Universal Derivations Kickoff:
A Collection of Harmonized Derivational Resources for Eleven Languages

Lukáš Kyjánek, Zdeněk Žabokrtský, Magda Ševčíková, Jonáš Vidra

Charles University, Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics

20th September 2019, DeriMo 2019

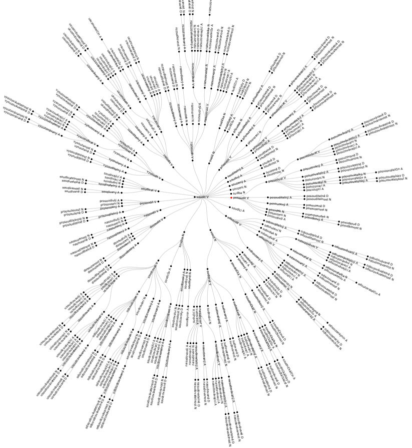
Let me choose any language,
for example English...

Let me choose any language,
for example English. . .

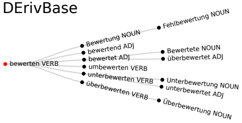
No!

Universal Derivations 0.5

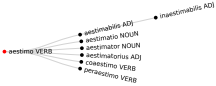
DeriNet



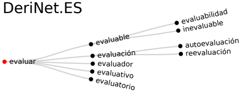
DErivBase



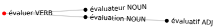
Word Formation Latin



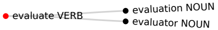
DeriNet.ES



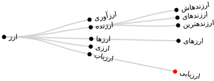
Démonette



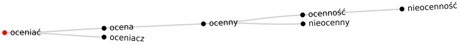
English WordNet



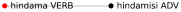
DeriNet.FA



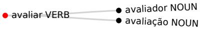
The Polish Word-Formation Network



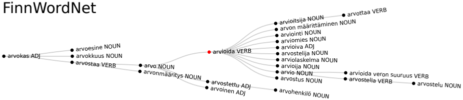
EstWordNet



NomLex-PT



FinnWordNet



Outline

- **Motivation**; the success story of Universal Dependencies
- **Diversity** of existing derivational resources
- Design decision on which our **harmonization** is based. . .
- . . . with a special attention paid to **trees**
- **Universal Derivations** collection – basic properties

Motivation

- Growing interest in derivational morphology in recent...
- 50+ existing derivational data resources for 20+ languages.
- Difficult to work with in a single experiment, because of
  - different **methodology**, different **formal model**,
  - different **file format**, incompatible **software tools** (tools for annotation, querying, visualization etc.)
  - published under various **licenses** (or unpublished), etc.

Multilingual language resources in other domains

Pushing to shared annotation schemes proved very fertile elsewhere, as:

Multilingual language resources in other domains

Pushing to shared annotation schemes proved very fertile elsewhere, as:

- new schemes become **less language dependent**. . .
- . . . and more **independent of local linguistic traditions**,

Multilingual language resources in other domains

Pushing to shared annotation schemes proved very fertile elsewhere, as:

- new schemes become **less language dependent**. . .
- . . . and more **independent of local linguistic traditions**,
- **sharing** software tools (for annotation, visualization, querying...) becomes possible,

Multilingual language resources in other domains

Pushing to shared annotation schemes proved very fertile elsewhere, as:

- new schemes become **less language dependent**. . .
- . . . and more **independent of local linguistic traditions**,
- **sharing** software tools (for annotation, visualization, querying...) becomes possible,
- **lower barrier** for under-resourced languages,

Multilingual language resources in other domains

Pushing to shared annotation schemes proved very fertile elsewhere, as:

- new schemes become **less language dependent**. . .
- . . . and more **independent of local linguistic traditions**,
- **sharing** software tools (for annotation, visualization, querying...) becomes possible,
- **lower barrier** for under-resourced languages,
- **typological studies** become simpler,

Multilingual language resources in other domains

Pushing to shared annotation schemes proved very fertile elsewhere, as:

- new schemes become **less language dependent**. . .
- . . . and more **independent of local linguistic traditions**,
- **sharing** software tools (for annotation, visualization, querying...) becomes possible,
- **lower barrier** for under-resourced languages,
- **typological studies** become simpler,
- competitions in **shared tasks** becomes a huge source of energy.

