Language technology for low-resource languages

Day 1/5

LOT 2018, Groningen

Yves Scherrer, University of Helsinki

Yves Scherrer
Post-doc, University of Helsinki

Nordic language processing laboratory (NLPL):
Head of machine translation activity

Finnish:
Neural machine translation and evaluation for Finnish and other morphologically rich languages

Historical Dutch/Slovene:
Normalization with character-level machine translation

Old Russian:
Tagging and normalization

European French:
Surveys, crowdsourcing and mapping of linguistic variation

Rusyn:
Multi-source annotation projection

Swiss German:
Construction of the multi-dialectal ArchiMob corpus
Dialectometrical studies with atlas and corpus data

About this course

- Course material available online:
  - https://github.com/yvesscherrer/lot
- Contact:
  - yves.scherrer@helsinki.fi
- Preparatory readings + discussions in class

Course program

- **Monday:** Definitions of low-resource languages in linguistics and computational linguistics, overview of the main language technology applications and their resource requirements
- **Tuesday:** Annotation projection using parallel corpora
- **Wednesday:** Delexicalisation and relexicalisation approaches
- **Thursday:** Closely related languages and language varieties - definitions, problems and solutions
- **Friday:** Multilingual modelling and zero-shot learning

Goals

- Get an overview of the data situation of various languages and assess the influence of the data situation on the development of language technologies
- Get to know various development methods of language technology systems for resource-poor languages
- Understand and discuss scientific papers on the subject

Today's lecture

- What is a low-resource language?
  - How does this relate to other concepts such as *minority language* or *endangered language*?
- Why should we care about NLP for low-resource languages?
- What are resources in NLP and how does their availability affect research?
- One approach for solving the low-resource problem: create resources

What is a low-resource language?

Related terms and concepts

- Scarce resource language
- Resource-scarce language
- **Low-resource language**
- Under-resourced language
- Lesser-resourced language
- Resource-poor language

- Minority language

- Endangered language

Language classification

- Socio-economical criteria
  - Number of L1 speakers
  - Access to internet
- Sociolinguistic criteria
  - «Competition» with other languages
  - Attitudes, official recognition, …

Endangered languages

Minority languages

- Language-technological criteria (Natural Language Processing, NLP)
  - Degree of informatisation
  - Research activity
  - Resource availability

Low-resource languages

Ethnologue: a meta-resource

- https://www.ethnologue.com/

Ethnologue: a meta-resource

- Hammarström, Harald (2015): Ethnologue 16/17/18th editions – A comprehensive review. *Language* 91, 723-737.
  - "While hundreds of spurious and missing languages can be documented for Ethnologue, it is at present still better than any other nonderivative work of the same scope, in all aspects but one."
  - "Ethnologue fails to disclose the sources for the information presented, at odds with well-established scientific principles."
  - "The classification of languages into families in Ethnologue is [...] found to be far off from that argued in the specialist literature on the classification of individual languages."

Socio-economical criteria

Table 2. Distribution of world languages by number of first-language speakers

Population range	Living languages			Number of speakers		
	Count	Percent	Cumulative	Total	Percent	Cumulative
100,000,000 to 999,999,999	8	0.1	0.1%	2,543,460,358	40.39988	40.39988%
10,000,000 to 99,999,999	80	1.1	1.2%	2,458,383,987	39.04854	79.44843%
1,000,000 to 9,999,999	305	4.3	5.5%	929,591,638	14.76547	94.21390%
100,000 to 999,999	937	13.2	18.7%	294,626,823	4.67980	98.89370%
10,000 to 99,999	1,811	25.5	44.2%	61,556,414	0.97775	99.87145%
1,000 to 9,999	1,978	27.8	72.0%	7,613,358	0.12093	99.99238%
100 to 999	1,062	14.9	87.0%	466,128	0.00740	99.99979%
10 to 99	338	4.8	91.7%	12,944	0.00021	99.99999%
1 to 9	137	1.9	93.7%	541	0.00001	100.00000%
0	204	2.9	96.5%	0	0.00000	100.00000%
Unknown	246	3.5	100.0%			
Totals	7,106	100.0		6,295,712,191	100.00000	

http://www.ethnologue.com/statistics/size

Sociolinguistic criteria

- Two main criteria:
  - Language development
    - Literacy, literature, language use in media, …
    - http://www.ethnologue.com/language-development
  - Language endangerment
    - Speaker population and trends, attitudes, domains of use, …
    - http://www.ethnologue.com/endangered-languages
- The **EGIDS scale** quantifies these two criteria:
  - Expanded Graded Intergenerational Disruption Scale (Lewis and Simons 2010)
  - Based on GIDS (Fishman 1991)

Sociolinguistic criteria

- EGIDS: Expanded Graded Intergene-rational Disruption Scale

Level	Label	Description
0	International	The language is widely used between nations in trade, knowledge exchange, and international policy.
1	National	The language is used in education, work, mass media, and government at the national level.
2	Provincial	The language is used in education, work, mass media, and government within major administrative subdivisions of a nation.
3	Wider Communication	The language is used in work and mass media without official status to transcend language differences across a region.
4	Educational	The language is in vigorous use, with standardization and literature being sustained through a widespread system of institutionally supported education.
5	Developing	The language is in vigorous use, with literature in a standardized form being used by some though this is not yet widespread or sustainable.
6a	Vigorous	The language is used for face-to-face communication by all generations and the situation is sustainable.
6b	Threatened	The language is used for face-to-face communication within all generations, but it is losing users.
7	Shifting	The child-bearing generation can use the language among themselves, but it is not being transmitted to children.
8a	Moribund	The only remaining active users of the language are members of the grandparent generation and older.
8b	Nearly Extinct	The only remaining users of the language are members of the grandparent generation or older who have little opportunity to use the language.
9	Dormant	The language serves as a reminder of heritage identity for an ethnic community, but no one has more than symbolic proficiency.
10	Extinct	The language is no longer used and no one retains a sense of ethnic identity associated with the language.

