NLP Tools for Low-Resource Languages

22 October 2013

Alexis Palmer, Michaela Regneri

UNIVERSITÄT
DES
SAARLANDES

And another question...

Why do we care?

- ✦ practical reasons
- ✦ theoretical reasons

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

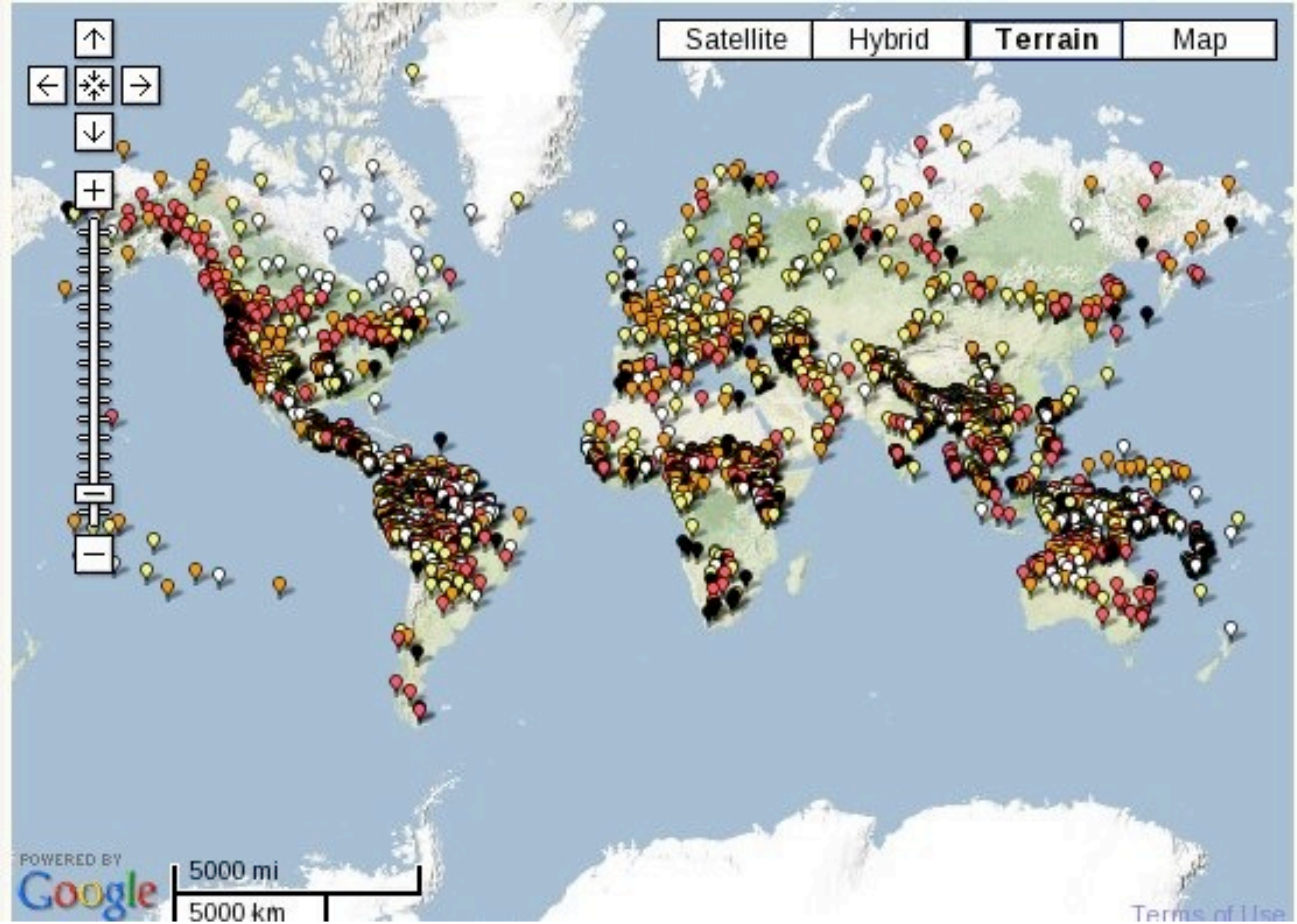
- six levels: safe, unsafe (or vulnerable), definitively endangered, severely endangered, critically endangered
- criteria go beyond number of speakers

