Syntactic Parsing across Languages, treebanks, and Domains

Sara Stymne

Uppsala University

Feb. 14, 2024

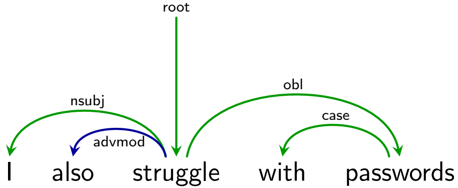
Overview

- Goal today: give an overview of research on dependency parsing across multiple:
  - Languages
  - Treebanks
  - Domains/genres
- Main focus on research 2017 and onwards
- This is one of my main research interests:
  - Going into details about my own work
  - Also trying to give a general overview of trends

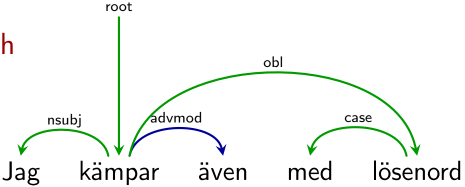
Intro

Languages have similarities

English

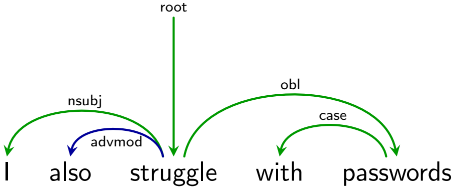


Swedish

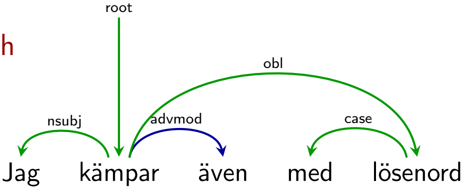


Languages have similarities

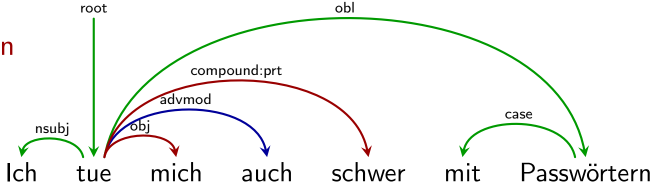
English



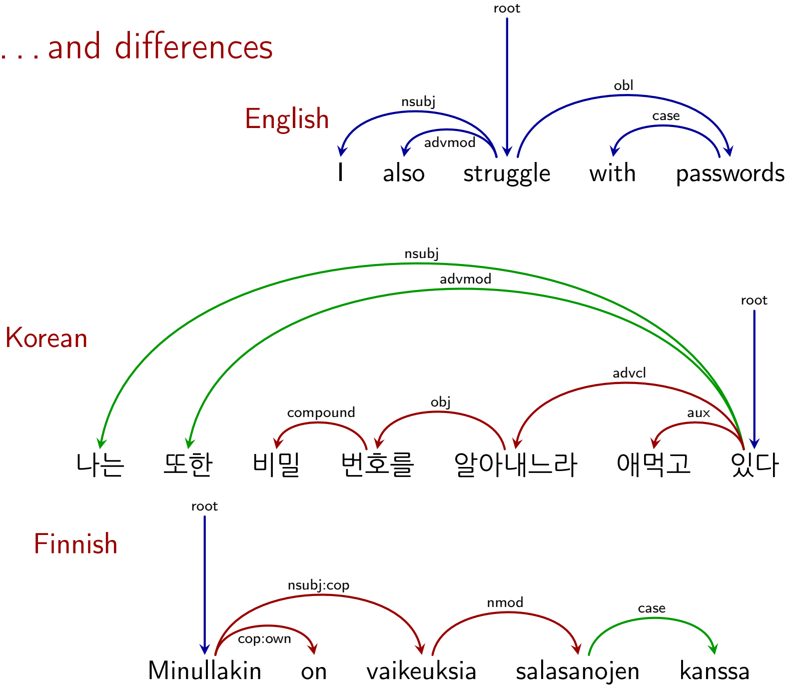
Swedish



German



. . . and differences



Multilingual parsing

- We can take advantage of language similarities!

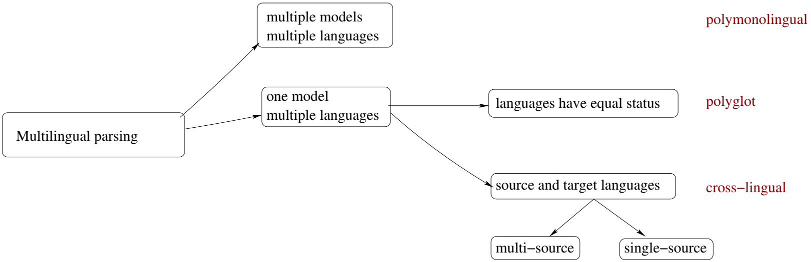


Figure by Miryam de Lhoneux

Cross-lingual parsing

- Popular in recent research
- Main purpose: improve parsing performance for a low-resource language by using data from another (related) language
  - Zero-shot
  - Few-shot
- Two main approaches:
  - Annotation transfer
  - Model transfer

Polyglot parsing

- Recently started to receive increased research interest
- Main purpose: improve parsing performance for a set of languages by using a joint model
- More diverse sets of languages:
  - Low-resource
  - Medium-resource
  - Large-resource(?)
- Main approach:
  - Joint training