http://www.ethnologue.com/about/language-status

Sociolinguistic criteria

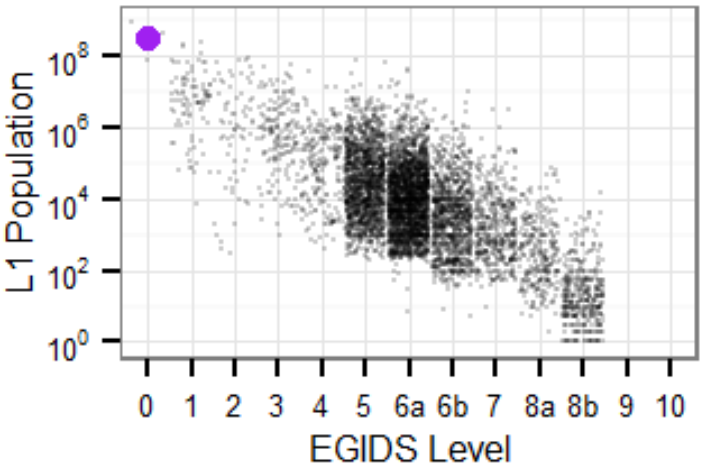
Table 4. Distribution of world languages by vitality status

	EGIDS	Living languages			Number of speakers				
		Count	Percent	Cumulative	Total	Percent	Cumulative	Mean	Median
International	0	6	0.1	0.1%	1,839,439,858	29.2173	29.2173%	306,573,310	335,148,868
National	1	95	1.3	1.4%	1,950,647,667	30.9837	60.2011%	20,533,133	6,915,000
Provincial	2	69	1.0	2.4%	703,103,562	11.1680	71.3691%	10,189,907	1,300,000
Wider Comm.	3	175	2.5	4.9%	553,839,750	8.7971	80.1662%	3,164,799	684,000
Educational	4	215	3.0	7.9%	159,717,369	2.5369	82.7031%	742,871	123,000
Developing	5	1,563	22.0	29.9%	616,297,678	9.7892	92.4923%	394,304	27,000
Vigorous	6a	2,549	35.9	65.7%	403,200,700	6.4044	98.8966%	158,180	10,000
Threatened	6b	1,055	14.8	80.6%	55,965,859	0.8890	99.7856%	53,048	3,000
Shifting	7	464	6.5	87.1%	12,705,302	0.2018	99.9874%	27,382	1,080
Moribund	8a	288	4.1	91.2%	721,897	0.0115	99.9988%	2,507	270
Nearly Extinct	8b	424	6.0	97.1%	72,549	0.0012	100.0000%	171	15
Dormant	9	203	2.9	100.0%	0	0.0000	100.0000%	0	0
	Totals	7,106	100.0		6,295,712,191	100.0000			

http://www.ethnologue.com/about/language-status
http://www.ethnologue.com/statistics/status

Sociolinguistic criteria

- Ethnologue Cloud:
  - Plots number of L1 speakers against EGIDS index
  - Purple dot: English
  - http://www.ethnologue.com/cloud/eng
  - What languages would you expect on the other edges of the graph?



  - Where would you place the languages you know?

Classification of languages Preliminary conclusions

- About 100 «big» languages with more than 10M speakers
  - Measure of commercial interest in languages
- About 100 languages with international or national diffusion and vitality
  - Lahnda/Panjabi: 82M speakers, but only EGIDS level 5
  - Latvian: 1.8M speakers, but EGIDS level 1
- This doesn't tell us anything about NLP…
  - Potential users of NLP tools must be able to use the language efficiently on a computer or a mobile device: **Informatization**

Informatization

- What does it take to use a language efficiently on a computer/mobile device?
  - What type of support would you find useful for a language?
  - What type of support do you find useful for *your* language?

- Example:
  - «45% of Indian language users face challenges in text input on chat applications. Adoption of input mechanisms such as voice to text support, local language keyboards and transliteration is expected to improve user experience.»

https://assets.kpmg.com/content/dam/kpmg/in/pdf/2017/04/Indian-languages-Defining-Indias-Internet.pdf P.30

Informatization

- Availability of basic tools and methods to use a language efficiently on electronic devices
  - Input methods (physical keyboards, mobile devices)
  - Interface language of applications (localization)
  - Simple dictionaries
  - Basic web search
  - OCR (character recognition)
  - Spelling correction
- There is a fine line between informatization and NLP:
  - Grammar correction
  - Speech interface (ASR + TTS)
  - Input prediction
  - Machine translation

How many languages are informatized?

Application	Languages
Microsoft Windows 7 input languages	
Android input languages	
Microsoft Windows 10 interface languages http://windows.microsoft.com/en-us/windows/language-packs	
KDE localization teams http://l10n.kde.org/teams-list.php	
Android interface languages	
Microsoft Office interface languages https://support.office.microsoft.com/en-us/article/Office-language-interface-pack-LIP-downloads-d63007c2-e8ae-41fd-8bfb-fce2857010e1	
Google Search (results) http://www.google.com/advanced_search	
Wiktionary languages with > 20 000 definitions (includes langages such as Latin, Esperanto, Translingual) https://en.wiktionary.org/wiki/Wiktionary:Statistics	

How many languages are informatized?