Evaluating language endangerment

Criteria to consider (UNESCO 2003)

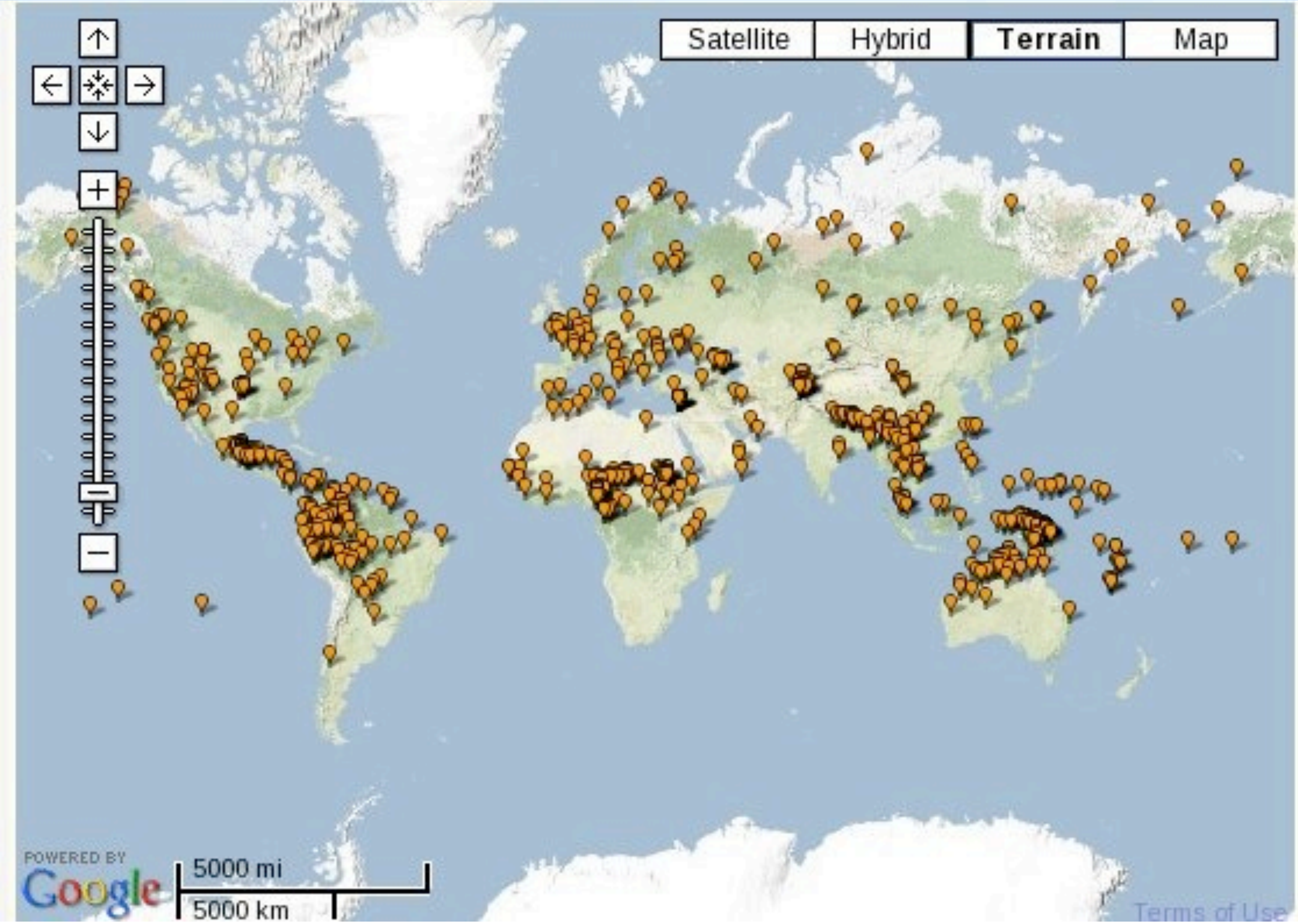
- Intergenerational language transmission
- Absolute number of speakers
- Proportion of speakers within the total population
- Trends in existing language domains
- Response to new domains and media
- Materials for language education and literacy
- Governmental and institutional attitudes and policies,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

