

December 16, 2022

International Conference on
Natural Language and Speech
Processing



Sous la co-tutelle de :

CNRS
ÉCOLE DES PONTS PARISTECH
UNIVERSITÉ GUSTAVE EIFFEL

Hybrid Language Processing in the Deep Learning Era

Eric Laporte, LIGM



Université
Gustave Eiffel

What can hybrid natural language processing do for you?

Pure neural nets: a bet on the future
What can symbolic resources be used for?
What are good symbolic resources for NLP?





Pure neural nets: a bet on the future

Deep learning only

'Deep learning is going to be able to do everything' (Hinton, 2020, interview by Hao, 2020)

'Yann LeCun (...) is betting on (...) machine learning models that can be trained without the need for human-labelled examples' (Dickson, 2022)

Deep learning is very successful in translation and other applications

Enthusiasm and an urge to explore its limits

Fans bet on a little more

They believe it will allow for doing NLP without any linguistic knowledge

Everything would be learnt from examples by neural models

Even what humans already know or know how to find out

'Symbolic models don't work' (1/4)

'One of the criticisms raised against lexicon-based methods is that the dictionaries are unreliable, as they are either built automatically or hand-ranked by humans' (Taboada *et al.*, 2011)

'Manually creating such rules (...) is [an] error-prone task' (Kim *et al.*, 2004)

In academia, the distrust of symbolic approaches is so widespread that some researchers don't even give any reasons for it

'The goal of our approach is to reproduce the segmentation (...) using as little resources as possible' (Gahbiche-Braham, 2013, p. 54)

'Symbolic models don't work' (2/4)

Objective evidence?

Precision and recall?

It depends on the tasks

Adaption time describes the time difference between the desired departure time of a passenger and his actual departure time

Disambiguating phrases

Neural models are better

From Edensor [village latitude=53.227°N longitude=1.625°W], you continue to Chatsworth House [building latitude=53°13'40"N longitude=1°36'36"W], up to the Hunting Tower [building], then south through the stunning Stand Wood [forest].

Fine-grained information about unambiguous place names

A simple dictionary is better

A neural model requires huge annotated data to learn with the same coverage

Not so simple

'Symbolic models don't work' (3/4)

Objective evidence? Precision and recall?

Performance depends on

- content of added information disambiguation / spotting phrases
- granularity fine-grained / basic (derogatory/laudatory)
- steps in processing chain segmentation / detection / disambiguation
- languages

Comparing systematically symbolic vs. neural models would involve varying all these parameters

Actual comparisons

In competitions and shared tasks (few researchers are experts in both neural nets and symbolic approaches)

Average out several types of annotation (Ye *et al.*, 2021)

Inconclusive

'Symbolic models don't work' (4/4)

Objective evidence?
Scalability and runtime efficiency?

'Throughput and memory footprint (...), while extremely important for commercial vendors, are typically not reported in NLP literature' (Chiticariu *et al.*, 2013, p.830)

Deep learning is computationally expensive for the moment

Not an objective flaw of symbolic models

Distrust of symbolic approaches is speculative

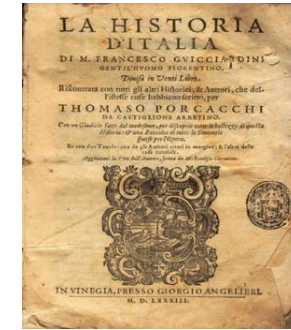
Neural and symbolic models: two approaches to NLP

A parallel with archaeologists and historians Two approaches to finding out about past events

Archaeologists and historians cooperate
They exchange information in mutual trust
They take advantage of any available results



By U.S. Air Force photo/Tech. Sgt. Shane A. Cuomo - <http://www.af.mil/shared/media/photodb/photos/080112-F-2034C-215.jpg>, Public Domain, <https://commons.wikimedia.org/w/index.php?curid=3514446>



Public Domain, <https://commons.wikimedia.org/w/index.php?curid=664067>

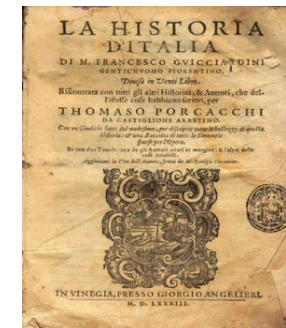
Indirect evidence of distrust of symbolic resources

*'The **external knowledge** ranges from linguistic, semantic, commonsense, factual, to domain-specific knowledge'* (Qiu et al., 2020)

Symbolic resources are often said to be 'external' to a hybrid system

Not really external: they are essential to its operation

Archaeologists don't say historical data are 'external', they say they are 'historical'



'Working on symbolic approaches is boring' (1/3)

*'The availability of [meaningful and annotated] data is (...) contingent on the **tedious and time-consuming** annotation job'* (Ahmed et al., 2020)

*'Building NLP datasets is often a **time-consuming and tedious** task'* (the PIAF project, 2020)

Chiticariu et al. (2013) interpret the frequent statement as *'What's the research in rule-based IE? Just go ahead and write the rules'*

Tasks that resist automation

- building resources
- revising pre-annotated data

Geeks mostly enjoy automating tasks

They find such problems uninteresting

They even find the solutions uninteresting



By Krassotkin - Own work, CC0,
<https://commons.wikimedia.org/w/index.php?curid=30819193>

'Working on symbolic approaches is boring' (2/3)

We are not all 100% geek

Disliking symbolic resources is information about dislikers

Other researchers enjoy the challenges of symbolic models and resources:

- coverage
- models
- updatability

The true message of '*tedious and time-consuming*': '***it is not an exciting task***'

Knowing what authors enjoy is interesting for their friends

In a scientific paper, it's exactly as relevant as knowing that they like broccoli or dance-pop

More relevant would be an argued **position** about

- the best approaches to symbolic models and resources
- the results of these approaches



By Fir0002 - Own work, GFDL 1.2, <https://commons.wikimedia.org/w/index.php?curid=5772317>

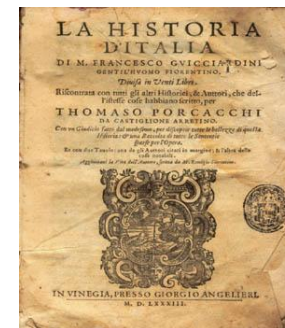


Alphaville - Big in Japan

'Working on symbolic approaches is boring' (3/3)

The *'tedious and time-consuming'* stereotype is not a scientific result

Historians don't say excavating sites is *'tedious and time-consuming'*
They are curious of the results
They just use them



Obfuscation of symbolic components

In academia

'Authors hid rules behind euphemisms such as "dependency restrictions", "entity type constraints", or "seed dictionaries" ' (Chiticariu et al., 2013, p. 828)

In business

Google does not communicate much about the query patterns in Google Search
Amazon communicates about the machine learning in Alexa, not much about the symbolic component

Discrepancy between choices and communication

This may give a sense these approaches are not being used, but they are

Distrust of symbolic resources is opinion, not science

What can these resources do for you?