Application	Languages
Microsoft Windows 7 input languages	130
Android input languages	261
Microsoft Windows 10 interface languages http://windows.microsoft.com/en-us/windows/language-packs	110
KDE localization teams http://l10n.kde.org/teams-list.php	76
Android interface languages	26
Microsoft Office interface languages https://support.office.microsoft.com/en-us/article/Office-language-interface-pack-LIP-downloads-d63007c2-e8ae-41fd-8bfb-fce2857010e1	74
Google Search (results) http://www.google.com/advanced_search	47
Wiktionary languages with > 20 000 definitions (includes langages such as Latin, Esperanto, Translingual) https://en.wiktionary.org/wiki/Wiktionary:Statistics	42

Popular NLP tools

Application	Languages
Microsoft Office proofing tools https://www.microsoft.com/en-US/download/details.aspx?id=35400	91
Google Translate https://translate.google.com/intl/en/about/languages/	103 (> 10000 pairs)
Apple Siri https://www.apple.com/ios/feature-availability/#siri	21

META-NET

- What does META-NET say about the European languages?
  - What types of languages (in terms of EGIDS) are considered?
  - Is informatization a problem?
  - Is language-technological support a problem?
  - What language technology applications are considered?

META-NET

What are common Language Technology applications?

Language technologies include: spelling and grammar checkers; web search; voice dialing; interactive dialogue systems (e. g., phone banking or train reservation systems); interactive assistants such as Apple's Siri or Google's voice search; crosslingual search in digital libraries (e. g., Europeana); term extraction; speech synthesis for navigation systems; recommender systems for online shops; automatic content summarisation; and machine translation systems such as Google Translate and Microsoft's Bing Translator.

http://www.meta-net.eu/vision/reports/meta-net-sra-version_1.0.pdf S. 4

- From a European perspective, the informatization problem is largely solved…
- But: "21 of 30 examined European languages are in danger of *digital extinction*"

META-NET

	Excellent	Good	Moderate	Fragmentary	Weak/None
Machine Translation		English	French, Spanish	Catalan, Dutch, German, Hungarian, Italian, Polish, Romanian	Basque, Bulgarian, Croatian, Czech, Danish, Estonian, Finnish, Galician, Greek, Icelandic, Irish, Latvian, Lithuanian, Maltese, Norwegian, Portuguese, Serbian, Slovak, Slovene, Swedish
Speech		English	Czech, Dutch, Finnish, French, German, Italian, Portuguese, Spanish	Basque, Bulgarian, Catalan, Danish, Estonian, Galician, Greek, Hungarian, Irish, Norwegian, Polish, Serbian, Slovak, Slovene, Swedish	Croatian, Icelandic, Latvian, Lithuanian, Maltese, Romanian

META-NET SRA 1.0 S. 29

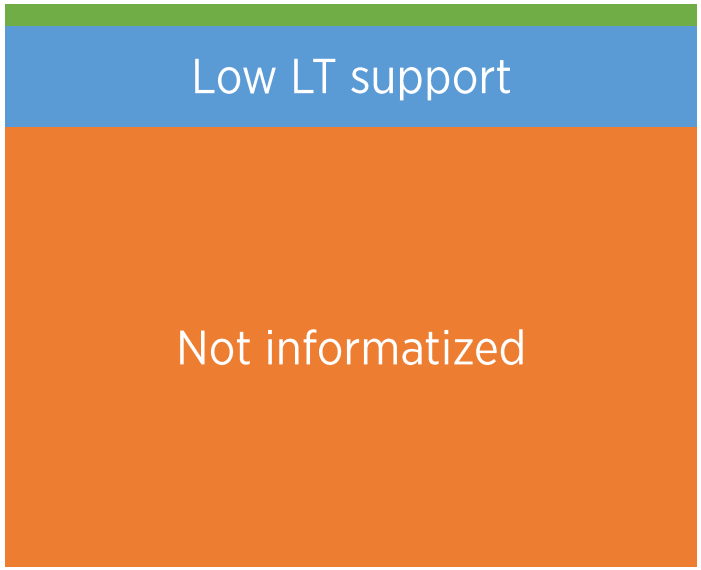
META-NET

	Excellent	Good	Moderate	Fragmentary	Weak/None
Text Analytics		English	Dutch, French, German, Italian, Spanish	Basque, Bulgarian, Catalan, Czech, Danish, Finnish, Galician, Greek, Hungarian, Norwegian, Polish, Portuguese, Romanian, Slovak, Slovene, Swedish	Croatian, Estonian, Icelandic, Irish, Latvian, Lithuanian, Maltese, Serbian
Language Resources		English	Czech, Dutch, French, German, Hungarian, Italian, Polish, Spanish, Swedish	Basque, Bulgarian, Catalan, Croatian, Danish, Estonian, Finnish, Galician, Greek, Norwegian, Portuguese, Romanian, Serbian, Slovak, Slovene	Icelandic, Irish, Latvian, Lithuanian, Maltese

META-NET SRA 1.0 S. 29

Classification of languages

	EGIDS	Living languages		
		Count	Percent	Cumulative
International	0	6	0.1	29.2173%
National	1	95	1.3	60.2011%
Provincial	2	69	1.0	71.3691%
Wider Comm.	3	175	2.5	80.1662%
Educational	4	215	3.0	82.7031%
Developing	5	1,563	22.0	92.4923%
Vigorous	6a	2,549	35.9	98.8966%
Threatened	6b	1,055	14.8	99.7856%
Shifting	7	464	6.5	99.9874%
Moribund	8a	288	4.1	99.9988%
Nearly Extinct	8b	424	6.0	100.0000%
Dormant	9	203	2.9	100.0000%
	Totals	7,106	100.0	