Globally, 2488 languages in danger



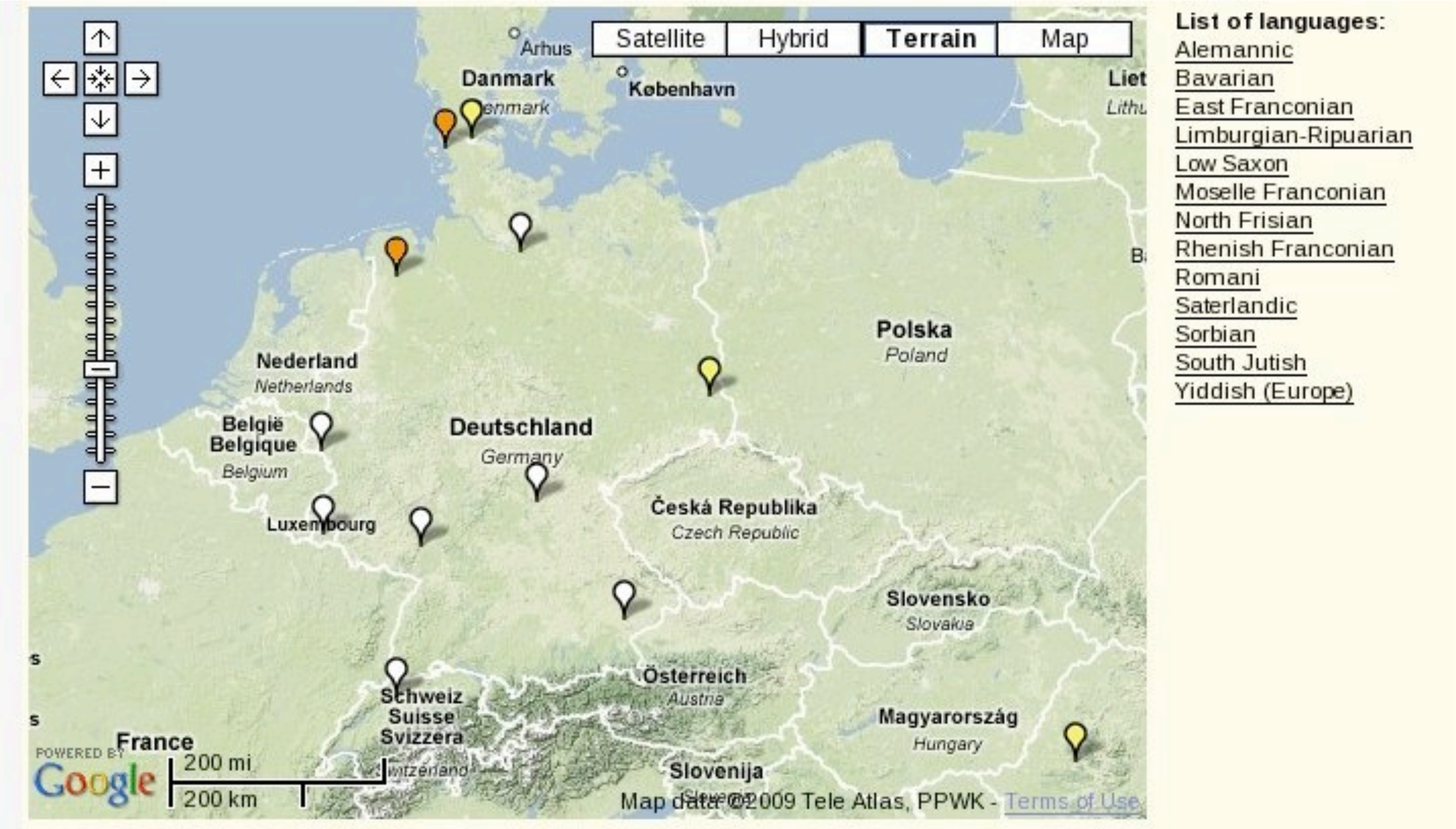
source: UNESCO Interactive Atlas of the World's Languages in Danger, 2009 edition

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

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
- ...

