# *Portability of ASR Technology to new Languages: multilinguality issues and speech/text resources*

*J. Kunzmann, K. Choukri, E. Janke, A. Kießling, K. Knill, L. Lamel, T. Schultz, and S. Yamamoto*

*Automatic Speech Recognition and Understanding ASRU, December 2001*

# *Topics which will be addressed*

Everybody speaks English: why bother with other languages?

Doing another language is simply training with other data: no science left?

Language portability: only an acoustic issue?

Multilingual ASR: what is it good for?

Data: what is available, what do we need?

Beyond ASR

# *Why bother with other languages?*

Myth: "Everyone speaks English, why bother?"

About 4500-6000 different languages exist in the world

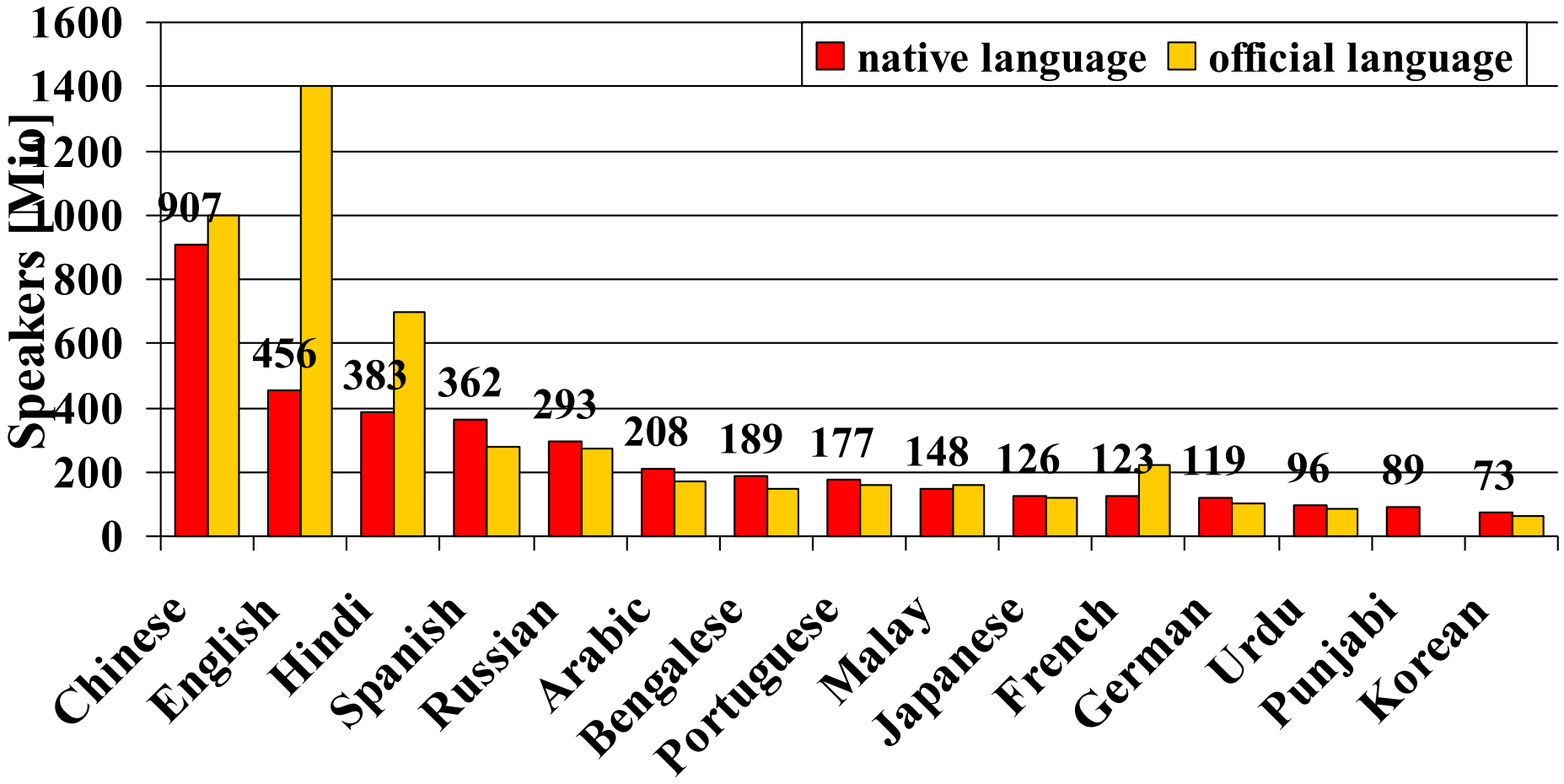
Number of languages on internet is increasing

English Internet pages: 80% -> 40% in 10 years

Users' mother tongue for acceptance

Non-native speech

# Top-15 Languages of the World




Webster‘s New Encyclopedic Dictionary, 1992

# *Another language? - No science*

Myth: ASR in another language - It's just training on another database - there is no science here

BUT: Other languages bring unseen challenges

Have we even seen "all" language characteristics?

