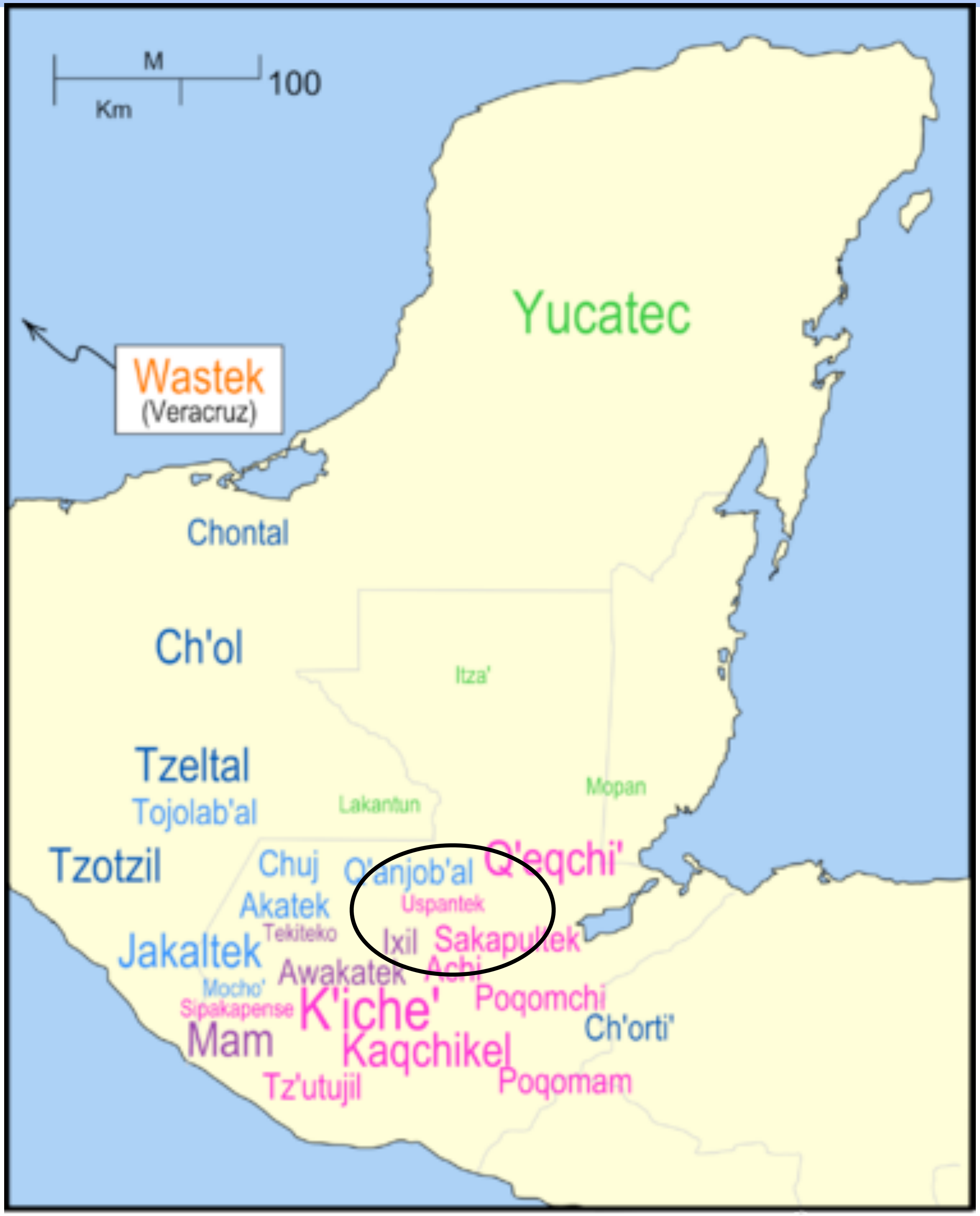
The realities

- Most projects are individual or small-group endeavors with very small budgets
- Each project seems to find its own workflow
- Basic workflow: collection, transcription, translation, detailed linguistic annotation (NOT a pipeline)
- Tangible end products: orthographies, grammars, dictionaries, language teaching and learning materials, collections of stories, websites, etc.
- Such materials support survival of the language
- Do they support CL/NLP???

Uspanteko : 1320 speakers, ‘unsafe’ status

- Uspantán, Quiché Department, Guatemala





Scenario: IGT for Uspanteko

- Corpus of texts in the Mayan language Uspanteko
  - Produced by OKMA (Oxlajuuj Keej Maya' Ajtz'iib')
  - 66 texts, mostly oral history, personal experience, and stories
  - Total 284K words of transcribed text, 74K words glossed
- IGT-XML: representational format specifically for IGT

	# texts	**# morphemes**
train	21	38802
dev	5	16792
test	6	18704

Types of resources

Data

- primary: audio, video, texts (archiving)
- machine-readable corpora
- data with annotations
- parallel corpora, comparable corpora

Linguistic resources

- traditional: grammars, dictionaries, word lists
- WordNet, other ontological resources
- treebanks, etc.

Tools

- user-oriented: spell checkers, input systems, etc.
- for NLP: tokenization, POS tagging, parsing, etc.

Challenges and approaches

Having to do with insufficiency of data

- create more data?
- leverage resource-rich languages
- use semi- or unsupervised methods
- use rule-based methods
- ...

Having to do with the nature of the data

- use linguistic knowledge to seed unsupervised models
- use linguistic knowledge to adapt models/approaches
- change the data to look more like familiar languages
- ...

Topics and scheduling

Topics

Data/resource creation

- annotation; crowd sourcing; active learning
- lexicon building
- “low-level” issues: orthography, character sets/encoding, spell checkers

POS tagging and morphological analysis

- unsupervised POS tag induction
- unsupervised morphological analysis (e.g. Morfessor)
- morph. by alignment and projection
- universal POS tag set, universal linguistic ontologies

Topics

Syntactic analysis

- grammar engineering [guest lecture]
- grammar induction
- parse projection; evaluation; treebanking

Other topics

- machine translation; crisis MT
- cross-lingual approaches to information retrieval, word-sense disambiguation, etc.
- leveraging resource-rich languages

Linguistic universals and typology

- inducing language classifications; linguistic universals
- empirically-driven linguistic typology

Scheduling

- 2 Nov: resource case studies; Bird/Simons [me]
- 9 Nov: no meeting
- 16 Nov: guest lecture, Antske Fokkens [grammar engineering, Grammar Matrix]

For next week:

- Bird and Abney on building a Universal Corpus
- Bird and Simons on requirements for good data
- Language resource case study (1-2 pages)
- Meet with me to finalize topic and schedule