What can symbolic resources be used for?

Symbolic models are used in business

'Knowledge-based techniques (a.k.a. symbolic) (...) have terrific value in the enterprise world. (...) The symbolic system (...) is significantly more efficient' (Scagliarini, 2022)

Symbolic models *'are much better positioned to reason their way through complex scenarios, (...) are better able to precisely represent relationships between parts and wholes, (...) are more robust and flexible in their capacity to represent and query large-scale databases'* (Marcus, 2022)

Such awareness in business is based on objective measurement of performance

Some reporter job

Amazon's virtual assistant Alexa has a strong, efficient symbolic model which has precedence over the neural model

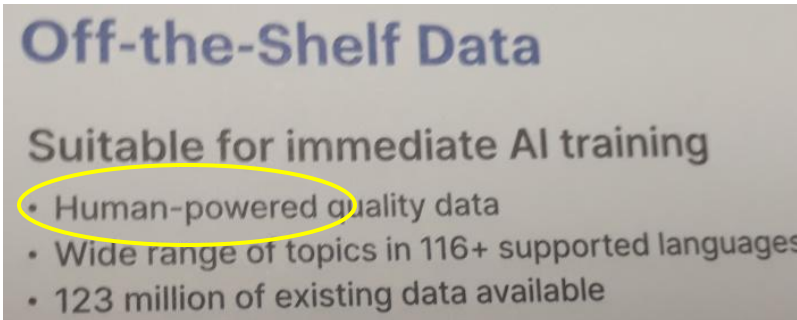
'Google Search (...) uses a pragmatic mixture of symbol-manipulating AI and deep learning' (Marcus, 2022)

Google patent no. WO 2017/024108 A1 uses language-specific *'predefined rules for matching [and interpreting] a (...) search query'*, called query patterns, e.g. *weather in Nplace*, i.e. a symbolic resource (Slawski, 2019)

Symbolic models are in demand in business

NLP businesses are seeking

- symbolic resources
- professionals able to build them
- methods to build them



Off-the-Shelf Data

Suitable for immediate AI training

- Human-powered quality data
- Wide range of topics in 116+ supported languages
- 123 million of existing data available

An ad at COLING 2022

Responsibility of academia towards business

- design methods
- provide and apply criteria of quality
- train professionals
- in some cases, build resources

Disliking symbolic resources is inconsistent with our responsibility

What are symbolic resources used for? What can they be used for?

Supervised deep learning depends on annotated data

'Among the known limits of deep learning is need for massive training data and lack of robustness in dealing with novel situations' (Dickson, 2022)

'Prioritizing the cultivation of new datasets and research communities around them could be essential to extending the present AI summer' (Wissner-Gross, 2016)

Labelled data

Deep learning requires even bigger data

New data is required to scale up to new tasks or requirements of quality: *'deep-learning systems (...) frequently stumble when confronted with novelty'*
(Marcus, 2022)

Weak supervision

Semi-supervision, indirect supervision, self-supervision

Human-labelled data is usually required at some stage

Annotating training data

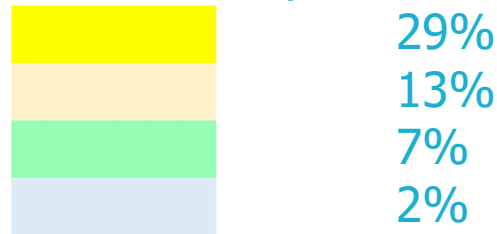
Same problem as the initial task to be automated, on a finite dataset

Annotation depends on automated pre-annotation (1/2)

Annotating data is repetitive

Example: labelling queries to a virtual assistant for tasks on demand

Recurrent situations are annotated each time they occur



Zipf's law

Human annotators get bored and make errors

Misuse of human labour

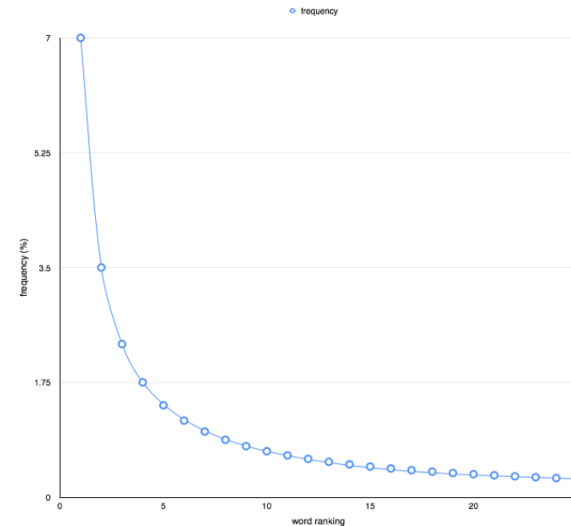
Crowdworkers and undergraduates

Assigning annotation to inexperienced workforce is worse

The problem is more when annotation is difficult

Automated pre-annotation + human revision

assistance cancelling an order I have made
assistance canceling the order I have made
I want to cancel the order I made
I do not know how I can cancel the order I made
help me to cancel an order I have made
I don't know how I could cancel the order I made
is it possible to cancel the order I have made?
would you give me information about canceling orders?
I have a problem with canceling the order I have made
I need help to cancel the order I made
I need help to cancel the last order I have made
I can't pay for the last order I have made
I would like to cancel my last order
where can I cancel my order?
help me canceling my order
question about canceling my order
question about canceling an order
problem with canceling the order I have made
could you help me cancelling an order I made?
question about canceling the last order I made
I want to know more about order cancellations
could you help me to cancel an order?
could you help me cancel the last order I have made?
where do I cancel the last order?
I try to cancel the order I made
help me cancelling an order I made
help me canceling the order I made
where can I cancel the order I have made?



hoakley <https://eclecticlight.co/2015/07/11/zipfs-law-deep-and-meaningful/>

Annotation depends on automated pre-annotation (2/2)

Annotating technical, legal or financial corpora

'The annotation process (...) requires the support of subject matter experts [which are] very expensive and scarce resources' (Scagliarini, 2022)

Automated pre-annotation + human revision

'Text annotation is dominating the global labeling market in 2022 to fine-tune AI's capacity to recognize patterns in text, voice, and semantic connection of annotated data. Plus, the development of text mining applications depends largely on pre-annotated text' (Kniazieva, 2022)

Pre-annotation depends on symbolic resources (1/2)

Train neural nets to pre-annotate the data required to train neural nets?