CL for LRL, from the perspective of LRL

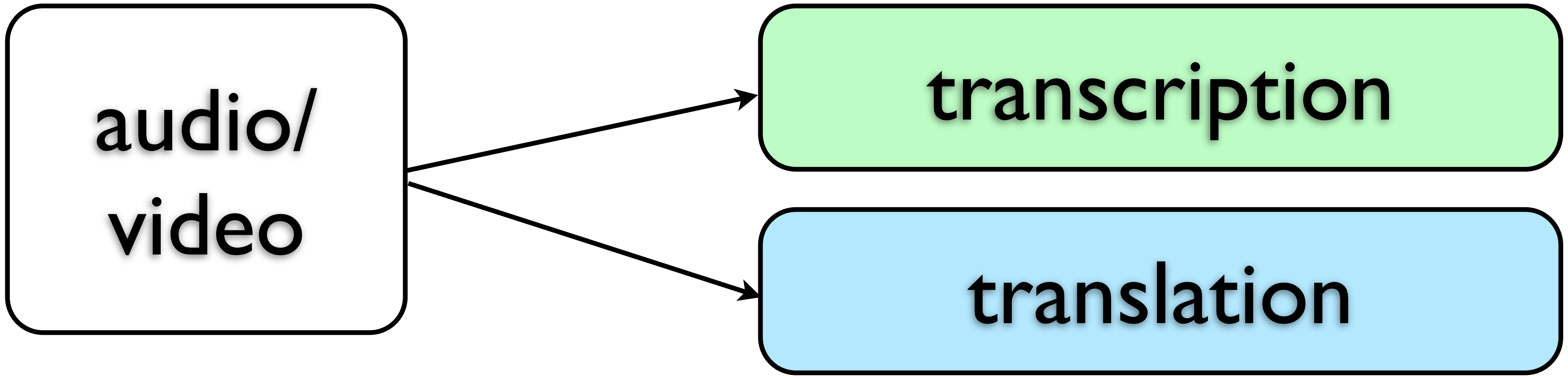
Some major concerns

- More annotated data, with “better” annotations
- Less time spent on low-level aspects of producing this data

Related themes

- Accessible technologies, easy to use
- Privacy, security of data, access and archiving
- Avoiding proprietary formats
- Funding, always

To start with a dream...