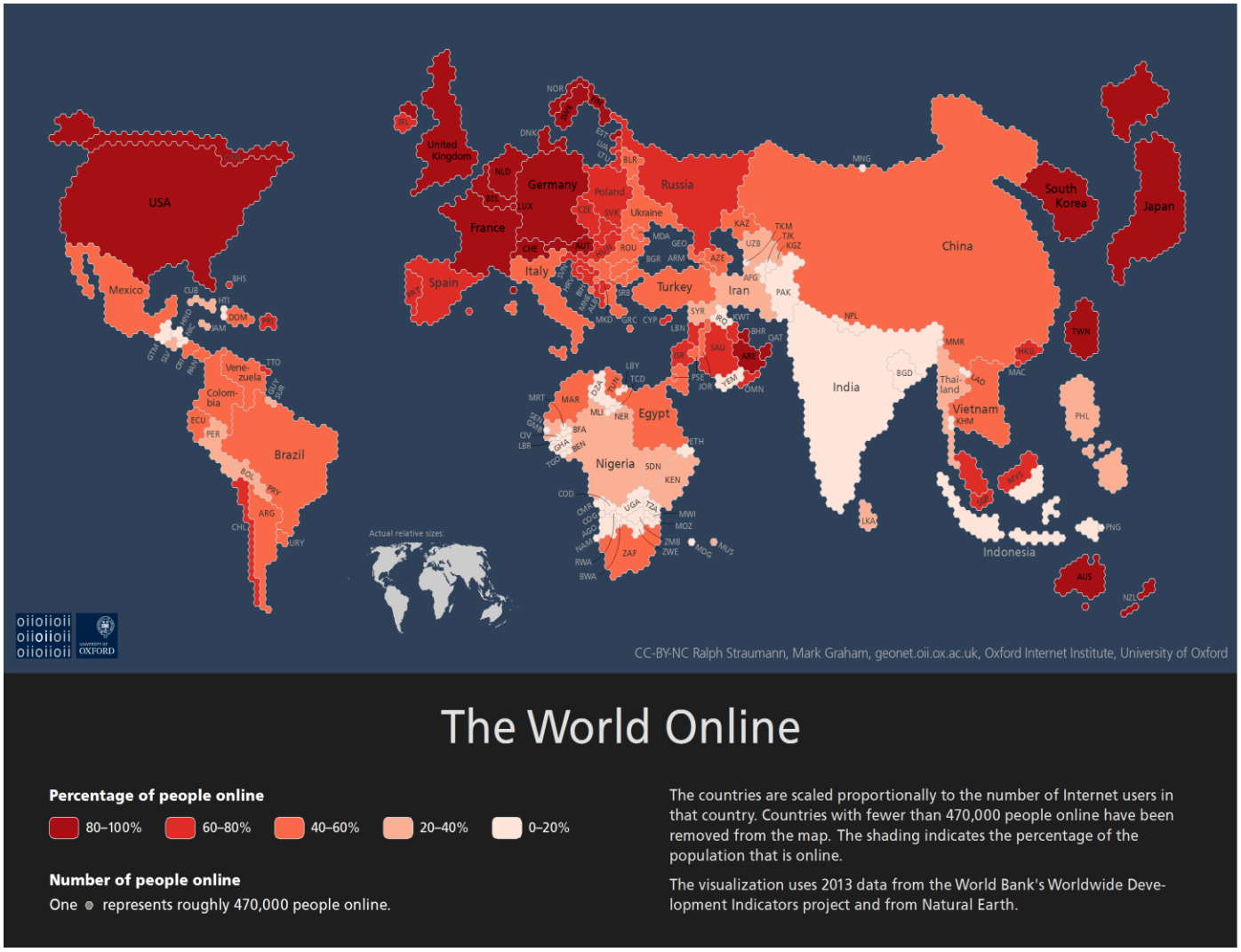


Classification of languages Conclusions

- About 100 «big» languages with more than 10M speakers
- About 100 languages with international or national diffusion and vitality
  - Lahnda/Panjabi: 82M speakers, but only EGIDS level 5
  - Latvian: 1.8M speakers, but EGIDS level 1
- About 100 languages with basic language-technological support
  - Luxemburgish: 400k speakers, Windows interface language
  - Yiddish: 1.5M speakers, Google Translate language