Ingenuous

Human generalisation and formalisation of observations on raw data

The '*Neural:Symbolic* → *Neural*' model of Kautz (2020)

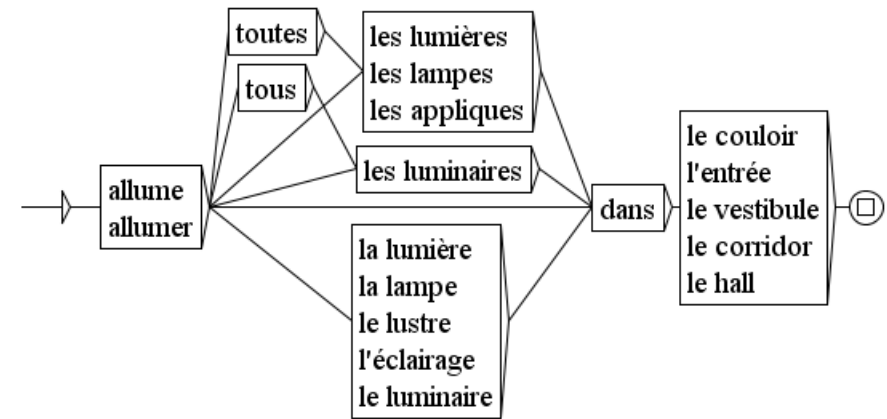
Rules and dictionaries

'Dictionary-based pre-annotation is a feasible and practical method to reduce the cost of annotation (...) without introducing bias in the annotation process' (Lingren *et al.*, 2014)

Data augmentation

Some techniques also involve symbolic resources (Wei, Zou, 2019; Feng *et al.*, 2020; Coulombe, 2020)

Supervised deep learning depends on symbolic resources



Rule for *turn on the light in the hallway* in French for a task-on-demand virtual assistant

Pre-annotation depends on symbolic resources (2/2)

Pre-annotating technical, legal or financial corpora

Generalisation by knowledge engineers

'Tools [for symbolic language processing] enable a specific type of expert, the so-called knowledge engineers, to write simple rules to develop the model'

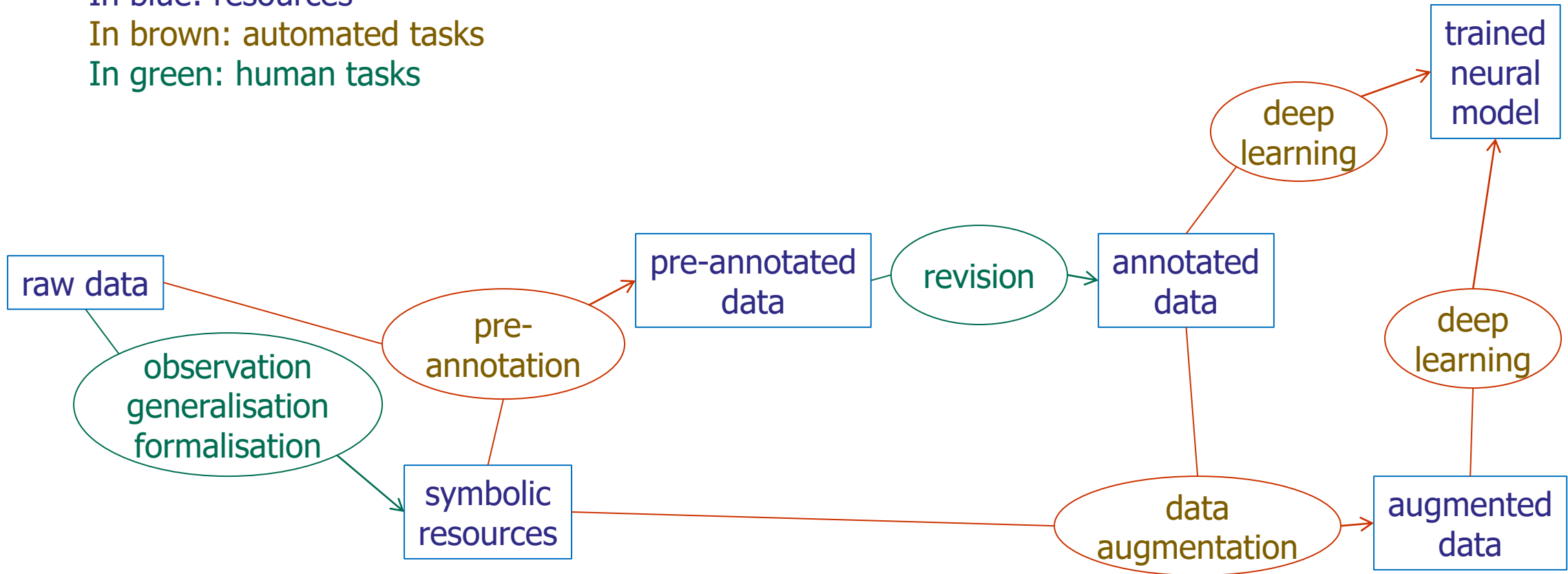
(Scagliarini, 2022)

Supervised deep learning depends on symbolic resources

In blue: resources

In brown: automated tasks

In green: human tasks



Bonus: human control on symbolic resources

Direct control on:

- knowledge encoded in symbolic resources
- definition and implementation of annotation guidelines