Cross-Lingual Parsing Methods

- Data transfer
  - Annotation projection (Hwa et al., 2005)
  - Machine translate treebanks (Tiedemann et al., 2014)
- Joint models (with language embeddings) (Ammar et al., 2016; Smith et al., 2018)
- Models based on multilingual representations:
  - Part-of-speech tags (delelxicalized parsing, Zeman and Resnik (2008))
  - Cross-lingual word clusters (Täckström et al., 2012)
  - Cross-lingual embeddings (Ammar et al., 2016; Ahmad et al., 2019)
  - Multilingual LMs (Kondratyuk and Straka, 2019; Üstün et al., 2020)

Cross-Lingual Parsing: Target

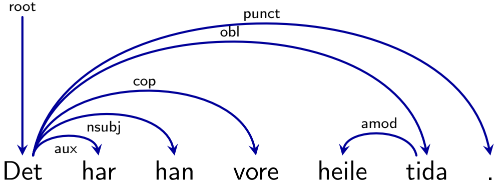
- Overall performance across a range of languages
  - UDify: trained on 75 languages (Kondratyuk and Straka, 2019)
  - UDapter: trained on 13 diverse languages, with typological features (Üstün et al., 2020)
- Performance for specific languages
  - 1 target, 1 source language (Vania et al., 2019)
  - 1 target + 3 source languages (Meechan-Maddon and Nivre, 2019)
  - Our work

Neural networks for cross-lingual and polyglot parsing

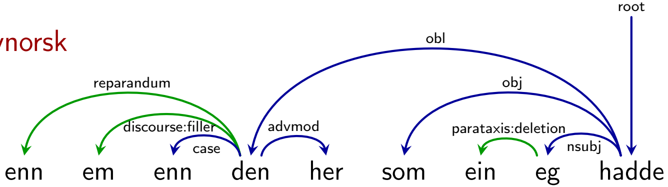
- Neural networks typically work well with multiple languages
- Cross-lingual systems can be viewed as multi-task systems
- Possible to share all or parts of an architecture
- Allows language representations as part of models
- Cross-lingual word embeddings an important resource

Within-language domain differences

Written Nynorsk



Spoken Nynorsk



Cross-domain parsing

- Even within a language, parsing can be affected by lack of data for some domain
- Cross-domain parsing can be approached as cross-lingual parsing
- Domain adaptation techniques
  - Few datasets with labeled data
  - Mainly unsupervised approaches

Cross-domain parsing

- Even within a language, parsing can be affected by lack of data for some domain
- Cross-domain parsing can be approached as cross-lingual parsing
- Domain adaptation techniques
  - Few datasets with labeled data
  - Mainly unsupervised approaches
- In this talk I will thus focus on cross-treebank parsing, partly covering domain differences

Parsing across different treebanks

Parsing with treebank embeddings

- I will now present our own work on treebank embeddings
- Add a represention of the treebank to each word
- An approach that works both across languages and treebanks
- Joint learning in a neural network setting
- Simple and effective!
- Stymne et al. (2018)
- Goal of this work: improve parsing for languages with multiple treebanks

Joint work


Miryam de Lhoneux


Aaron Smith


Joakim Nivre

Cross-Treebank Parsing Approaches

- Single treebank training
- Concatenation
- Concatenation + fine tuning
- Adversarial learning
- Treebank embeddings

Mono-treebank

- Train each treebank on its own
- Apply to each treebank's test data
- For extra test set, pick one of these models

Mono-treebank

- Train each treebank on its own
- Apply to each treebank's test data
- For extra test set, pick one of these models

- Simple, but does not take advantage of all available data
- Has separate models for each treebank

Concatenation

- Concatenate all training data from all treebanks for a language (Björkelund et al., 2017; Das et al., 2017)
- Use this model for all test sets from that language

Concatenation

- Concatenate all training data from all treebanks for a language (Björkelund et al., 2017; Das et al., 2017)
- Use this model for all test sets from that language

- Simple, but does not take the differences between treebanks into account
- Needs only one model for all treebanks

Concatenation + fine tuning

- Concatenate all training data from all treebanks for a language and train a joint model
- For each individual treebank, fine tune the joint model, by training more on only that treebank (Che et al., 2017, Shi et al., 2017)
- For extra test set, pick one of these models

Concatenation + fine tuning

- Concatenate all training data from all treebanks for a language and train a joint model
- For each individual treebank, fine tune the joint model, by training more on only that treebank (Che et al., 2017, Shi et al., 2017)
- For extra test set, pick one of these models

- Needs more training than previous suggestion
- Has separate models for each treebank

Adversarial learning

- Proposed for this scenario by Sato et al. (2017)
- Use an adversarial task of treebank identification during training
- Use both treebank-specific structures and a shared structure for the adversarial task

Adversarial learning

- Proposed for this scenario by Sato et al. (2017)
- Use an adversarial task of treebank identification during training
- Use both treebank-specific structures and a shared structure for the adversarial task

- Quite complex architecture
- Needs only one model for all treebanks, but a treebank representation for input sentences

Adversarial learning

- Proposed for this scenario by Sato et al. (2017)
- Use an adversarial task of treebank identification during training
- Use both treebank-specific structures and a shared structure for the adversarial task

- Quite complex architecture
- Needs only one model for all treebanks, but a treebank representation for input sentences
- Not explored in this work, but shown to give some gains

Treebank embeddings

- We can apply language embeddings to the monolingual case, getting "treebank embeddings"
- Treebank embeddings can learn to represent important differences between treebanks in the same language
- This model can also easily be extended to include more languages