Have we seen most of the language characteristics?

Do we have the big picture?

How do the differences effect ASR?

# *Language Characteristics*

What is a word? "the written string between two blanks"

Exp: Osman-lı-laç-tır-ama-yabil-ecek-ler-imiz-den-miş-siniz

Inflection system?

Effects for ASR: language modeling

- text processing, #words in text
- vocabulary size, OOV-rates

- Performance Comparison?

# *Language Characteristics*

Grapheme-to-phoneme relation / writing system

العربى български català 中国话 hrvatski česky
english ελληνικα עברית हिंदी italiano 日本語
한국어 românește русский српски ภาษาไทย

- No written form at all!

Effects for ASR: Pronunciation dictionary

# *Language Characteristics*

Linguistic structure

- Phoneme system (number/confusability)
- Tonality, stress pattern
- Phonotactics (mora, consonant clusters)
- Coarticulation

Effect for ASR:

- Myth: IPA for real?
- What kind of acoustic units?
- Suprasegmental modeling?

# *Questions to the audience*

Everybody speaks English: Why bother?

For how many languages are speech interfaces needed?

ASR in another language: no science?

Have we seen most of the language characteristics already?

Do we have the big picture?

# *Language Portability*

Standard porting steps: do they work?

- Audio data, text data, pronunciation model
- (for some applications and languages -> yes)

What are suitable acoustic models for bootstrapping ?

- Language independent ASR ?

What phoneme set to use ?

What lexicon ?

# *Why we need portability of ASR technologies to N languages*

Portability of ASR system/technology for human-machine interface to N languages

When ASR system/technology is applied to other languages,

- lack of speech corpus for acoustic modeling
- lack of spoken language corpus for language modeling

Portability of ASR system/technology for multilingual speech translation

**Extension to multilingual speech communication**

# *Portability of ASR technology for multilingual speech translation*

- Speech translation = speech recognition + machine translation + other functions
- Speech recognition requires a huge speech corpus.
- Machine translation technology is shifting from rule-based technology to corpus-based technology such as Stochastic MT or Example based MT.
- Corpus-based MT technology requires a huge sentence aligned bilingual spoken language corpus.
- One of the key issues is creation of sentence aligned corpus.
- Some huge bilingual text corpora available
- Lack of bilingual spoken language corpora

# *Multilinguality*

What is multilinguality?

- Seen/unseen languages
- Non-native speech and language
- Multiple systems with language switching

Should we be building language independent models ?

Is multilingual pronunciation modelling possible ?

Is multilingual language modelling sensible ?

# *Data*

What data do we need ? Make a wish

- speech, transcriptions, lexicon, text corpora
- number of languages
- amount of speech and text data (#hours, #speakers, #words)
- application domain

What is available ?