Multilingual language resources in other domains

Pushing to shared annotation schemes proved very fertile elsewhere, as:

- new schemes become **less language dependent**. . .
- . . . and more **independent of local linguistic traditions**,
- **sharing** software tools (for annotation, visualization, querying...) becomes possible,
- **lower barrier** for under-resourced languages,
- **typological studies** become simpler,
- competitions in **shared tasks** becomes a huge source of energy.

Perhaps the most convincing example: Universal Dependencies!

A brief history of multilingual treebank collections

Some steps in the evolution:

- 2006: 13 languages in the **CoNLL-X** shared task dataset
- 2011: 29 languages in **HamleDT**
- 2019: 85 languages in **Universal Dependencies**

The case of Universal Dependencies

- UD is an obvious success as for the number of languages.
- Resulting from collaboration of a (still growing) community!
- What can we learn from this harmonization story?

Lesson No. 1: gain project momentum from snowballing

- A **positive feedback** effect (snowballing, rich-get-richer principle):
    - the more languages are covered, the more attractive the collection becomes, and the more new languages added . . .

Lesson No. 1: gain project momentum from snowballing

- A **positive feedback** effect (snowballing, rich-get-richer principle):
  - the more languages are covered, the more attractive the collection becomes, and the more new languages added . . .
- Why CoNLL 2006, 2007, or 2009 or HamleDT were not sufficient to start the snowballing?

Lesson No. 1: gain project momentum from snowballing

- A **positive feedback** effect (snowballing, rich-get-richer principle):
  - the more languages are covered, the more attractive the collection becomes, and the more new languages added . . .
- Why CoNLL 2006, 2007, or 2009 or HamleDT were not sufficient to start the snowballing?
- Hard to say.
  - Maybe **super-critical** initial energy investment is needed.
  - Maybe an attractive **brand** matters most. Maybe the **licensing** policy.
  - Maybe they were just lucky.

Lesson No. 1: gain project momentum from snowballing

- A **positive feedback** effect (snowballing, rich-get-richer principle):
  - the more languages are covered, the more attractive the collection becomes, and the more new languages added . . .
- Why CoNLL 2006, 2007, or 2009 or HamleDT were not sufficient to start the snowballing?
- Hard to say.
  - Maybe **super-critical** initial energy investment is needed.
  - Maybe an attractive **brand** matters most. Maybe the **licensing** policy.
  - Maybe they were just lucky.
- Evolution is unpredictable. Still, snowballing can help a lot.

Lesson No. 2: simplicity is the key

[with a little bit of exaggeration]

- **Better simple** than perfectly linguistically adequate.
  - Trees are clearly insufficient for syntax? Who cares, trees are simple, let's start with trees, and the other things can be solved later.
- **Better simple** than expressive.
  - Multilayer schemes are powerful, but complex. Let's start with a single structure for a sentence, the rest will be solved later.
- **Better simple** than flexible.
  - XML is versatile, but non-trivial to process. Let's stick to a simple plain-text file format with a fixed number of columns.