Treebank embeddings

- We can apply language embeddings to the monolingual case, getting "treebank embeddings"
- Treebank embeddings can learn to represent important differences between treebanks in the same language
- This model can also easily be extended to include more languages

- Simple, and takes the differences between treebanks into account
- Needs only one model for all treebanks, but a treebank representation for input sentences

Cross-treebank parsing approaches

- Comparison of different approaches:

Approach	Number models	Simple	Sensitive to Differences	Pools data
Mono-treebank	Many	Yes	Yes	No
Concatenation	1	Yes	No	Yes
Concat+fine tuning	Many	No	Yes	Yes
Adversarial learning	1	No	Yes	Yes
TB embeddings	1	Yes	Yes	Yes

Proxy Treebanks

- For all methods, except concatenation, we need to define which treebank an input sentence comes from (at test time)
- We call this a **proxy** treebank
  - single/concat+ft: for choosing a model
  - tb-emb: for setting a treebank embedding

Experiments

- 9 languages with at least two UD training treebanks + PUD
- Comparing four methods for handling multiple treebanks
- BiLSTM-based transition-based dependency parser (de Lhoneux et al., 2017)
- Using UD version 2.1 treebanks
- All results are shown as LAS scores

UUparser

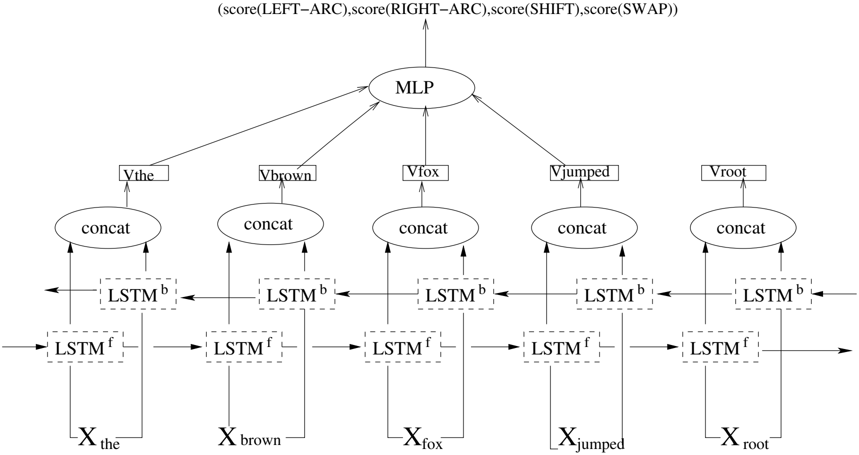


Figure by Miryam de Lhoneux

Overall results – matching test sets

Language	Treebank	Size	mono	concat	c+ft	tb-emb
Czech	PDT	68495	86.7	87.5^{+}	**88.3***	87.2^{+}
	CAC	23478	86.0	87.8^{+}	88.1^{+}	**88.5**$^{+}$
	FicTree	10160	84.3	89.3^{+}	**89.5**$^{+}$	89.2^{+}
	CLTT	860	72.5	86.2^{+}	**86.9**$^{+}$	86.0^{+}
English	EWT	12543	82.2	82.1	82.5	**83.0**
	LinES	2738	72.1	76.7^{+}	**77.3**$^{+}$	**77.3**$^{+}$
	ParTUT	1781	80.5	83.5^{+}	85.4^{+}	**85.7**$^{+}$
Finnish	FTB	14981	76.4$^{\times}$	74.4	80.1*	**80.6***
	TDT	12217	78.1$^{\times}$	70.6	**80.6***	80.3*
French	FTB	14759	83.2	83.2	83.9*	**84.1***
	GSD	14554	84.5	84.1	85.3	**85.6**$^{\times}$
	Sequoia	2231	84.0	86.0^{+}	**89.8***	89.1*
	ParTUT	803	79.8	80.5	89.1*	**90.3***
Italian	ISDT	12838	87.7	**87.9**	87.7	87.6
	PoSTWITA	2808	71.4	76.7^{+}	76.8^{+}	**77.0**$^{+}$
	ParTUT	1781	83.4	89.2^{+}	**89.3**$^{+}$	88.8^{+}
Portuguese	GSD	9664	88.3	87.3	89.0*	**89.1***
	Bosque	8331	84.7	84.2	86.2$^{\times}$	**86.3***
Russian	SynTagRus	48814	90.2$^{\times}$	89.4	**90.4**$^{\times}$	**90.4**$^{\times}$
	GSD	3850	74.7$^{\times}$	73.4	79.8*	**80.8***
Spanish	AnCora	14305	87.2$^{\times}$	86.2	87.5$^{\times}$	**87.6**$^{\times}$
	GSD	14187	84.7	83.0	85.8$^{\times}$	**86.2***
Swedish	Talbanken	4303	79.6	79.1	80.2	**80.6**$^{\times}$
	LinES	2738	74.3	76.8	**77.3**$^{+}$	77.1^{+}
Average			81.4	82.7^{+}	**84.9***	**84.9***

Overall results - PUD sets

PUD: parallel dataset without any training data

Language	mono	concat	c+ft	tb-emb
Czech	**81.7**	**81.7**	81.6	81.2
English	80.7	80.0	81.7*	**81.9***
Finnish	78.6×	73.0	**81.3***	80.9*
French	79.1	79.4	80.2*	**80.3***
Italian	77.4	86.0	85.8+	**86.1**+
Portuguese	75.2	76.8+	77.5+	**77.6**+
Russian	70.1×	68.7	77.6*	**78.0***
Spanish	79.8	79.9	80.8+	**80.9***
Swedish	70.3	72.0+	73.2*	**73.6***
Average	77.9	77.5	80.0*	**80.1***

Extension to cross-lingual parsing

- Use treebank embeddings for treebanks from more than one language
- Typically works better for closely related languages
- Open questions:
  - Language mix
  - Model size

What about genre/domain?