Indirect control on performance of hybrid systems

Impact on training data and trained models

Quality improvement

Trace errors to a readable symbolic resource like a dictionary, rule, or graph

Fix the resource

Rerun the pre-annotation

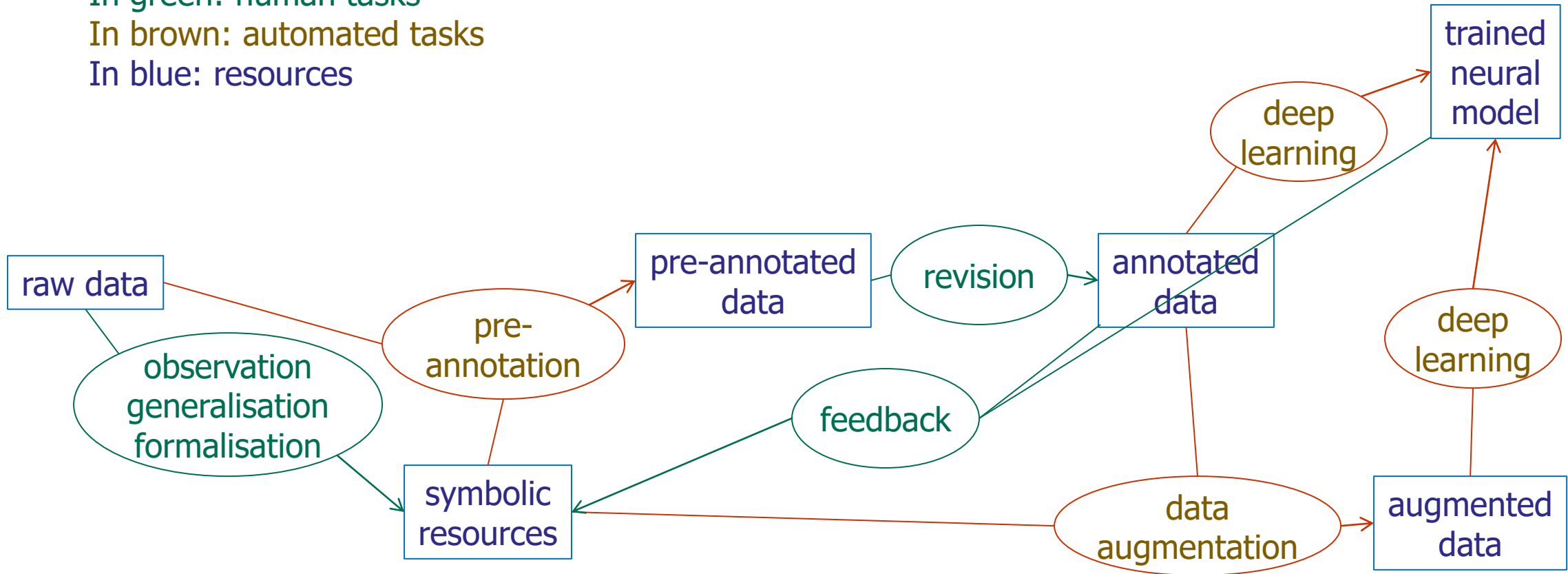
Revise the data

Retrain the model

Symbolic resources provide a means of controlling the quality of systems

'We (...) don't know what to do about [deep learning systems] (except to gather more data) if they come up with the wrong answers' (Marcus, 2022)

In green: human tasks
In brown: automated tasks
In blue: resources



Tokenization

Converting text into a string of tokens

All NLP, even tasks where training data is raw text:

- translation
- word prediction
- next sentence prediction

A small component with a strong impact on output

English, French...

Tokens can be space-delimited words

Languages with restricted graphical delimitation

- Arabic
- Agglutinative languages
- Chinese
- Vietnamese...

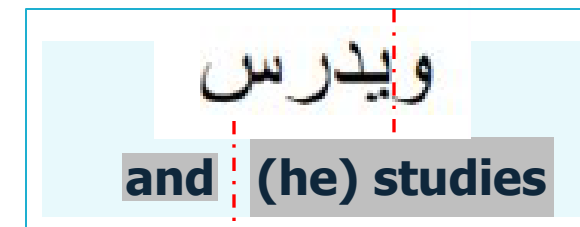
Many meaningful units are subwords

The lexicon of words is bigger than the lexicon of meaningful units

Tokenizing into words produces **data scarcity** (artefact)



5 words - 5 tokens



1 word - 2 or 3 tokens

Relaxing synchronicity constraints in tokenization

Tokens can be subwords

Tokens can contain spaces

- Vietnamese
- multiword expressions

phụ nữ 'woman'
en. *on time*

Tokens can be non-concatenative

Semitic languages: pattern and root



Neural tokenization

Token limits learnt from combinatorial statistics in raw data (Kudo, Richardson, 2018; Nguyen-Vo *et al.*, 2021)

Relaxing synchronicity constraints in tokenization

Tokens can be subwords

Tokens can contain spaces

- Vietnamese *phụ nữ* "woman"
- multiword expressions en. *on time*

Tokens can be non-concatenative

Semitic languages: pattern and root



Hybrid tokenization

Dictionaries and rules define possible delimitations

Supervised model chooses best path among lattice of possible delimitations (Sak *et al.*, 2011)

Neural models are as applicable to lattices as to sequences (Sperber *et al.* 2017)

With this approach, the resources can take into account semantic contribution of tokens

Can manage non-concatenative tokens (Neme, Paumier, 2020)

Approach not mentioned in Alyafeai *et al.* (2022)

Bonus: human control on definition of elementary units

Different forms, same item
fr. peuvent / pouvaient

Same form, different items
en. hopeful / awful

Balancing deep learning with symbolic resources

The proportion between deep learning and symbolic resources need not be 100% vs. 0%

Leverage the best from computers' strengths and the best from humans' strengths

Assign what computers can do to the neural component, and what they cannot do to a human-made symbolic component

What can and cannot computers do?

The **human-in-the-loop** trend

Tries to find the right balance as well but with direct human interaction (Wang, 2021)

Human labour is better used through symbolic resources because they can be reused

Computers' strengths

'Solv[ing] tasks by statistical approximation and learning from examples' (Marcus, 2022)

Weaknesses

Translation of a sentence into a near-equivalent of its negation

fr. *Cela n'empêche que vous êtes en retard* 'Still, you're late'

translated into *That doesn't mean you're late* (Google translate, October 14, 2022)

Challenges

Tasks that involve reasoning

Finding *'who the prime minister of the UK was when Theresa May was born'*
(Lenat, 2019)

Humans' strengths

'Provid[ing] vast amounts of background knowledge' (Workshop on Semiparametric Methods in NLP: Decoupling Logic from Knowledge)
'Curating and carefully documenting datasets' (Bender *et al.*, 2021)
Reference books with checked information, e.g. conjugation textbooks

Building symbolic resources with explicit formal knowledge for

- pre-annotation
- tokenization

Revising pre-annotated data

Workshops on hybrid NLP

Workshop on Semiparametric Methods in NLP: Decoupling Logic from Knowledge (Spa-NLP)

Workshop on Pattern-based Approaches to NLP in the Age of Deep Learning (Pan-DL)

Workshop on Knowledge Augmented Methods for NLP

Workshop on Ten Years of BabelNet and Multilingual Neurosymbolic Natural Language Understanding

The background features a dark blue upper section and a teal lower section. Large, white, curved geometric shapes are positioned in the top right, while lighter teal curved shapes are in the bottom right. The text is centered in the dark blue area.

What are good symbolic resources?

Are there good symbolic resources?

'Manipulating symbols (...) is treated as a dirty word in deep learning' (Marcus, 2022)