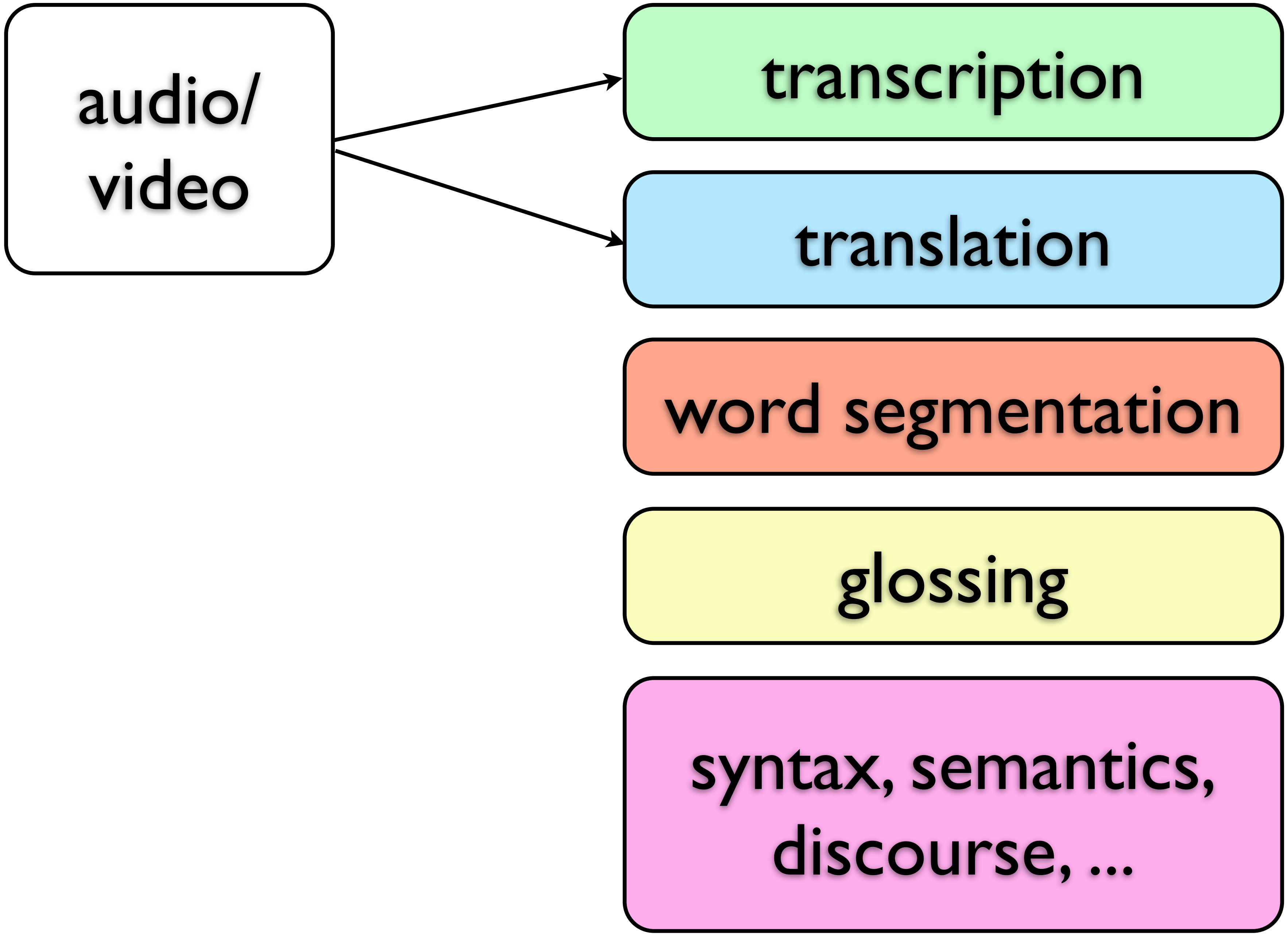


audio/ video

transcription

translation

To start with a dream...