Cross-Lingual Parsing across Domains

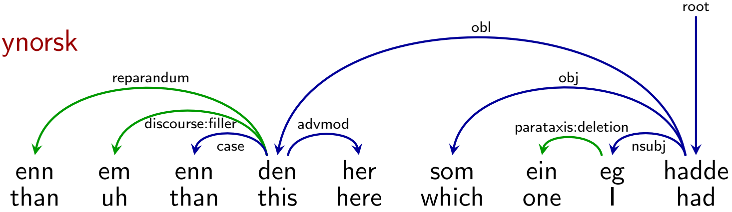
- Stymne (2020) *Cross-Lingual Domain Adaptation for Dependency Parsing*. Workshop on Treebanks and Linguistic Theories (TLT)
- Improve dependency parsing for specific text types:
  - Twitter
  - Transcribed speech

Cross-Lingual Parsing across Domains

- Stymne (2020) *Cross-Lingual Domain Adaptation for Dependency Parsing*. Workshop on Treebanks and Linguistic Theories (TLT)
- Improve dependency parsing for specific text types:
  - Twitter
  - Transcribed speech
- By treebank combination:
  - In-language out-of-domain data
  - In-domain data from other languages

Example: Transcribed Speech

Spoken Nynorsk



Experiments

- Languages
  - **Speech**: French, Norwegian, and Slovenian //Low-resource: Naija and Komi-Zyrian
  - **Twitter**: English, Italian, and Hindi–English code-switching
- Labelled attachment score for evaluation
- More results in the paper

Combining treebanks

Same L IND	Same L OOD	Other L IND	Other L OOD	Spoken Fr	Spoken No	Spoken Sl	Twitter It	Twitter En	Twitter HiEn	**Mean**
–	X	–	–	63.4	52.8	46.9	62.8	55.7	25.0	51.1
–	X	–	X	64.3	**54.4**	47.6	63.4	54.6	24.9	51.5
–	X	X	–	**64.5**	52.0	**52.7**	**65.5**	**58.9**	**25.7**	**53.2**

Combining with matching treebanks

Same L		Other L		Spoken			Twitter			
				Fr	No	Sl	It	En	HiEn	**Mean**
IND	OOD	IND	OOD							
–	X	–	–	63.4	52.8	46.9	62.8	55.7	25.0	51.1
–	X	–	X	64.3	**54.4**	47.6	63.4	54.6	24.9	51.5
–	X	X	–	**64.5**	52.0	**52.7**	**65.5**	**58.9**	**25.7**	**53.2**
X	–	–	–	76.6	74.3	65.8	82.3	74.7	65.0	73.1
X	–	X	–	76.1	73.9	65.3	81.8	76.3	64.1	72.9
X	X	–	–	**84.0**	78.3	71.8	84.2	**82.8**	**67.6**	78.1
X	X	X	–	83.7	**78.7**	**72.7**	**84.5**	82.1	67.2	**78.2**

Low-resource languages

	Related OOD		Related OOD + other IND	
	Interp	Ensemble	Interp	Ensemble
Komi Zyrian	14.8	18.4	**19.0**	18.7
Naija	28.0	27.4	**30.0**	28.3

Discussion

- Combining treebanks across languages and domains is feasible
- Small, but quite consistent gains from adding in-domain treebanks from other languages

Discussion

- Combining treebanks across languages and domains is feasible
- Small, but quite consistent gains from adding in-domain treebanks from other languages
- These experiments were performed with a somewhat old RNN-based parser
  - Müller-Eberstein et al. (2021) also suggests that matching data for genre across languages is useful, with an mBERT-based parser
  - We are currently working on this
    - Tentative results: in-genre data often helps, but mainly in combination with other genres as well
    - In-language data more important than in-genre data
    - UD-MULTIGENRE: variant of UD split into genre-specific subset (Danilova and Stymne, 2023)

Transfer Language Choice

Cross-Lingual Parsing Targeting a Specific Language

- **Problem**: Which language(s) to transfer from?
- Common strategy: Select a language that belongs to the same language family or has a small phylogenetic distance in the language family tree to the task language (Cotterell and Heigold, 2017; Dehouck and Denis, 2019; Meechan-Maddon and Nivre, 2019; Vania et al., 2019)

Cross-Lingual Parsing Targeting a Specific Language

- **Problem**: Which language(s) to transfer from?
- Common strategy: Select a language that belongs to the same language family or has a small phylogenetic distance in the language family tree to the task language (Cotterell and Heigold, 2017; Dehouck and Denis, 2019; Meechan-Maddon and Nivre, 2019; Vania et al., 2019)
- Not all languages have a closely related language with a treebank
- Not all languages in a single language family share the same linguistic properties

Options for Transfer Language Choice

- Some strategies explored in our work
- de Lhoneux et al. (2017a):
  - Genetic distance
  - Geographical closeness
  - Sharing the same script
  - Dev performance in a zero-shot setting
- Smith et al. (2018):
  - Genetic distance
  - Clustering treebank/language embeddings from a small model trained on all available training languages
- Stymne (2020)
  - Matching domain/genre