Dislike of symbolic resources: a reason not to develop them

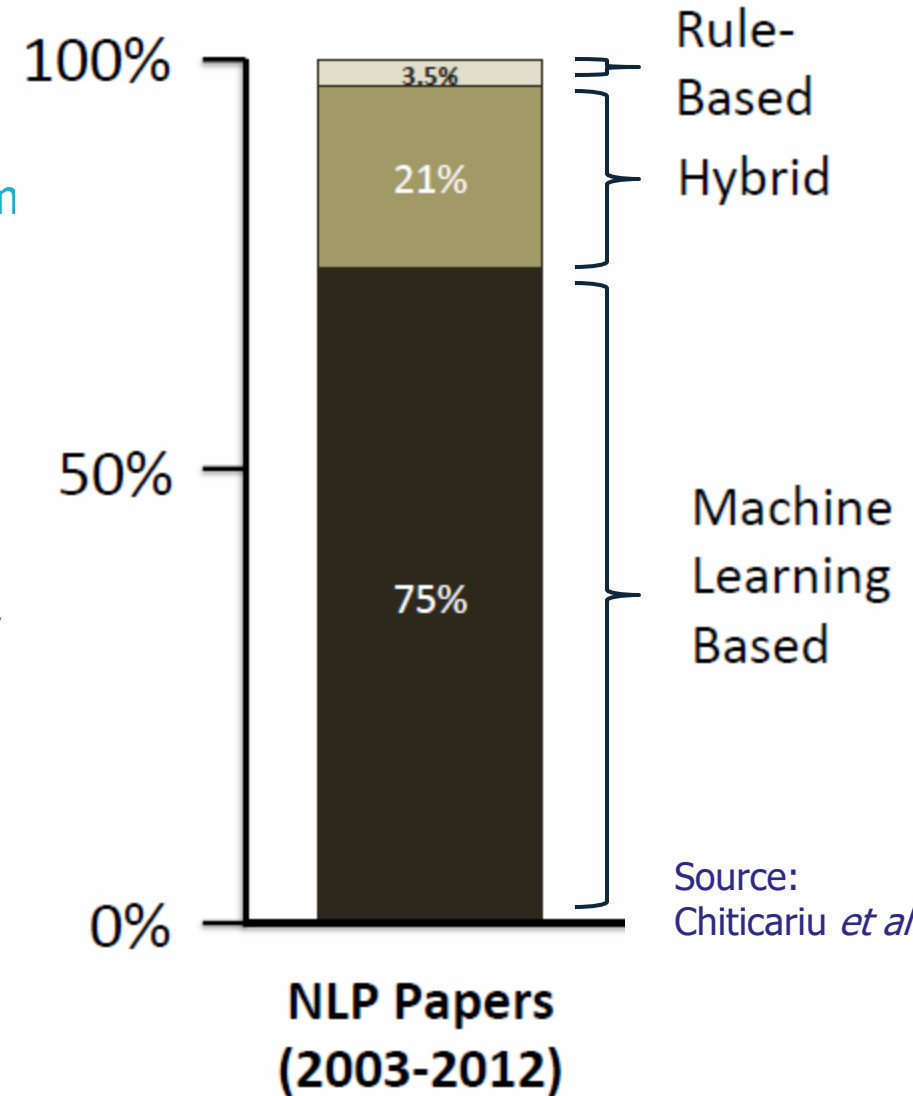
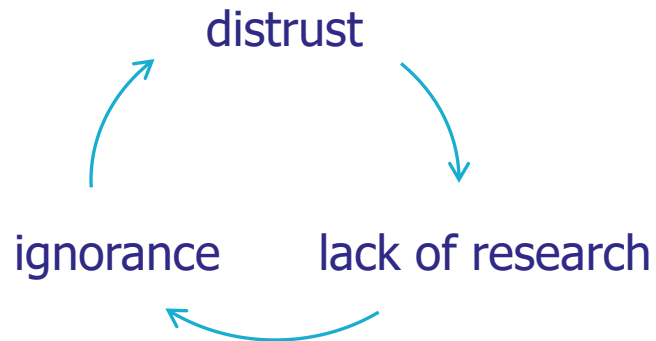
Distrust: a reason not to use them

Symbolic resources remain underinvestigated

A vicious circle

Ignorance in linguistics

'A dictionary, also called lexicon and gazetteer, is a set of words sharing the same semantics' (Wang et al., 2021)



Source:
Chiticariu et al., 2013

Distrust of symbolic resources is a prejudice and a misconception

Many experts consider **obvious** you can't trust symbolic models
This view is speculative

Scientists must beware of obviousness and collective perception

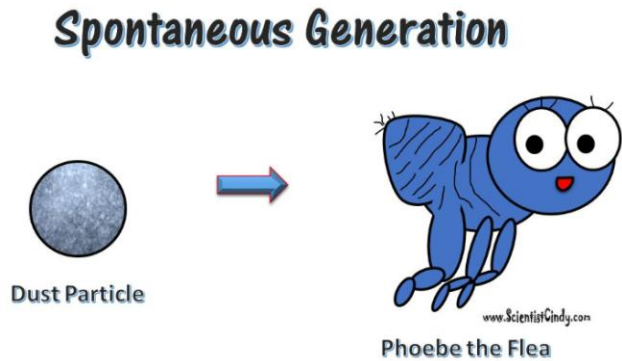
For centuries, it was obvious that fleas might be generated by dust

It is a duty for us scientists to doubt

Take into account mainstream view, but **think critically**

My work and experience taught me that symbolic resources can do things other
methods still cannot

But how to explain such distrust?



<http://www.scientistcindy.com/>

Prestige or quality?

Linguistics for NLP is not prestigious

'Belittling unfashionable ideas that haven't yet been fully explored is not the right way to go' (Marcus, 2022)

Descriptive linguistics has not been fully explored

Potential for scientific innovation is more important than reputation

We scientists should not be fooled by levels of prestige

We are able to check whether the prestige is deserved

Take into account current trends, but don't be biased

The true reasons why symbolic approaches are distrusted

Dramatic progress achieved by deep learning Weaknesses of research on symbolic resources

'There is an enormous untapped opportunity for researchers to make the rule-based approach more principled, effective, and efficient' (Chiticariu et al., 2013, p. 830)

Challenges to linguistics for NLP

Maintainability

Observation

Rules and coverage

Modelling and discretisation

Challenge 1: Maintainability (1/2)

Symbolic resources should be easy to edit and maintain

Possible human errors

Language evolution

Extension to new tasks

Like software source code

Compactness, simplicity and clarity

Compactness and simplicity

Dictionaries of inflected forms are full of duplicate information

púlpito,N001
pulsação,N102
pulsante,A301
pulsão,N102
pulsar,N004
pulsar,V005

Dictionary of lemmas

púlpito,púlpito.N:ms
púlpitos,púlpito.N:mp
pulquérroma,pulcro.A:Sfs
pulquérrimas,pulcro.A:Sfp
pulquérrimo,pulcro.A:Sms
pulquérrimos,pulcro.A:Smp
pulsa,pulsar.V:P3s
pulsa,pulsar.V:Y2s
pulsação,pulsação.N:fs
pulsações,pulsação.N:fp
pulsada,pulsar.V:K
pulsadas,pulsar.V:K
pulsado,pulsar.V:K
pulsados,pulsar.V:K
pulsai,pulsar.V:Y2p
pulsais,pulsar.V:P2p
pulsam,pulsar.V:P3p
pulsamos,pulsar.V:J1p