Diversity across word-formation resources

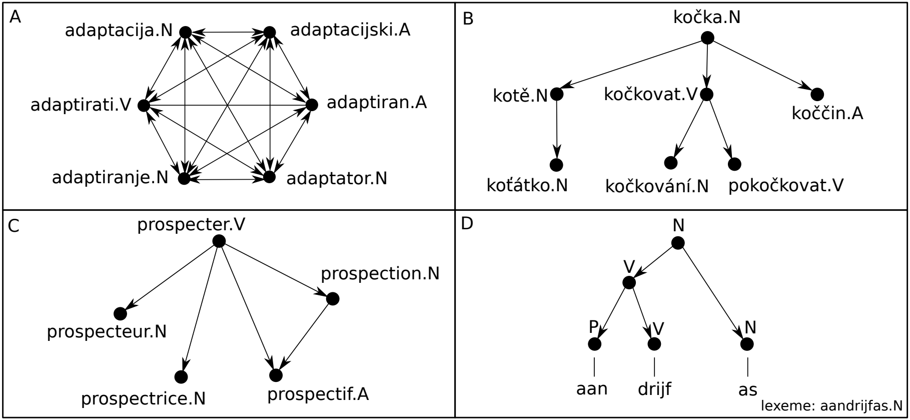
- OK, lessons taken, so let's return to word formation.
- How diverse the existing resources actually are?
- Let's have a look at how a derivational family is represented formally.

Representation of derivational families in existing resources

We observed basically four distinct approaches in which derivational family is represented

1. just as an unstructured set,
2. or as a rooted tree,
3. or as a less constrained graph, e.g. as a weakly connected graph,
4. or just implicitly, by overlaps in constituency trees representing internal structure of a word
5. LEARNED YESTERDAY: morpheme-centric graphs (LiLa)

How do the existing resources represent a derivational family?



Universal Derivations (UDer)

- a newly created collection of word-formation resources
- trying to go as multilingual as possible
- admittedly imitative title
- a shameless attempt at replicating the UD success story
- the current version (UDer 0.5) publicly available in the LINDAT/Clarin repository.

UDer's design decision

- a **lexeme-centric graph-based** approach inherited from DeriNet 2.0:
  - a node represents a lexeme
  - an oriented edge represents a derivational relation
  - a (rooted) tree represents a derivational family
  - the whole vocabulary of a language is represented by a forest
  - additional links can be stored as extra non-tree edges
  - space for other annotation components (morpheme segmentation, semantic labels, etc.)

Why trees?

- Just three conditions implied:
  - acyclic
  - single-rooted
  - connected

Why trees?

- Just three conditions implied:
  - acyclic
  - single-rooted
  - connected
- Is there any risk that some of them is violated in our data?!?

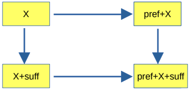
Condition 1: always acyclic?

Condition 1: always acyclic?

- Sometimes violated ☹

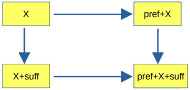
Condition 1: always acyclic?

- Sometimes violated ☹
- Example: a systematic pattern, in which adding a prefix, or adding a suffix, or adding both, produces valid lexemes

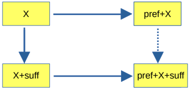


Condition 1: always acyclic?

- Sometimes violated ☹
- Example: a systematic pattern, in which adding a prefix, or adding a suffix, or adding both, produces valid lexemes

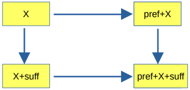


- Luckily, there's a simple workaround ☺: let's **store only a tree-shaped skeleton** (chosen preferably according to some rules) and consider it a shortcut representation for a richer structure.

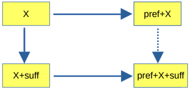


Condition 1: always acyclic?

- Sometimes violated ☹
- Example: a systematic pattern, in which adding a prefix, or adding a suffix, or adding both, produces valid lexemes



- Luckily, there's a simple workaround ☺: let's **store only a tree-shaped skeleton** (chosen preferably according to some rules) and consider it a shortcut representation for a richer structure.



- They do it too in UD! (argumentation by a logical fallacy, hopefully nobody notices): e.g. coordination structures are cyclic, but they're represented as trees in UD.

Condition 2: always single-rooted?

- Sometimes violated too ☹

Condition 2: always single-rooted?

- Sometimes violated too ☹
- Example: composition.

Condition 2: always single-rooted?

- Sometimes violated too ☹
- Example: composition.
- Workaround ☺: let's allow inserting **“second-class” edges**

Condition 2: always single-rooted?