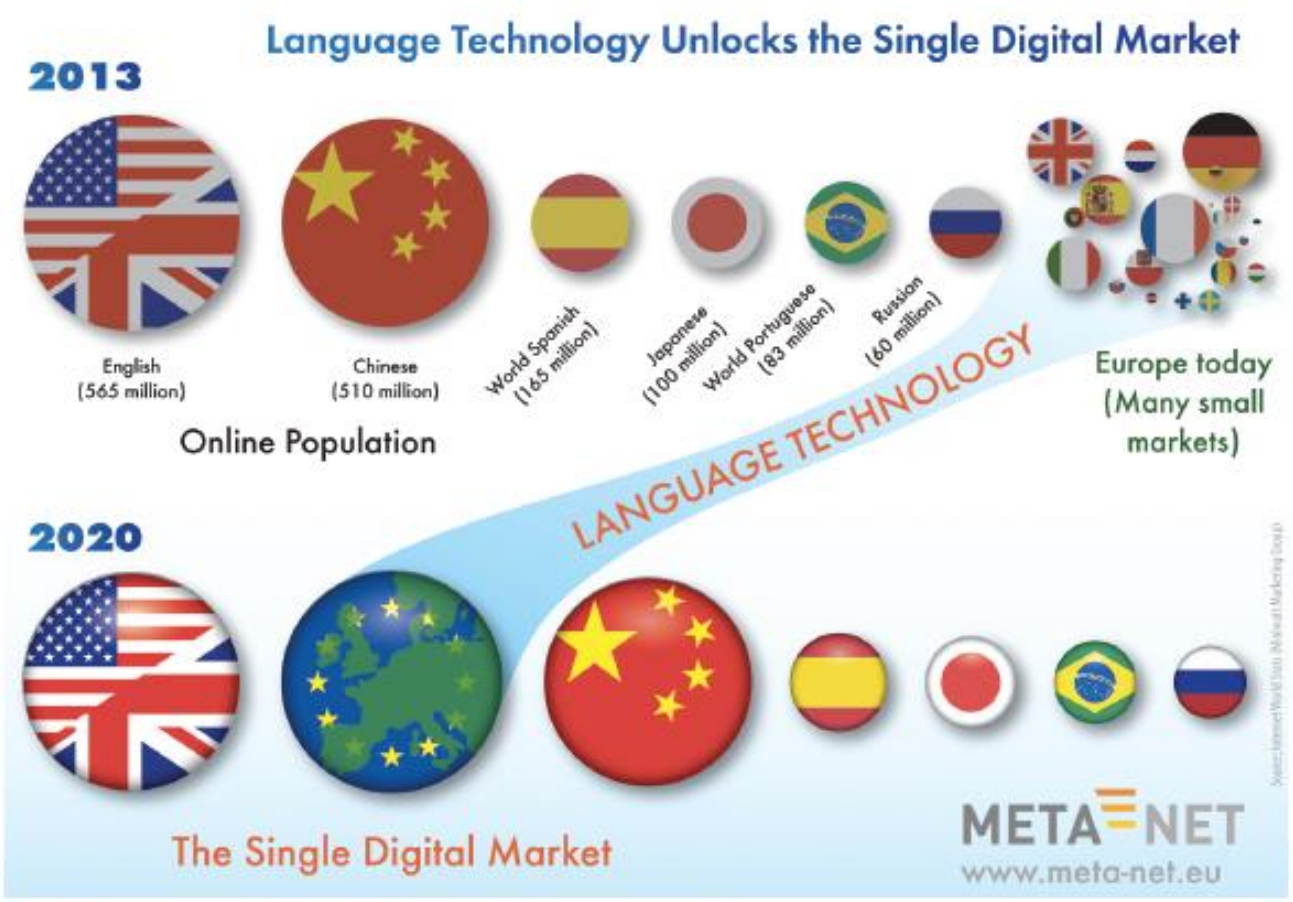
Why should we care about low-resource languages?

Commercial value



Graph from
Y. Tsvetkov

Commercial value



Social good

- See Tsvetov's slides for illustrations:
  - Translation systems to improve access to information
  - Speech interfaces to adapt to low literacy rates
  - Educational applications
  - Emergency response applications
  - Monitoring democratic processes
  - Help a language climb up the EGIDS ladder
  - Access to and preservation of culture

Linguistics

- Automatically annotate datasets to make them accessible for corpus-linguistic research
- Counteract bias towards English (and other Germanic/Romance languages) in LT research

What are resources in NLP?

How does their availability affect research in NLP?

(Utterly simplified) history of language technology

- 1980s, early 1990s:
  - Rule-based systems
  - A linguist writes rules on the basis of their own language competence and with the help of publications (dictionaries, grammars)
  - There are no fundamental differences between low-resource and high-resource languages
- 1990s:
  - Rule-based systems
  - Electronically available datasets (dictionaries) simplify the linguist's work
  - Low-resource languages: no available datasets – no simplification of the workflow

(Utterly simplified) history of language technology

- Since 2000:
  - Statistical systems + annotated corpora (= resources)
  - The systems automatically learn the regularities present in the corpora
  - Division of labor: linguists annotate corpora, computer scientists develop learning algorithms
    - But: nobody really likes to annotate corpora…
  - Low-resource languages are gradually disconnected from cutting-edge research because corpus creation does not follow
  - The problem of «low-resource languages» is relatively recent

NLP research (Sept 2017)

15 languages with most L1 speakers (Ethnologue)	Speakers (Mio.)	Papers in ACL Anthology	Datasets in LRE Map
Chinese [+ variants]	**1197**	**8480**	185
Spanish	**414**	4180	199
English	**335**	**24'000**	**1498**
Hindi	260	1420	49
Arabic [+ variants]	237	3400	148
Portuguese	203	1350	68
Bengali	193	473	14
Russian	167	1680	36
Japanese	122	5960	167
Javanese	84.3	35	0
Lahnda/Punjabi [+ variants]	82.6	15	5
German	78.2	6660	**269**
Korean	77.2	1820	10
French	75.0	**6910**	**305**
Telugu	74.0	216	5

http://www.ethnologue.com/statistics/size http://www.aclweb.org/anthology/ http://www.resourcebook.eu/searchll.php