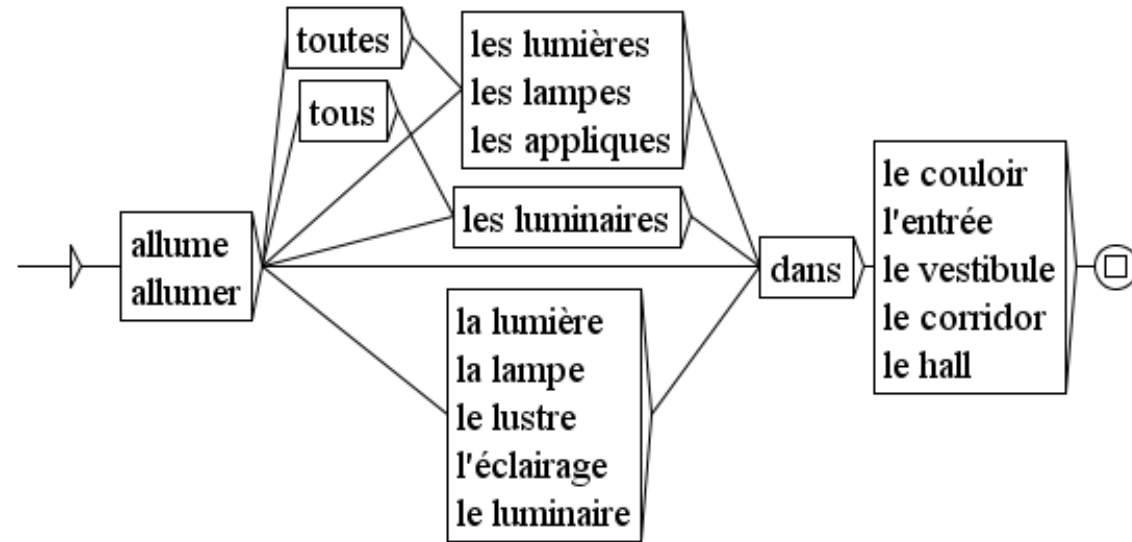
Dictionary of inflected forms

Maintainability (2/2)

(allume + allumer) ((toutes + <E>) (les lumières + les lampes + les appliques) + (tous + <E>) (les luminaires) + <E> + (la lumière + la lampe + le lustre + l'éclairage + le luminaire)) dans (le couloir + l'entrée + le vestibule + le corridor + le hall)

Clarity

Finite automata vs. regular expressions



Not all linguists address the challenge of maintainability

But some do, you can check in their papers and resources

Challenge 2: Observation

Cucurbitaceae plural noun

Cu-cur-bi-ta-ce:ae (.)kyū,kərbə'tāsē,ē

: a family of chiefly herbaceous tendril-bearing vines (order

Erroneous observations

'One of the criticisms raised against lexicon-based methods is that the dictionaries are unreliable, as they are either built automatically or hand-ranked by humans' (Taboada *et al.*, 2011)

C. sativus known as cucumber is a Cucurbitaceae from India

?The lake missed its freeze date by 2 days

No Google hit for *"missed its freeze date"*

Gunflint Lake missed its average ice-out date of May 7

Possible observation methods

- Authentic attestations

But absence of evidence is not evidence

May be unrealistic for statistical reasons

- Introspection

Requires training, caution and peer reviewing (judge and party)

Not all linguists are cautious about that

But you can check in their papers

Challenge 3: Rules and coverage (1/2)

All languages contain chaos

Broken plural in Arabic

160 combinations of a singular pattern, a plural pattern and a root alternation (Neme, Laporte, 2013)

Singular pattern	Plural pattern	Root alternation
<i>1a2o3</i>	<i>Oa1o2aAo3</i>	<i>1y3</i>
<i>naAob</i> 'tooth'	<i>OanoyaAob</i> 'teeth'	

Partial coverage does not help

Describing broken plural only for trilateral roots

Setting semantic classes of nouns from singular patterns in Arabic

Does not work for all nouns, e.g. *qalam* 'pen'

Practically any rule has exceptions

Impossible to find a correct rule without checking the whole lexicon

Finding rules involves a big descriptive task

Class	Singular pattern	Example	In English
Agent	<i>mu1a22i3</i>	<i>muDaxGim</i>	'amplifier'
Instrument	<i>mi1o2a3</i>	<i>miEowal</i>	'mattock'
Place	<i>ma1o2a3ap</i>	<i>makotabap</i>	'library'

Rules and coverage (2/2)

Observational challenges

coverage

accuracy

interaction between levels of analysis (fr. *sa femme / son épouse*)

sense-related distinctions (en. *on time / on that day*)

Modelling-related challenges

comparing models

rule chaining or not (fr. *peuvent/pouvaient*)

rule hierarchy or not

updatability

Human driving force: curiosity

Not all linguists accept the descriptive part of the task

You can check in their papers and resources

Making the rules: a noble task?

Scanning the lexicon: a menial task?

But then who should do the descriptive job?

Challenge 4: Modelling and discretisation (1/3)

Inventory of lexical entries

Example: spelling correction with diagnosis

For a polysemic word:

- *I miss my years in Brazil*
- *I missed the deadline by 2 days*
- *The missile missed its target by a hair*

A discrete number of senses but how many?

Inventory of linguistic features

Lexical features that each entry can have or not:

- possible contexts
- form variations
- syntactic operations applicable

Discretisation

Required for a lexical database

Different senses of a word have different features

If they are represented in a single entry, you don't know which features it has

	$N_0=:N\text{-hum}$	<i>for Ntime</i>	passive
N_0 miss N_1 sense 1	-	+	+
N_0 miss N_1 sense 2	+	-	+
N_0 miss N_1 sense 3	+	-	+

	$N_0=:N\text{-hum}$	<i>for Ntime</i>	passive
N_0 miss N_1	+	+	+

Modelling and discretisation (2/3)

Approximation of linguistic reality

Decisions may be partly arbitrary

- two senses or one? → *I missed the deadline by 2 days*
I missed the third episode
- encode two senses if they differ by at least a feature represented in the model
- what about when a sense is in the process of splitting in two?

- dubious usage → *?I missed the third episode by a day*

The price to be paid for an operational model

Not all linguists accept to address this challenge

Reluctant to make arbitrary decisions

Especially corpus linguists

Focus on careful and rigorous observation

Find discretisation a risky territory

You can check in their papers and resources

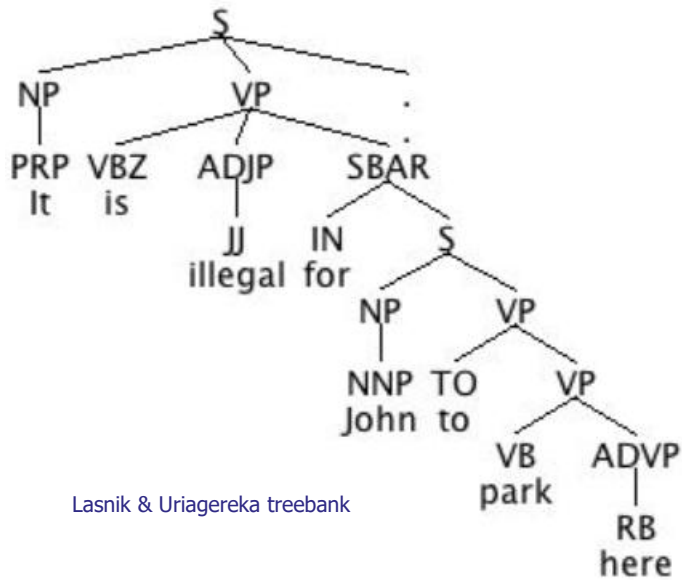
Modelling and discretisation (3/3)