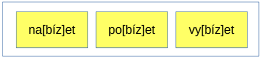
- Sometimes violated too ☹
- Example: composition.
- Workaround ☺: let's allow inserting **"second-class" edges**
- They do it too in UD: secondary predication ("She declared the cake beautiful").

Condition 3: always connected?

- Sometimes violated too ☹

Condition 3: always connected?

- Sometimes violated too ☹
- Example: *nabízet* (to offer) and *pobízet* (to urge) feel as siblings, but no *bízet*.



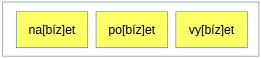
na[bíz]et

po[bíz]et

vy[bíz]et

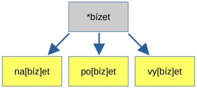
Condition 3: always connected?

- Sometimes violated too ☹
- Example: *nabízet* (to offer) and *pobízet* (to urge) feel as siblings, but no *bízet*.



na[bíz]et po[bíz]et vy[bíz]et

- Workaround ☺: introduce **fictitious lexemes**

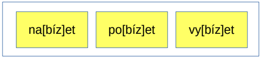


*bízet

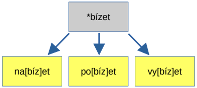
na[bíz]et po[bíz]et vy[bíz]et

Condition 3: always connected?

- Sometimes violated too ☹
- Example: *nabízet* (to offer) and *pobízet* (to urge) feel as siblings, but no *bízet*.



- Workaround ☺: introduce **fictitious lexemes**



- They do it too in UD: “Sue likes coffee and Bill tea.” – an additional node inserted

Once again, why trees?

- A tree is an **irresistibly attractive** data structure.
- Compared to unrestricted graphs, “treeness” simplifies all kinds of algorithmic processing.
- It simplifies any evaluation attempts too, such as measuring inter-annotator agreement or success of cross-lingual projection.

Common, why trees? Seriously!

Common, why trees? Seriously!

- Perhaps the most influential reason: the law of the hammer

Common, why trees? Seriously!

- Perhaps the most influential reason: the law of the hammer

Law of the hammer

A cognitive bias:

- If our basic tool is a hammer, one tends to look for nails.
- In our case: after a decade or two in treebanking, one sees trees everywhere around.

Common, why trees? Seriously!

- Perhaps the most influential reason: the law of the hammer

Law of the hammer

A cognitive bias:

- If our basic tool is a hammer, one tends to look for nails.
- In our case: after a decade or two in treebanking, one sees trees everywhere around.

Conclusion: rooted trees fit derivation perfectly, Q.E.D. ☺

What if the input resource is not tree-based?

- we can't have a cake and eat it
  - harmonization means reducing the diversity
  - e.g., if a weakly connect graph is used to represent a family, we extract its tree-shaped skeleton
- compromise: other information not lost, but stored on a less prominent place

Resources integrated in Universal Derivations 0.5

- Démonette 1.2 (French)
- DeriNet 2.0 (Czech)
- DeriNet.ES (Spanish)
- DeriNet.FA (Persian)
- DErivBase 2.0 (German)
- English WordNet 3.0 (English)
- EstWordNet 2.1 (Estonian)
- FinnWordNet 2.0 (Finnish)
- Nomlex-PT 2017 (Portuguese)
- Polish WFN (Polish)
- Word Formation Latin (Latin)

UDer 0.5 – basic statistical properties

		After harmonization			
Resource	**Language**	**Lexemes**	**Relations**	**Families**	**License**
Démonette 1.2	French	21,290	13,808	7,482	CC BY-NC-SA
DeriNet 2.0	Czech	1,027,665	808,682	218,383	CC BY-NC-SA
DeriNet.ES	Spanish	151,173	36,935	114,238	CC BY-NC-SA
DeriNet.FA	Persian	43,357	35,745	7,612	CC BY-NC-SA
DErivBase 2.0	German	280,775	44,830	235,945	CC BY-SA 3.0
English WordNet 3.0	English	13,813	7,855	5,958	CC BY-NC-SA
EstWordNet 2.1	Estonian	988	507	481	CC BY-SA 3.0
FinnWordNet 2.0	Finnish	20,035	13,687	6,348	CC BY 3.0
Nomlex-PT 2017	Portuguese	7,020	4,201	2,819	CC BY 4.0
Polish WFN 0.5	Polish	262,887	189,217	73,670	CC BY-NC-SA
Word Formation Latin	Latin	29,708	22,641	5,320	CC BY-NC-SA

UDer 0.5 – basic statistical properties, cont.