Systematic Transfer Language Choice

- Lin et al. (2019) *Choosing Transfer Languages for Cross-Lingual Learning*. ACL
- Investigate the impact of different factors on transfer language choice
- Create a ranker, LangRank, for ranking transfer languages based on these features
- Apply this to four NLP tasks
  - Machine translation (joint training)
  - POS-tagging (joint training)
  - Entity linking (zero shot)
  - Dependency parsing (zero shot)

Features

- **Dataset features:**
  - Dataset size, type-token ratio, word and subword overlap
- **Linguistic Distances:** based on the URIEL typological database (Littell et al., 2017) information-rich vector identifications of languages drawn from typological, geographical, and phylogenetic databases:
  - WALS (Dryer and Haspelmath, 2013)
  - Ethnologue (Lewis, 2009)
  - Glottolog (Nordhoff and Hammarström, 2011)
  - PHOIBLE (Moran and McCloy, 2014)

Linguistic Distances

- **Geographic distance** (d_{geo}): The spherical distance among languages on Earth's surface, mainly based on abstractions of locations from Glottolog
- **Genetic distance** (d_{gen}): The genealogical distance among languages, based on the world language family tree from Glottolog
- Cosine distance of feature vectors:
  - **Phonological distance** (d_{pho}): Phonological vectors from WALS and Ethnologue
  - **Inventory distance** (d_{inv}) Phonological vectors from PHOIBLE
  - **Syntactic distance** (d_{syn}): Syntactic vectors from WALS
  - **Featural distance** (d_{fea}): Combinations of all other feature vectors

Transfer Language Choice as a Ranking Problem

	Method	MT	EL	POS	DEP
dataset	word overlap o_w	28.6	30.7	13.4	52.3
	subword overlap o_{sw}	29.2	–	–	–
	size ratio s_{tf}/s_{tk}	3.7	0.3	9.5	24.8
	type-token ratio d_{ttr}	2.5	–	7.4	6.4
ling. distance	genetic d_{gen}	24.2	50.9	14.8	32.0
	syntactic d_{syn}	14.8	46.4	4.1	22.9
	featural d_{fea}	10.1	47.5	5.7	13.9
	phonological d_{pho}	3.0	4.0	9.8	43.4
	inventory d_{inv}	8.5	41.3	2.4	23.5
	geographic d_{geo}	15.1	49.5	15.7	46.4
	LANGRANK (all)	51.1	**63.0**	**28.9**	**65.0**
	LANGRANK (dataset)	**53.7**	17.0	26.5	**65.0**
	LANGRANK (URIEL)	32.6	58.1	16.6	59.6

Average Normalized discounted cumulative gain @3
From (Lin et al., 2019, p. 3130)

Example Decision Tree

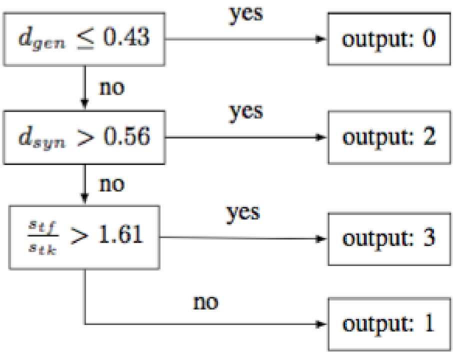


Figure 4: An example of the decision tree learned in the machine translation task for Galician as task language.

From Lin et al. (2019, p. 3132)

Going Beyond Parsing

- Fine-tuning large multilingual LMs useful across many tasks
  - NLI, QA, Paraphrases, semantic similarity, NER, POS, parsing, ...
  - Devlin et al. (2019); Wu and Dredze (2019); Lauscher et al. (2020) ...
- Typical transfer language: English
  - Mainly due to the availability of training data for many tasks
- Recent discussion of this choice:
- Lauscher et al. (2020)
  - Some tendency for structurally similar languages to transfer best
- Turc et al. (2021)
  - Across tasks, German and Russian tend to be better than English, even when machine-translated from En

Uppsala at CoNLL Shared Task, 2018

CoNLL Shared task 2018

- Shared task on multilingual dependency parsing from raw text to universal dependencies
- Used the UD data, with multiple treebanks for many languages

CoNLL Shared task 2018

- Shared task on multilingual dependency parsing from raw text to universal dependencies
- Used the UD data, with multiple treebanks for many languages
- Most teams trained a parser per treebank
- Some teams suggested more advanced strategies, but none did any comparison between methods
- Some teams employed cross-lingual strategies (mainly to small treebanks)

UUparser

- BiLSTM-based feature extractor (Kiperwasser and Goldberg, 2016)
- Transition-based (and graph-based)
  - Arc-hybrid + SWAP
  - Static-dynamic oracle
- Cross-lingual models
  - With language/treebank embeddings
- de Lhoneux et al. (2017b); Smith et al. (2018)

UUp@CoNLL'18 Shared Task

- 82 treebanks, 34 models
- Multilingual models with small groups of languages
- Grouped languages based on:
  - Relatedness
  - Clustering of treebank embeddings
- Comparison with a monolingual model
- Metric: LAS

Treebank size	Mono	TB embeddings	Diff
Big	79.6	80.3	+0.7
Small	60.1	63.6	+3.5
Low-resource	17.7	25.5	+7.8
All	70.7	72.3	+1.6