Consequences on treebanks

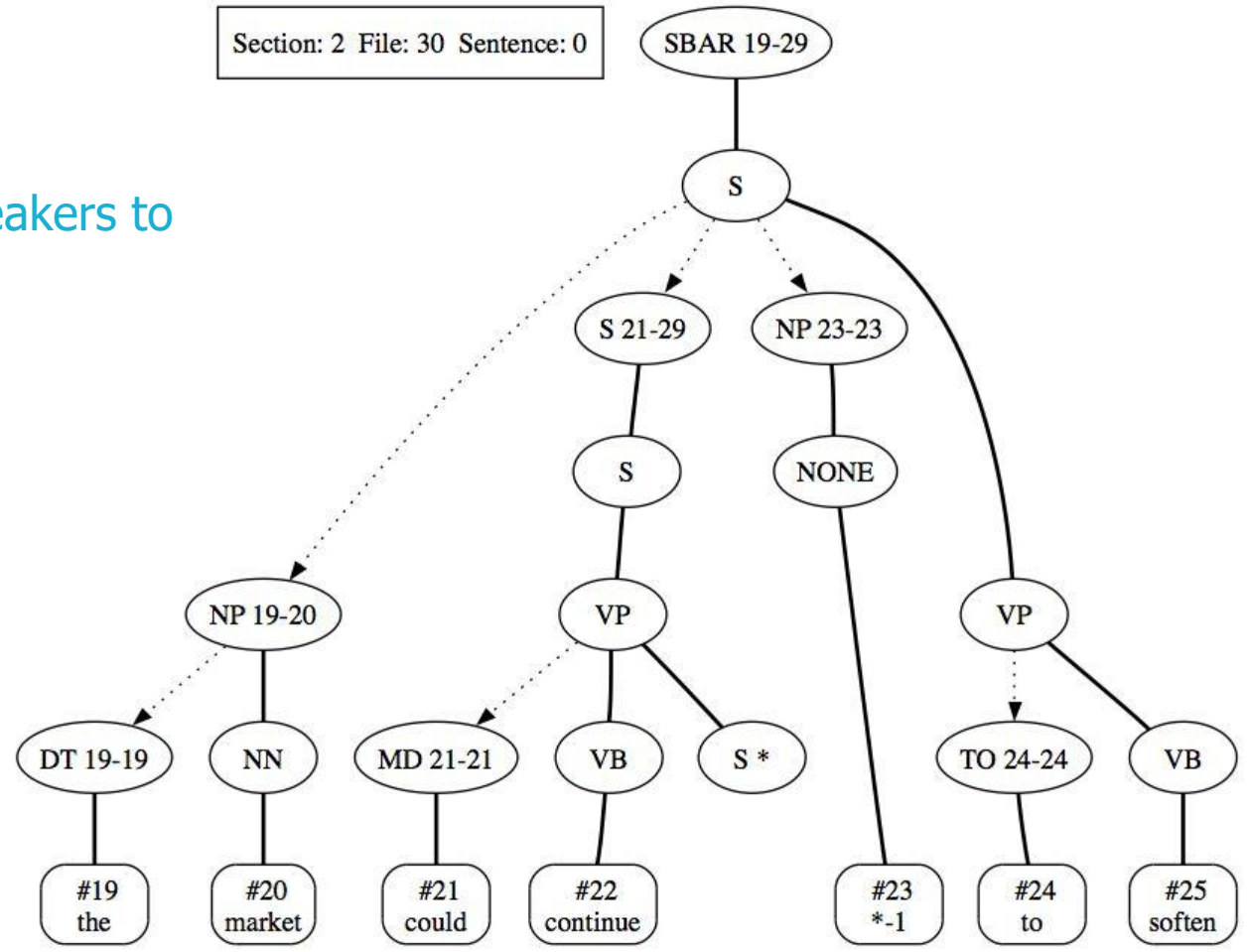
They never include identifiers of

- lexical entries
- the syntactic operations applied by speakers to produce sentences

They should



Lasnik & Uriagereka treebank



Shen, Champollion, Joshi, Mannem

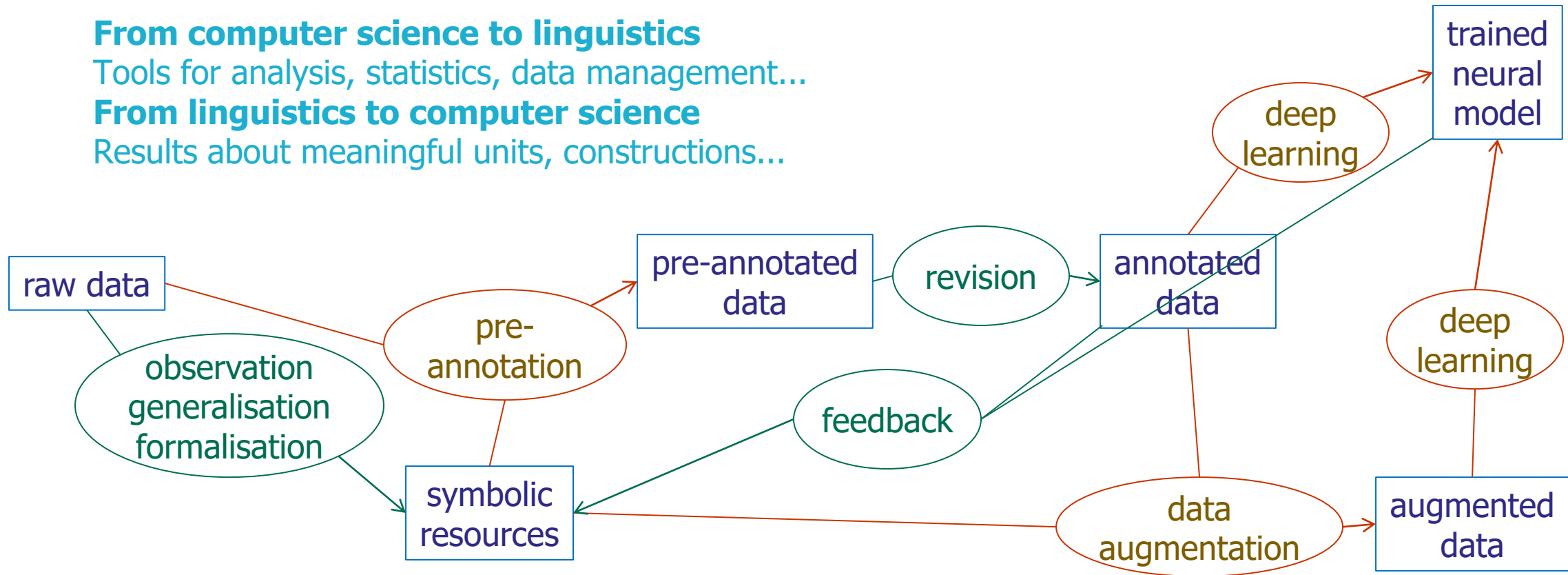
A model of cooperation between computer science and linguistics

From computer science to linguistics

Tools for analysis, statistics, data management...