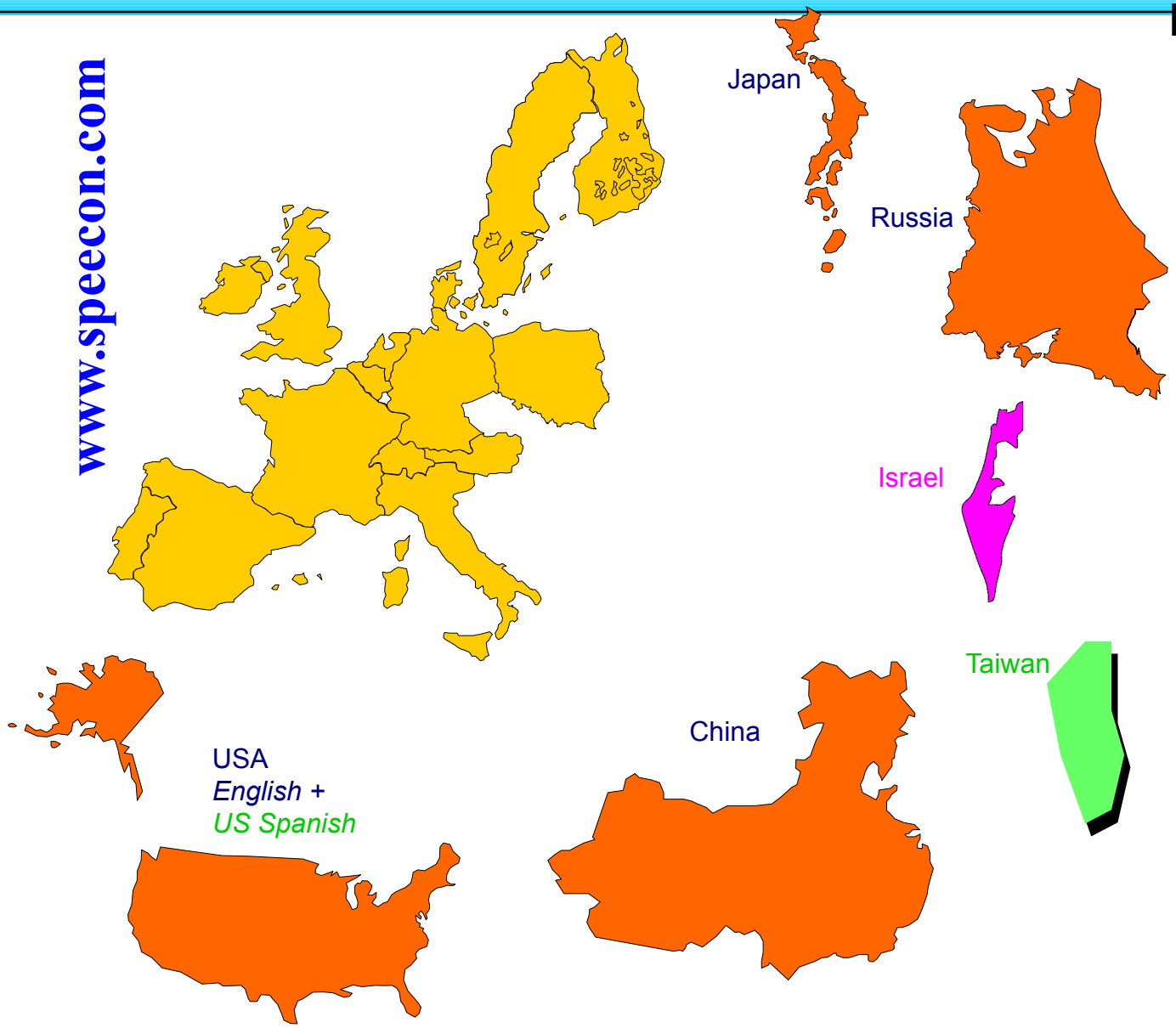
Do we have the right data?

# *Data: What is available?*

- What is available?
  - From ELRA and LDC
    - Transcribed speech data in >20 languages
    - Pronunciation dictionaries in the order of 10 languages
    - text corpora > 20 languages
  - GlobalPhone
- What is planned?
  - Speecon, OrienTel
  - Bilingual data ATR

# THE WORLD ACCORDING TO SPEECON

www.speecon.com




## Languages

- Danish
- Dutch
- UK-English
- US-English
- Finnish
- Flemish
- French French
- German & Austrian German
- Swiss German
- Hebrew
- Italian
- Japanese
- Mandarin Chinese
- Polish
- Portuguese
- Russian
- Spanish
- Swedish
- Mandarin Taiwan
- US Spanish
- ...

# GlobalPhone

Arabic
Ch-Mandarin
Ch-Shanghai
English
French
German
Japanese
Korean
Croatian
Portuguese
Russian
Spanish
Swedish
Tamil
Turkish