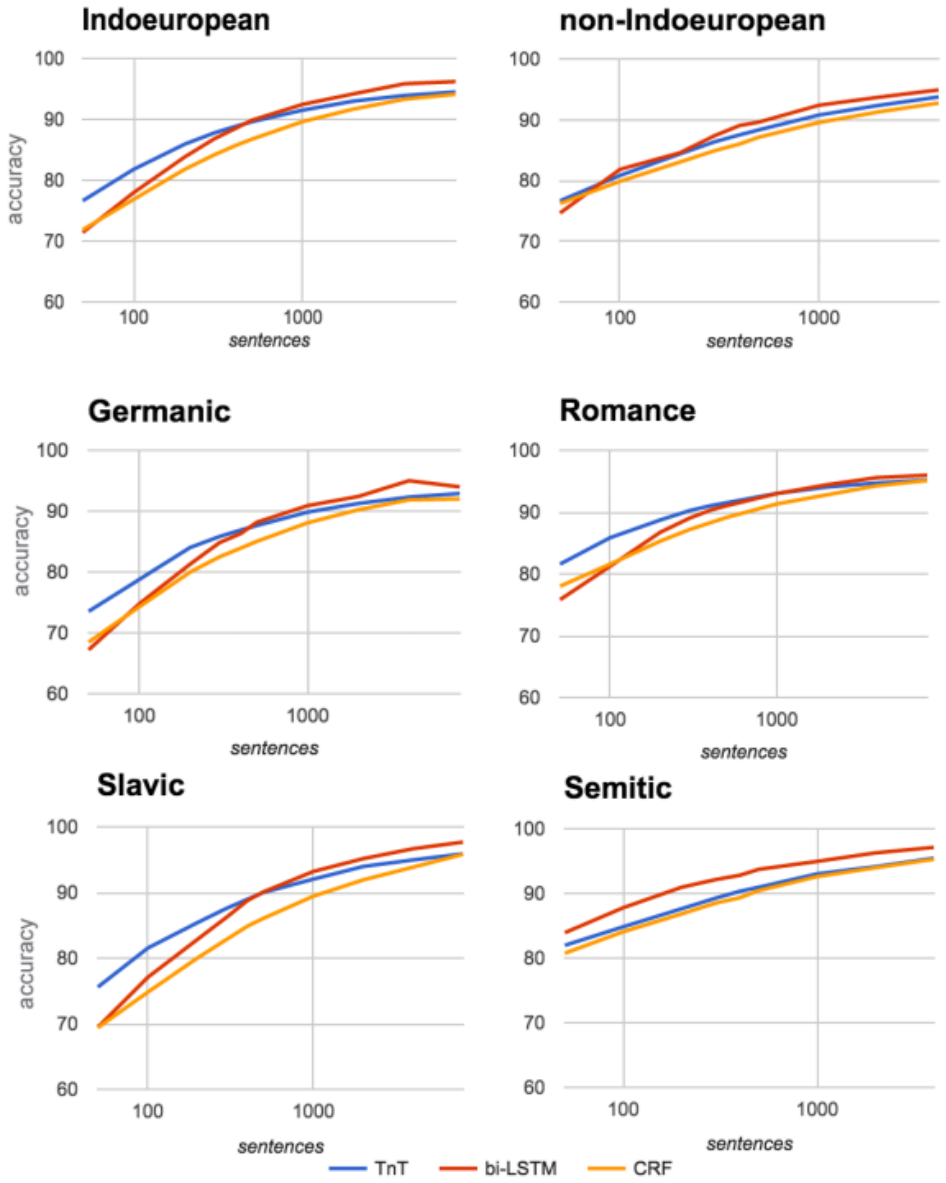
What resources for what applications?

- Supervised learning:
  - Training: examples + labels -> model
  - Testing: model + examples -> labels

- Rules of thumb:
  - The better the training examples, the better the model (but what is a “good” training corpus?)
  - The smaller the number of labels, the easier the task (i.e. the better the model)

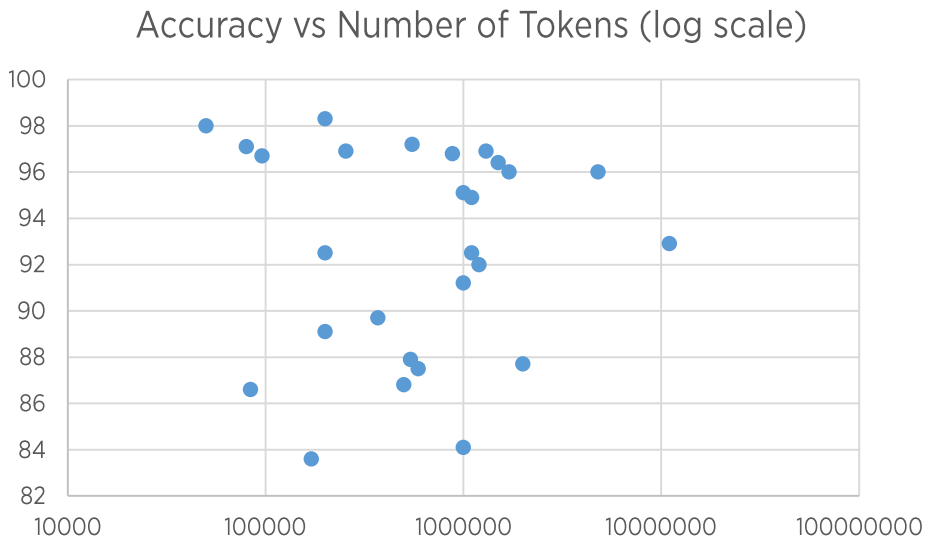
Part-of-speech tagging

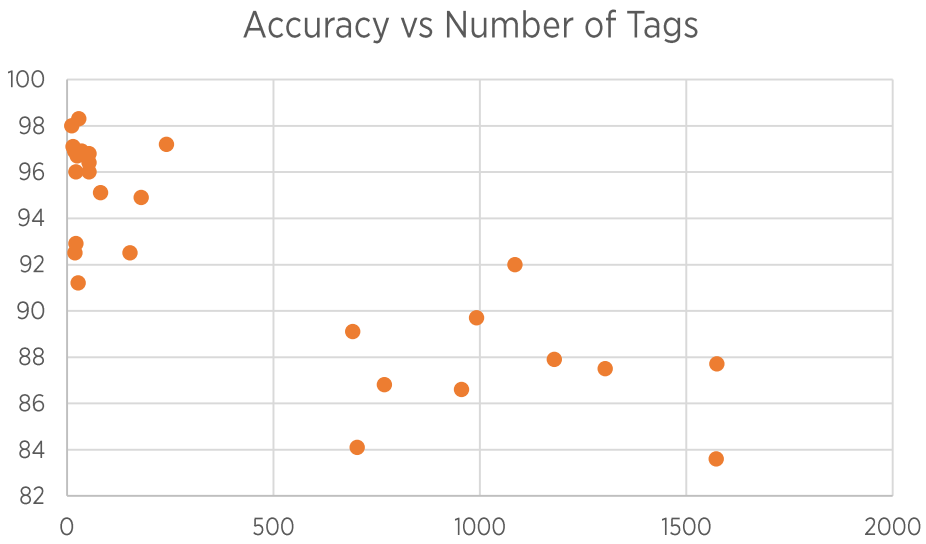
- Universal Tagset: 17 distinct labels
- 1000 training sentences ~ 90% accuracy