audio/ video

transcription

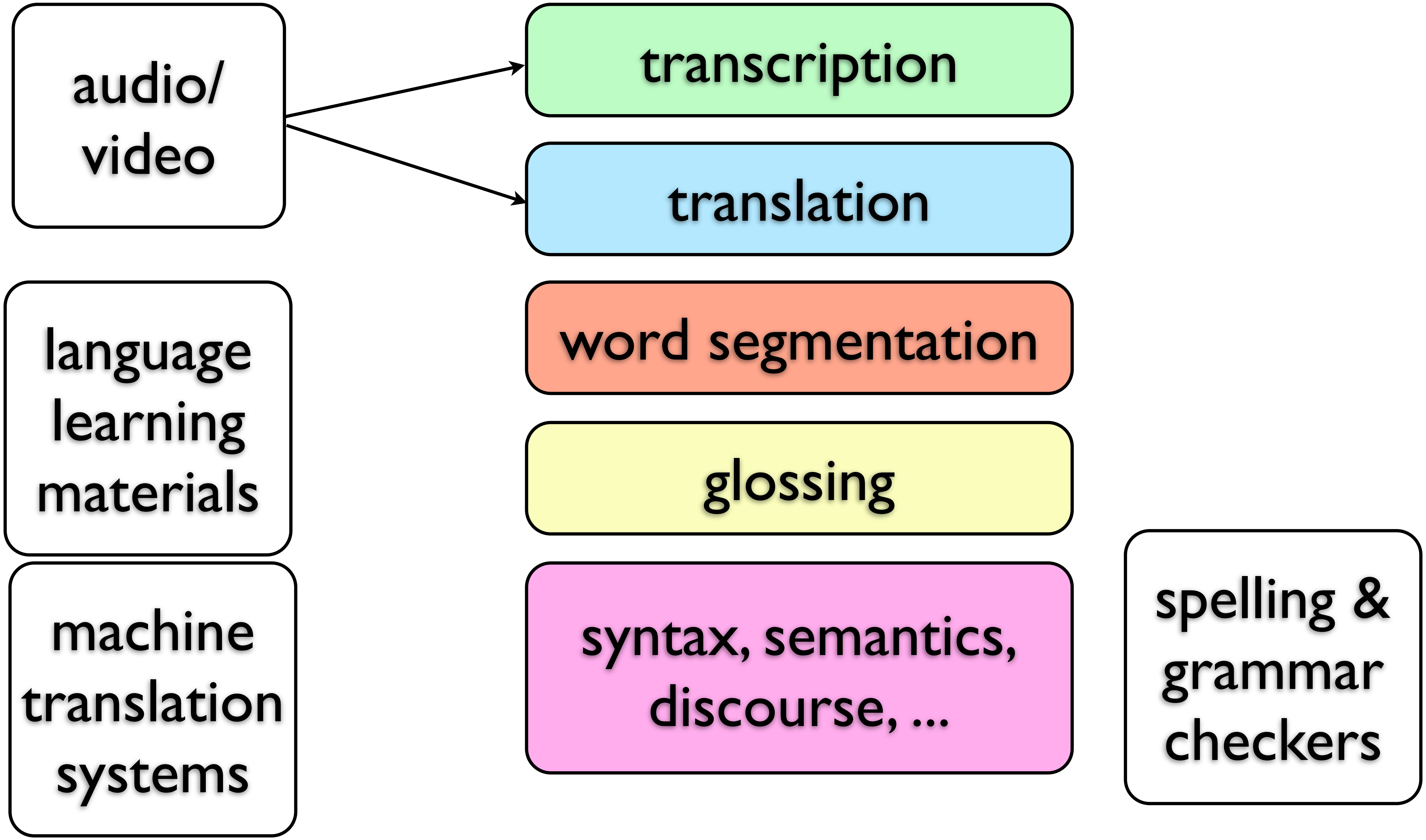
translation

word segmentation

glossing

syntax, semantics, discourse, ...

To start with a dream...



audio/ video

transcription

translation

word segmentation

glossing

syntax, semantics, discourse, ...

language learning materials

machine translation systems

spelling & grammar checkers

Challenges: data, linguistic knowledge

Access to data is the #1 challenge

- data may not exist
- data may be inaccessible
- data may not be machine readable
- data may be inconsistently transcribed, translated, annotated

Linguistic knowledge not always complete

- changes/differences in orthography
- ongoing analysis
- multiple sources or annotators

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???

Project ideas and teams

UNIVERSITÄT
DES
SAARLANDES

A few organizational points

- ✦ course homepage:
  www.coli.uni-saarland.de/courses/cl4lrl-swp
- ✦ our wiki:
  wiki.coli.uni-saarland.de/cl4lrl/swproject
- ✦ access credentials:
  - username: cl4lrl
  - password: Aid1aiji
- ✦ make an account to edit wiki

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.

Data

Data sources:

- Four Mayan languages: annotated with translations into Spanish, POS tags, morphological segmentation and glosses
- Pali: POS labels, ongoing annotation project in Trier
- Speech: for 10 different languages, 10h recorded speech plus transcription and pronunciation lexicon
- others that you find (or create?)

Possible types of projects

Just a few ideas:

- Specific tool for specific language: POS tagger, morphological analyzer, spell checker, etc.
- Speech recognition for a given language
- Annotation tool or interface
- Corpus interface tool
- Wikipedia-based tools: e.g. named entity recognition, ontology creation, etc.
- Something with Twitter, blogs, Facebook, etc.
- Something else.....

For next week

- ✦ rough project proposal from each team
  - what you want to develop
  - what resources & tools you might require
  - availability of those resources & tools
- ✦ wiki page for each team