CoNLL 2018, Scandinavian languages

Treebank	Mono	TB embeddings	Diff	
Danish	79.7	80.1	+0.4	
Norwegian BM	87.7	88.3	+0.6	
Norwegian NN	86.2	87.4	+1.2	
Norwegian NN Spoken	55.5	59.7	+4.2	
Swedish TB	83.3	84.3	+1.0	
Swedish LinES	78.3	80.5	+2.2	
Swedish PUD	75.5	78.2	+2.7	
Faroese	40.0	41.7	+1.7	Zero-shot

CoNLL 2018 sample of languages

Treebank	Mono	TB embeddings	Diff
Russian	89.4	89.0	-0.4
Russian	59.3	65.5	+6.2
Ukraine	81.4	82.7	+1.3
Persian	83.2	83.4	+0.2
Kurmanji	7.6	29.5	+21.9
Ancient Greek	63.0	65.2	+2.2
Ancient Greek	71.6	72.2	+0.6
Gothic	60.6	63.4	+2.8
Latin	82.6	83.0	+0.4
Latin	49.9	58.3	+8.4
Latin	63.9	64.1	+0.2
Old Church Slavonic	70.3	70.4	+0.1

Discussion

- Training in groups of languages typically helped
  - More for languages with little data
  - Often also smaller gains for languages with more data
- Preliminary experiments showed that it was better to use smaller groups of closer languages, than larger groups

Discussion

- Training in groups of languages typically helped
  - More for languages with little data
  - Often also smaller gains for languages with more data
- Preliminary experiments showed that it was better to use smaller groups of closer languages, than larger groups
- Later work shows that later transformer-based parsers may work as well with massively multilingual training, as with smaller designed language groups (van der Goot and de Lhoneux, 2021)

More about Language Choice

What about more diverse languages?

- Yifei Zhang (2021) *The Influence of M-BERT and Sizes on the Choice of Transfer Languages in Parsing*. Master thesis, Uppsala.
- Explores correlations with linguistic distances from URIEL, investigating:
  - mBERT versus randomly initialized embeddings
  - Influence of training data size
- UUparser variant (Attardi et al., 2020), with embeddings from mBERT

Languages

- Target languages:
  - Afrikaans, Greek, Vietnamese
  - 10K training tokens
- Transfer languages:
  - Czech, Dutch, French, German, Ancient Greek, Arabic, Urdo, Bulgarian, Russian, Hebrew, Chinese, Japanese, Korean, Hindi
  - 100K training tokens

Languages

- Target languages:
  - Afrikaans, Greek, Vietnamese
  - 10K training tokens
- Transfer languages:
  - Czech, Dutch, French, German, Ancient Greek, Arabic, Urdo, Bulgarian, Russian, Hebrew, Chinese, Japanese, Korean, Hindi
  - 100K training tokens

	af_afribooms			**el_gdt**			**vi_vtb**		
	rd	**mb**	diff	**rd**	**mb**	diff	**rd**	**mb**	diff
Monolingual	63.76	68.56	4.8	70.91	75.78	4.87	49.58	43.98	-5.6

Joint Learning Experiments

	af_afribooms			el_gdt			vi_vtb		
	rd	**mb**	diff	**rd**	**mb**	diff	**rd**	**mb**	diff
Monolingual	63.76	68.56	4.8	70.91	75.78	4.87	49.58	43.98	-5.6
nl_alpino	77.97	80.37	2.40	78.71	82.78	4.07	67.14	68.40	1.26
de_gsd	74.75	79.56	**4.81**	77.86	82.68	4.82	65.47	67.78	2.31
cs_pdt	75.43	79.92	4.49	79.44	84.48	**5.04**	66.72	69.06	**2.34**
fr_gsd	**78.45**	**81.85**	3.40	**82.23**	**85.85**	3.62	**69.57**	**70.95**	1.38
ar_padt	71.70	74.07	2.37	73.94	78.22	4.28	62.49	63.98	1.49
ur_udtb	72.32	74.57	2.25	74.22	77.18	2.96	62.95	61.26	-1.69
ru_syntagrus	74.34	78.95	**4.61**	77.78	83.21	**5.43**	65.25	66.81	1.56
bg_btb	77.16	**80.71**	3.55	80.77	84.91	4.14	68.11	**69.52**	1.40
he_htb	73.81	75.78	1.97	76.45	79.02	2.57	64.43	64.25	-0.18
ko_kaist	75.33	77.54	2.21	77.15	81.57	4.42	65.28	63.77	-1.51
ja_gsd	**79.23**	80.37	1.14	**82.83**	**85.04**	2.21	**71.31**	68.05	-3.26
zh_gsd	69.82	69.07	-0.75	72.24	71.33	-0.91	61.27	58.42	-2.85
hi_hdtb	76.06	79.37	3.31	78.42	82.72	4.3	61.26	67.42	**6.16**
grc_proiel	70.42	69.32	-1.1	72.41	72.31	-0.11	60.69	55.45	-5.24
AVERAGE	74.77	77.24	2.47	77.46	80.81	3.35	65.14	65.36	0.22