B. Plank, A. Søgaard, Y. Goldberg: *Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss*. ACL 2016.

Full morphological tagging





T. Horsmann, T. Zesch: *Do LSTMs really work so well for PoS tagging? A replication study.* EMNLP 2017.

Dependency parsing

- CoNLL 2017 Shared Task on Multilingual Dependency Parsing:
- 55 “big” treebanks
  - Training data > test data, dev data available
  - Best system average: 81.77 LAS
- 8 “small” treebanks
  - Training data <= test data, no dev data available
  - Best system average: 61.49 LAS
- 4 surprise languages
  - 20 sentences training/dev data per language
  - Best system average: 47.54 LAS

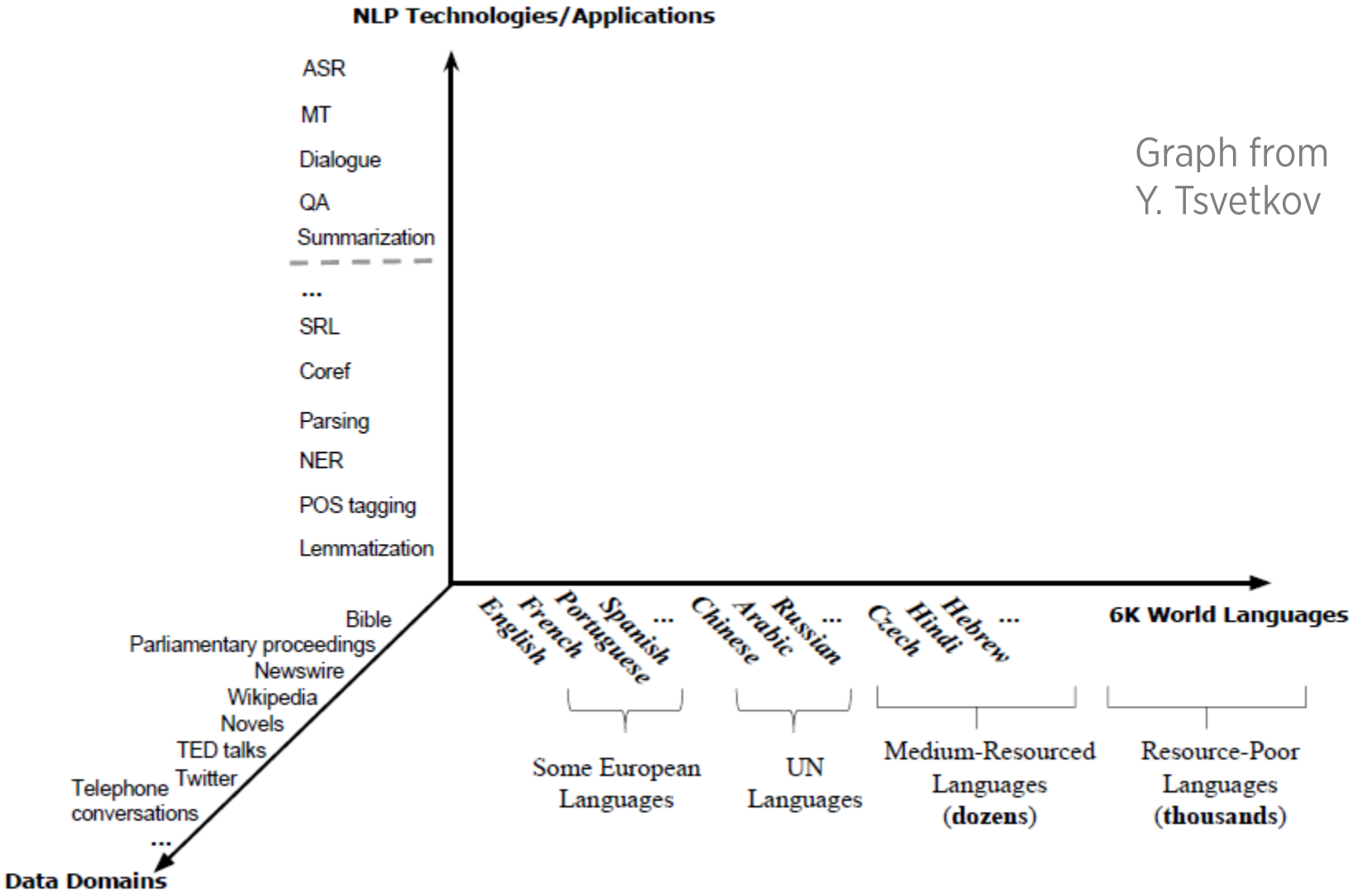
D. Zeman et al.: *CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. CoNLL 2017.