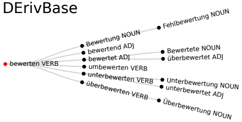
	Singleton		Tree	Tree	Part-of-speech distribution [%]				
Resource	**nodes**	**#Nodes**	**depth**	**out-degree**	**Noun**	**Adj**	**Verb**	**Adv**	**Other**
Démonette 1.2	69	2.8 / 12	1.1 / 4	1.8 / 8	63.0	2.5	34.5	–	–
DeriNet 2.0	96,208	4.7 / 1638	0.8 / 10	1.1 / 40	44.0	34.8	5.5	15.7	–
DeriNet.ES	98,325	1.3 / 35	0.2 / 5	0.3 / 14	–	–	–	–	–
DeriNet.FA	0	5.7 / 180	1.5 / 6	3.3 / 114	–	–	–	–	–
DErivBase 2.0	215,823	1.2 / 51	0.1 / 7	0.1 / 13	85.5	9.9	4.6	–	–
En. WordNet 3.0	65	2.3 / 6	1.0 / 1	1.3 / 6	56.9	–	43.1	–	–
EstWordNet 2.1	21	2.1 / 3	1.0 / 2	1.0 / 3	15.9	29.0	7.9	47.2	–
FinnWordNet 2.0	3	3.2 / 36	1.5 / 9	1.5 / 13	55.3	29.2	15.5	–	–
Nomlex-PT 2017	17	2.5 / 7	1.0 / 1	1.5 / 7	59.8	–	40.2	–	–
Polish WFN 0.5	41,332	3.6 / 214	1.0 / 8	1.1 / 38	–	–	–	–	–
Word Form. Latin	63	5.6 / 130	1.5 / 6	3.0 / 42	46.0	27.4	23.8	–	2.8

Universal Derivations sample, once again

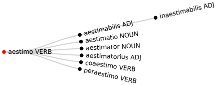
DeriNet

soudit V
posoudit V

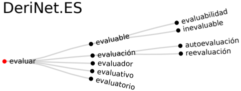
DErivBase



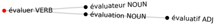
Word Formation Latin



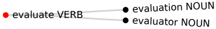
DeriNet.ES



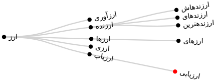
Démonette



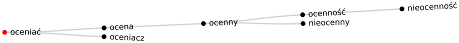
English WordNet



DeriNet.FA



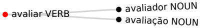
The Polish Word-Formation Network



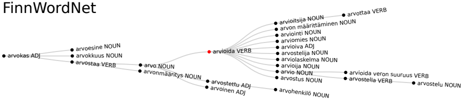
EstWordNet

hindama VERB
hindamisi ADV

NomLex-PT



FinnWordNet



Future perspectives

- We are not dogmatic about UDer's design decisions, not at all.
- Our main **ambition: to provide the initial momentum** and start the snowballing effect.
- Maybe our lexeme-centric representation will serve only as **"Wittgenstein's ladder"**, and will be replaced
  - by a paradigm-centric approach,
  - by a morpheme-centric approach,
  - or by something completely new . . . who knows?

Take home message

- There's a collection of derivational databases for **11 languages** converted into the **same format**.
- Publicly available in the **LINDAT/Clarin** repository under CC.
- **Searchable** using an online query interface.
- We will be happy if you start using it . . .
- . . . and we will be even happier if you allow include your data.

Acknowledgements

We would like to thank
all brave men and women
who made their own derivational resources
publicly available under open licenses.

Time for a demo?

Thank you!

https://ufal.mff.cuni.cz/universal-derivations