From linguistics to computer science

Results about meaningful units, constructions...



The processing chain for annotating data requires co-operation between linguists and computer scientists

Disliking symbolic resources is a weakness

How to obtain symbolic data

Don't do it yourself if you're not a linguist



Don't trust just any provider of symbolic language resources
Check the maintainability, the observation, the coverage, the modelling

Conclusion

Pure neural nets: a bet on the future
What can symbolic resources be used for?
What are good symbolic resources for NLP?

What if **pure** neural nets turn out not to be the best approach to **all** NLP tasks?

Symbolic resources **help** in

- preparing data at less expense
- guiding the evolution of systems
- handling the right elementary units

To select **satisfactory symbolic resources** for NLP, check

- maintainability
- the carefulness of observational processes
- coverage
- if they follow a discrete model

Synergy between machine learning and symbolic resources

References (1/3)

- Ahmed L, Ahmad K, Said N, Qolomany B, Qadir J, Al-Fuqaha A, 2020. Active Learning Based Federated Learning for Waste and Natural Disaster Image Classification, *IEEE Access*, vol. 8, pp. 208518-208531, doi: 10.1109/ACCESS.2020.3038676.
- Almeida H, Meurs MJ, Kosseim L, Tsang A, 2016. Data Sampling and Supervised Learning for HIV Literature Screening. *IEEE Trans Nanobioscience* 15(4):354-361, doi: 10.1109/TNB.2016.2565481.
- Alyafeai Z, Al-shaibani MS, Ghaleb M, Ahmad I, 2022. Evaluating Various Tokenizers for Arabic Text Classification, *Neural Processing Letters*, doi: 10.1007/s11063-022-10990-8.
- Bender E, Gebru T, McMillan-Major A, 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *FACCT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.
- Chiticariu L, Li Y, Reiss FR, 2013. *Rule-Based Information Extraction is Dead! Long Live Rule-Based Information Extraction Systems!* EMNLP, pp. 827--832. Association for Computational Linguistics.
- Coulombe C, 2020. Techniques d'amplification des données textuelles pour l'apprentissage profond. Ph.D. thesis, Télé-université.
- Dickson B, 2022. *Meta's Yann LeCun on his vision for human-level AI*. Blog post, <https://bdtechtalks.com/2022/03/07/yann-lecun-ai-self-supervised-learning/>
- Feng SY, Gangal V, Kang DY, Mitamura T, Hovy E, 2020. Genaug: Data augmentation for finetuning text generators. arXiv preprint arXiv:2010.01794.
- Gahbiche-Braham S, 2013. *Amélioration des systèmes de traduction par analyse linguistique et thématique. Application à la traduction depuis l'arabe*. PhD, Université Paris Sud.
- Hao K, 2020. "AI pioneer Geoff Hinton: 'Deep learning is going to be able to do everything.'" *MIT Technology Review*.

References (2/3)

- Kautz H, 2020. The Third AI Summer, *AAAI*, https://www.youtube.com/watch?v=_cQITY0SPiw.
- Kim S, Lewis P, Martinez K, 2004. The Impact of Enriched Linguistic Annotation on the Performance of Extracting Relation Triples. In: Gelbukh, A. (eds) *Computational Linguistics and Intelligent Text Processing. CICLing 2004*. LNCS 2945. Springer, https://doi.org/10.1007/978-3-540-24630-5_68
- Kniazieva Y, 2022. *What Does the Future Hold for the Data Annotation Industry?* Wilmington, Delaware: LabelYourData, https://labeledyourdata.com/articles/trends-in-data-annotation-market-forecast-2022#_data_annotation_tools
- Kudo T, Richardson J, 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing, *EMNLP (System Demonstrations)*, pages 66–71.
- Lenat D, 2019. What AI Can Learn From Romeo & Juliet. Blog post, *Cognitive World*, Forbes.
- Lingren T, Deleger L, Molnar K, Zhai H, Meinzen-Derr J, Kaiser M, Stoutenborough L, Li Q, Solti I, 2014. Evaluating the impact of pre-annotation on annotation speed and potential bias: natural language processing gold standard development for clinical named entity recognition in clinical trial announcements. *Am Med Inform Assoc* 21:406–413.
- Marcus G, 2022. *Deep Learning Is Hitting a Wall. What would it take for artificial intelligence to make real progress?* Blog post, <https://nautil.us/deep-learning-is-hitting-a-wall-238440/>
- Neme AA, Laporte E, 2013. Pattern-and-root inflectional morphology: the Arabic broken plural, *Language Sciences* 40:221–250.
- Neme AA, Paumier S, 2020. Restoring Arabic vowels through omission-tolerant dictionary lookup, *LRE* 54:487-551.
- Nguyen-Vo TH, Truong D, Nguyen LHB, Dinh D, 2021. Exploring Subword Segmentation Methods in English-Vietnamese Neural Machine Translation. *Advances in Intelligent Information Hiding and Multimedia Signal Processing*, Springer, p. 324–330.

References (3/3)

- The PIAF Project, 2020. *French language keeping pace with AI: FlauBERT, CamemBERT, PIAF*. Blog post, <https://piaf.etalab.studio/francophonie-ia-english/>
- Qiu X, Sun T, Xu Y, Shao Y, Dai N, Huang X, 2020. Pre-trained Models for Natural Language Processing: A Survey. *Science China Technological Sciences* 63, doi: 10.1007/s11431-020-1647-3.
- Sak H, Güngör T, Saraçlar M, 2011. Resources for Turkish morphological processing, *LRE* 45:249-261.
- Scagliarini L, 2022. *Machine Learning and Deep Learning work... in theory*. Blog post, <https://www.linkedin.com/pulse/machine-learning-deep-work-theory-luca-scagliarini/>
- Sculley D, Holt G, Golovin D, Davydov E, Phillips T, Ebner D, Chaudhary V, Young M, 2014. *Machine Learning: The High Interest Credit Card of Technical Debt*. NIPS 2014 Workshop.
- Slawski B, 2019. *Query Pattern Generation at Google*. Blog post, <https://gofishdigital.com/blog/query-pattern-generation/>
- Sperber M, Neubig G, Niehues J, Waibel A, 2017. Neural Lattice-to-Sequence Models for Uncertain Inputs. *EMLNP*, 1380–1389.
- Wang ZJ, Choi D, Xu Sh, Yang D, 2021. Putting Humans in the Natural Language Processing Loop: A Survey. *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 47–52.
- Wei J, Zou K, 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. arXiv preprint arXiv:1901.11196.
- Wissner-Gross A, 2016. Datasets over algorithms. *Edge*, <https://www.edge.org/response-detail/26587>
- Ye Z, Liu P, Fu J, Neubig G, 2021. Towards More Fine-grained and Reliable NLP Performance Prediction. *EACL*.

Eric Laporte

eric.laporte@univ-eiffel.fr

+33 1 60 95 75 52