<u>Multilingual Database</u>

Uniformity

Widespread languages

Newspaper domain

+ Large text corpora

<u>Total sum of resources</u>

15 languages so far

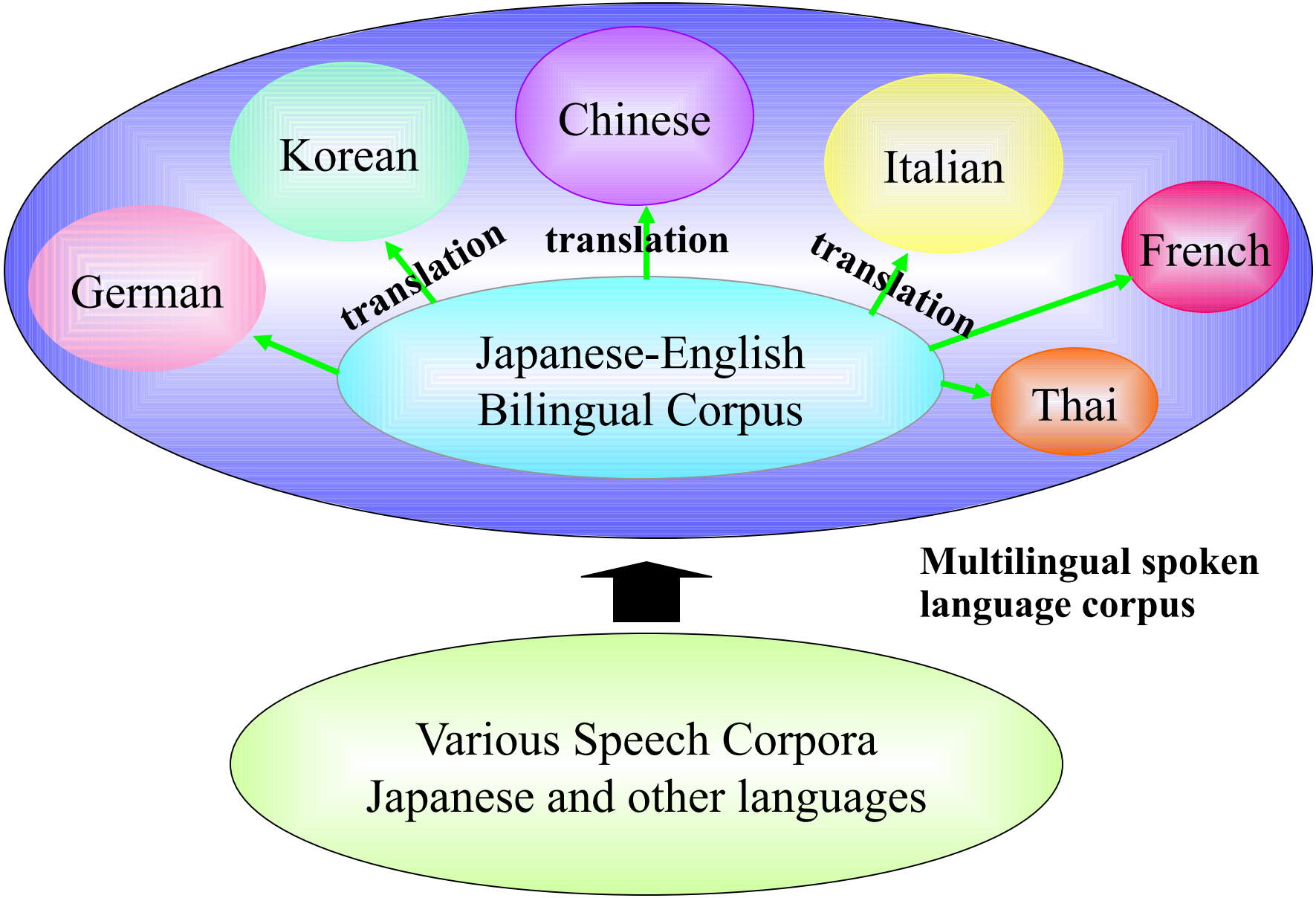
Fully transcribed

(15x20) ⌈ 300 h speech

⌈ 1400 native speakers

Ready, Soon available

# *Speech/Text Resources being collected at ATR*



# *Speech/Text Resources being collected at ATR*

For both ASR and MT

- Bilingual conversation aided by human translators
  - 16,000 utterances
- Bilingual conversation via speech translation systems
  - under construction

For MT

- Text of bilingual conversation
  - 500,000 utterances
- Expanding with various methods including paraphrasing
  - Extension rate is high for paraphrasing.

# *Data: Do we have the right set?*

Do we have the right data?

- What goal do you want to achieve ?
- How much data do we need ?
- Scripts / ready to go data ?

What do we need ?

You can’t always get what you want, you get what you need (Rolling Stones)

# *Beyond ASR*

We cross cultural borders

Are concepts the same across languages?

Defining concepts

- Time concepts: when does the day start
- Politeness concepts:

What is the relationship between “words” and concepts ?

Generation

# *Topics which may have been addressed*

Everybody speaks English: why bother with other languages?

Doing another language is simply training with other data: no science left?

Language portability: only an acoustic issue?

Multilingual ASR: what is it good for?

Data: what is available, what do we need?

Beyond ASR