Correlations with linguistic distances

		d_{geo}	d_{gen}	d_{inv}	d_{syn}	d_{pho}	d_{fea}
af	rd	-0.3998	0.0207	**-0.6443**	0.086	0.598	**-0.4536**
	mb	**-0.4097**	-0.2067	**-0.8089**	-0.1014	0.6197	**-0.6789**
el	rd	**-0.4351**	-0.1921	**-0.6222**	0.0019	**-0.5156**	**-0.429**
	mb	**-0.5316**	-0.0342	**-0.6094**	**-0.5999**	**-0.5746**	**-0.6188**
vi	rd	-0.168	–	-0.1944	-0.3067	**-0.4769**	-0.2654
	mb	-0.2547	–	**-0.482**	-0.036	-0.0901	**-0.5639**

Correlations, variations with size

mBERT Joint

		d_{geo}	d_{gen}	d_{inv}	d_{syn}	d_{pho}	d_{fea}
af	all	**-0.4097**	-0.2067	**-0.8089**	-0.1014	**0.6197**	**-0.6789**
	half	-0.2732	-0.2108	**-0.6966**	-0.1412	**0.6291**	**-0.5791**
el	all	**-0.5316**	-0.0342	**-0.6094**	**-0.5999**	**-0.5746**	**-0.6188**
	half	**-0.4777**	0.3	**-0.7217**	-0.1833	**-0.5678**	**-0.5201**
vi	all	-0.2547	–	**-0.482**	-0.036	-0.0901	**-0.5639**
	half	-0.2096	–	**-0.4589**	-0.1488	-0.1646	**-0.5426**

Conclusion

- Joint parsing
  - Nearly all transfer languages lead to improvements over monolingual baseline in all settings
  - Some languages, e.g. French, transfer well to all target languages
- Transfer language choice shows some variation based on
  - Zero-shot versus joint
  - Target language
  - Embedding type
  - Relatively stable across training set sizes

Wrapping up

Summary

- An increasing interest in cross-lingual and polyglot parsing
- Much research focused on low-resource scenarios
- I mainly discussed our work, based on UUparser with treebank embeddings
  - Can be used for both cross-treebank and multilingual parsing
  - Simpler than many other proposed methods
  - No external resources or processing needed
  - Gives good results both with small and large treebanks
  - Could potentially be extended to domains

Current trends

- This lecture mainly focused on my research
- A lot of other work on multilingual parsing
- The overall dominating parsing algorithm right now is graph-based parsing, CLU-algorithm, on top of fine-tuning an LM
  - This works well in a multilingual setting, based on a multilingual LM (e.g. mBERT, XLM-R)
- Many current state-of-the-art tools are general-purpose fine-tuning toolkits, like Trankit (Nguyen et al., 2021) or Machamp (van der Goot et al., 2021)

Practicalities

Coming up

- Monday, Feb. 19: supervision
- Wednesday, Feb. 21: lecture on Earley's algorithm
  - Recorded lectures + exercise available
- Deadlines:
  - Assignment 2: Feb. 22
  - Project proposal: Feb 26
  - Assignment 3: March 4
  - Seminar 2: March 4

Assignment 3

- In assignment 3, you will use UUparser with treebank embeddings
  - Based on the Kiperwasser and Goldberg (2016) parser that we will discuss in seminar 2
  - No multilingual signal, so you will only explore it in a few-shot setting (with some target language data)
  - Allows experiment to run on our Linux cluster, on CPUs
- Compare two transfer languages you think are good or bad for a chosen target
- Try out some different types of evaluation and error analysis

Project

- Project should have a practical component, e.g. implementation or empirical study
- You also need to connect it to at least one research paper
- Common projects
  - Implement Earley's algorithm
  - Cross-lingual dependency parsing: extension of assignment 3
- Also other ideas available, or propose your own project
- Individual or pair projects
  - Sign up to a group in Studium
  - If you want to work in a pair: you need to find a partner yourself
  - Do not sign up with a peer unless you have decided to work together

Project proposal

- Due February 26
- Around 1/2 A4-page, describing your project plan
- Main purposes:
  - Get you started on your projects
  - Allow Sara to do feasibility assessments of your project ideas
- In case your plans change for some reason after handing in the proposal – get in touch with Sara to discuss the potential change

Final project seminar

- Discuss your project in smaller groups
- No slides of formal presentations
- Students working in pairs present independently
- We will move the final seminar
  - Suggestion: March 25, 9–12

References

References I

Ahmad, W., Zhang, Z., Ma, X., Hovy, E., Chang, K.-W., and Peng, N. (2019). On difficulties of cross-lingual transfer with order differences: A case study on dependency parsing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2440–2452, Minneapolis, Minnesota. Association for Computational Linguistics.

Ammar, W., Mulcaire, G., Ballesteros, M., Dyer, C., and Smith, N. A. (2016). Many languages, one parser. *Transactions of the Association for Computational Linguistics*, 4:431–444.

References II

Attardi, G., Sartiano, D., and Simi, M. (2020). Linear neural parsing and hybrid enhancement for enhanced Universal Dependencies. In *Proceedings of the 16th International Conference on Parsing Technologies and the IWPT 2020 Shared Task on Parsing into Enhanced Universal Dependencies*, pages 206–214, Online. Association for Computational Linguistics.

Cotterell, R. and Heigold, G. (2017). Cross-lingual character-level neural morphological tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 748–759, Copenhagen, Denmark. Association for Computational Linguistics.

Danilova, V. and Stymne, S. (2023). UD-MULTIGENRE – a UD-based dataset enriched with instance-level genre annotations. In Ataman, D., editor, *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 253–267, Singapore. Association for Computational Linguistics.