Statistical machine translation

- Data sizes and results from WMT 2017:

File	Size	CS-EN	DE-EN	FI-EN	LV-EN	RU-EN	TR-EN	ZH-EN
Europarl v7	628MB	✓	✓					
Europarl v8 2.5M sentences	238MB			✓	✓			
Common Crawl corpus	876MB	✓	✓			✓		
News Commentary v12	162MB	✓	✓			✓		✓
CzEng 1.6	3.1GB	✓						
Yandex Corpus	121MB					✓		
Wiki Headlines	9.1MB			✓		✓		
SETIMES2	44 MB						✓	
UN Parallel Corpus V1.0	3.6 GB					✓		✓
Rapid corpus of EU press releases	156 MB		✓	✓	✓			
LETA translated news	2 MB				✓			
Digital Corpus of European Parliament	123 MB				✓			
Online Books	309 kB				✓			
Best results X → EN (BLEU)		30.9	35.1	20.5	21.9	34.7	20.1	26.4
Best results EN → X (BLEU)		22.8	28.3	20.7	21.1	29.8	18.1	36.3

Task and domain dependence

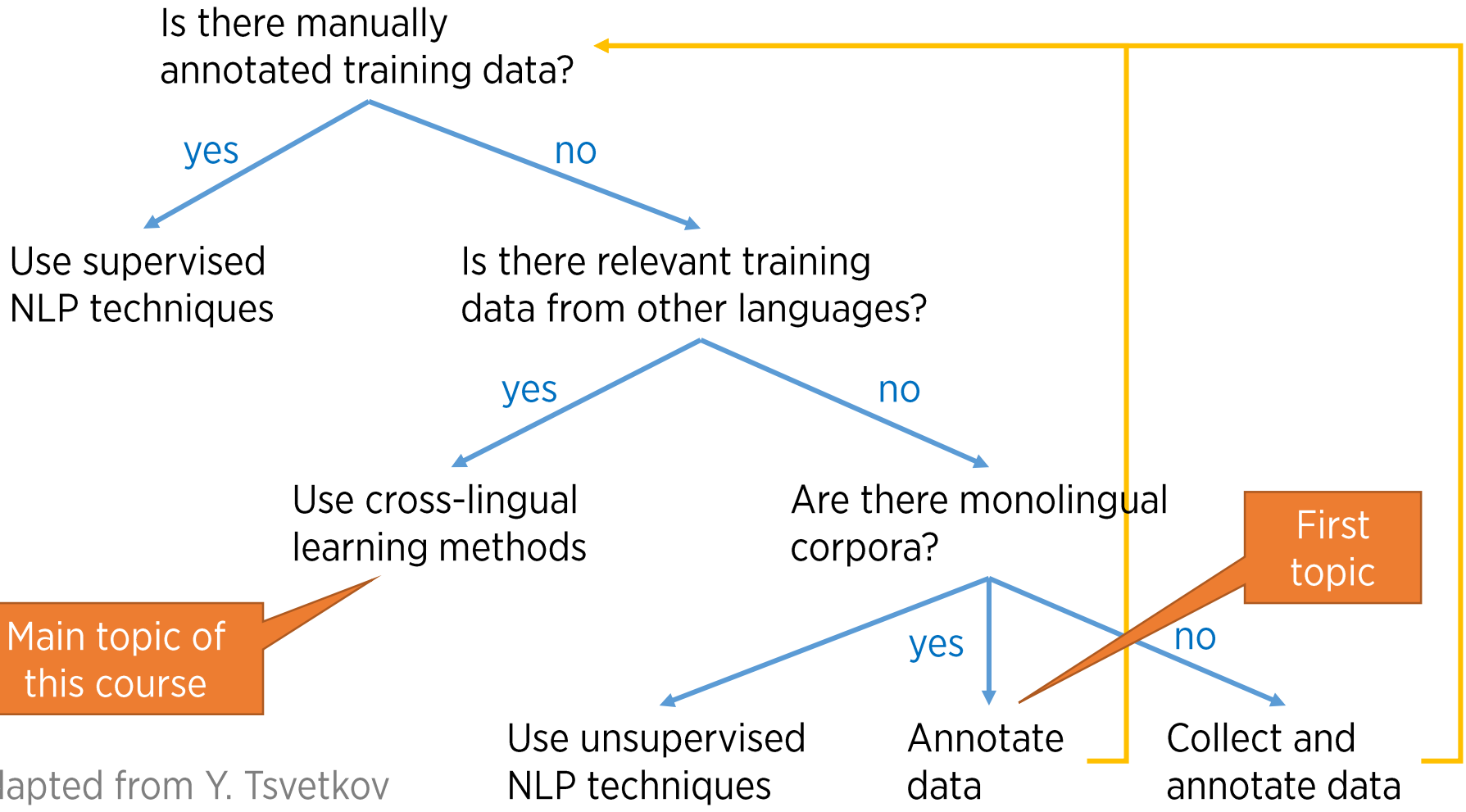


Graph from
Y. Tsvetkov

What resources for what applications?

- Resources are task and domain dependent
- A resource is rarely too big for NLP
- Rather than asking «How much data do I need?», one should ask «How bad can I allow my application to be?»

What resources can we get?



Adapted from Y. Tsvetkov

Tomorrow...

Readings

- Jörg Tiedemann & Željko Agić (2016): *Synthetic treebanking for cross-lingual dependency parsing*. JAIR 55. (Sections 1 and 2)
  - What is the difference between *model transfer* and *data transfer*?
  - What is the difference between *annotation projection* and *treebank translation*? What are the advantages of the latter according to the authors?
  - What are the problems regarding cross-lingual parsing *evaluation*?

Readings

- David Yarowsky & Grace Ngai (2001): *Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora.* Proceedings of NAACL-HLT.
  - What is word alignment?
  - What can go wrong when using direct projection (see Figure 2)? What correction techniques do the authors propose?
  - Annotation projection assumes that it is easier to obtain a parallel corpus (and a tagger for the high-resource language) than a directly annotated corpus for the low-resource language. Do you agree with this assumption?