References III

de Lhoneux, M., Shao, Y., Basirat, A., Kiperwasser, E., Stymne, S., Goldberg, Y., and Nivre, J. (2017a). From raw text to universal dependencies - look, no tags! In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 207–217, Vancouver, Canada. Association for Computational Linguistics.

de Lhoneux, M., Stymne, S., and Nivre, J. (2017b). Arc-hybrid non-projective dependency parsing with a static-dynamic oracle. In *Proceedings of the 15th International Conference on Parsing Technologies*, pages 99–104, Pisa, Italy. Association for Computational Linguistics.

References IV

Dehouck, M. and Denis, P. (2019). Phylogenic multi-lingual dependency parsing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 192–203, Minneapolis, Minnesota. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

References V

Dryer, M. S. and Haspelmath, M. (2013). *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Hwa, R., Resnik, P., Weinberg, A., Cabezas, C., and Kolak, O. (2005). Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering*, 11(3):311–325.

Kiperwasser, E. and Goldberg, Y. (2016). Simple and accurate dependency parsing using bidirectional LSTM feature representations. *Transactions of the Association for Computational Linguistics*, 4:313–327.

Kondratyuk, D. and Straka, M. (2019). 75 languages, 1 model: Parsing universal dependencies universally. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2779–2795, Hong Kong, China. Association for Computational Linguistics.

References VI

Lauscher, A., Ravishankar, V., Vulić, I., and Glavaš, G. (2020). From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Lewis, M. P., editor (2009). *Ethnologue: Languages of the World*. SIL International, Dallas, Texas, USA, sixteenth edition.

Lin, Y.-H., Chen, C.-Y., Lee, J., Li, Z., Zhang, Y., Xia, M., Rijhwani, S., He, J., Zhang, Z., Ma, X., Anastasopoulos, A., Littell, P., and Neubig, G. (2019). Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

References VII

Littell, P., Mortensen, D. R., Lin, K., Kairis, K., Turner, C., and Levin, L. (2017). URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain. Association for Computational Linguistics.

Meechan-Maddon, A. and Nivre, J. (2019). How to parse low-resource languages: Cross-lingual parsing, target language annotation, or both? In *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, pages 112–120, Paris, France. Association for Computational Linguistics.

Moran, S. and McCloy, D. (2014). PHOIBLE Online. Max Planck Institute for Evolutionary Anthropology, Leipzig.

References VIII

Müller-Eberstein, M., van der Goot, R., and Plank, B. (2021). Genre as weak supervision for cross-lingual dependency parsing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4786–4802, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nguyen, M. V., Lai, V., Veyseh, A. P. B., and Nguyen, T. H. (2021). Trankit: A light-weight transformer-based toolkit for multilingual natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*.

Nordhoff, S. and Hammarström, H. (2011). Glottolog/langdoc: Defining dialects, languages, and language families as collections of resources. In *Proceedings of the First International Workshop on Linked Science 2011*, volume 783 of *CEUR Workshop Proceedings*.

References IX

Smith, A., Bohnet, B., de Lhoneux, M., Nivre, J., Shao, Y., and Stymne, S. (2018). 82 treebanks, 34 models: Universal dependency parsing with multi-treebank models. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 113–123, Brussels, Belgium. Association for Computational Linguistics.

Stymne, S. (2020). Cross-lingual domain adaptation for dependency parsing. In *Proceedings of the 19th International Workshop on Treebanks and Linguistic Theories*, pages 62–69, Düsseldorf, Germany. Association for Computational Linguistics.

Stymne, S., de Lhoneux, M., Smith, A., and Nivre, J. (2018). Parser training with heterogeneous treebanks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 619–625, Melbourne, Australia. Association for Computational Linguistics.

References X

Täckström, O., McDonald, R., and Uszkoreit, J. (2012). Cross-lingual word clusters for direct transfer of linguistic structure. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 477–487, Montréal, Canada. Association for Computational Linguistics.

Tiedemann, J., Agić, Ž., and Nivre, J. (2014). Treebank translation for cross-lingual parser induction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 130–140, Ann Arbor, Michigan. Association for Computational Linguistics.

Turc, I., Lee, K., Eisenstein, J., Chang, M.-W., and Toutanova, K. (2021). Revisiting the primacy of english in zero-shot cross-lingual transfer. *ArXiv*, abs/2106.16171.

References XI

Üstün, A., Bisazza, A., Bouma, G., and van Noord, G. (2020). UDapter: Language adaptation for truly Universal Dependency parsing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2302–2315, Online. Association for Computational Linguistics.

van der Goot, R. and de Lhoneux, M. (2021). Parsing with pretrained language models, multiple datasets, and dataset embeddings. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, pages 96–104, Sofia, Bulgaria. Association for Computational Linguistics.

van der Goot, R., Üstün, A., Ramponi, A., Sharaf, I., and Plank, B. (2021). Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197, Online. Association for Computational Linguistics.

Vania, C., Kementchedjhieva, Y., Søgaard, A., and Lopez, A. (2019). A systematic comparison of methods for low-resource dependency parsing on genuinely low-resource languages. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1105–1116, Hong Kong, China. Association for Computational Linguistics.

References XIII

Wu, S. and Dredze, M. (2019). Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Zeman, D. and Resnik, P. (2008). Cross-language parser adaptation between related languages. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*.

Zhang, Y. (2021). The influence of M-BERT and sizes on the choice of transfer languages in parsing. Master thesis, Uppsala University, Sweden.