



HAL
open science

Universals of Linguistic Idiosyncrasy in Multilingual Computational Linguistics (Dagstuhl Seminar 23191)

Timothy Baldwin, William Croft, Joakim Nivre, Agata Savary, Sara Stymne, Ekaterina Vylomova

► **To cite this version:**

Timothy Baldwin, William Croft, Joakim Nivre, Agata Savary, Sara Stymne, et al.. Universals of Linguistic Idiosyncrasy in Multilingual Computational Linguistics (Dagstuhl Seminar 23191). Dagstuhl Reports, 2023, 13 (5), pp.22-70. 10.4230/DagRep.13.5.22 . hal-04322958

HAL Id: hal-04322958

<https://hal.science/hal-04322958v1>

Submitted on 8 Dec 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Universals of Linguistic Idiosyncrasy in Multilingual Computational Linguistics

Timothy Baldwin^{*1}, William Croft^{*2}, Joakim Nivre^{*3},
Agata Savary^{*4}, Sara Stymne^{†5}, and Ekaterina Vylomova^{†6}

- 1 MBZUAI – Abu Dhabi, AE. tb@ldwin.net
- 2 University of New Mexico – Albuquerque, US. wacroft@icloud.com
- 3 Uppsala University, SE. joakim.nivre@lingfil.uu.se
- 4 University Paris-Saclay, CNRS – Orsay, FR.
agata.savary@universite-paris-saclay.fr
- 5 Uppsala University, SE. sara.stymne@lingfil.uu.se
- 6 The University of Melbourne, AU. ekaterina.vylomova@unimelb.edu.au

Abstract

The Dagstuhl Seminar 23191 entitled “Universals of Linguistic Idiosyncrasy in Multilingual Computational Linguistics” took place May 7–12, 2023. Its main objectives were to deepen the understanding of language universals and linguistic idiosyncrasy, to harness idiosyncrasy in treebanking frameworks in computationally tractable ways, and to promote a higher degree of convergence in universalism-driven initiatives to natural language morphology, syntax and semantics.

Most of the seminar was devoted to working group discussions, covering topics such as: representations below and beyond word boundaries; annotation of particular kinds of constructions; semantic representations, in particular for multiword expressions; finding idiosyncrasy in corpora; large language models; and methodological issues, community interactions and cross-community initiatives. Thanks to the collaboration of linguistic typologists, NLP experts and experts in different annotation frameworks, significant progress was made towards the theoretical, practical and networking objectives of the seminar.

Seminar May 7–12, 2023 – <https://www.dagstuhl.de/23191>

2012 ACM Subject Classification Computing methodologies → Artificial intelligence

Keywords and phrases computational linguistics, morphosyntax, multiword expressions, language universals, idiosyncrasy

Digital Object Identifier 10.4230/DagRep.13.5.22

* **Editor / Organizer**

† **Editorial Assistant / Collector**



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 4.0 International license

Universals of Linguistic Idiosyncrasy in Multilingual Computational Linguistics, *Dagstuhl Reports*, Vol. 13, Issue 5, pp. 22–70

Editors: Timothy Baldwin, William Croft, Joakim Nivre, Agata Savary, Sara Stymne, and Ekaterina Vylomova



Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Executive Summary

Agata Savary (University Paris-Saclay, CNRS – Orsay, FR)

Timothy Baldwin (MBZUAI – Abu Dhabi, AE)

Joakim Nivre (Uppsala University, SE)

William Croft (University of New Mexico – Albuquerque, US)

License © Creative Commons BY 4.0 International license
© Agata Savary, Timothy Baldwin, Joakim Nivre, William Croft

The Dagstuhl Seminar 23191 entitled “Universals of Linguistic Idiosyncrasy in Multilingual Computational Linguistics” was an accomplishment of long-standing efforts, initiated as early as in October 2018. We submitted at that time a Dagstuhl Seminar proposal which was selected to take place in Dagstuhl on June 21–26, 2020. Due to the Corona/COVID-19 pandemic, the event was first re-scheduled to August 29 to September 3, 2021, and finally transformed into a reduced online seminar under the same title on August 30–31, 2021 [1]. Despite its very reduced format, the seminar achieved part of its objectives and provided a proof of concept of the initial proposal. Following the encouragement from the participants, we re-submitted roughly the same proposal in November 2021 for a full-fledged on-site event. It was then selected to take place in Dagstuhl on **May 7–12, 2023**.

The objectives, following the initial 2018 proposal, were threefold:

- **Theoretical:** To deepen the understanding of language universals, and of linguistic idiosyncrasy in particular, so as to further promote unified modelling while preserving diversity.
- **Practical:** To harness idiosyncrasy in treebanking frameworks, in computationally tractable ways and, thus, to foster high quality NLP tools for very many languages.
- **Networking:** To promote a higher degree of convergence to universalism-driven initiatives, while focusing on three main aspects of language modelling: morphology, syntax, and semantics.

The program of the event followed the Dagstuhl model:

- A list of **recommended readings** was published prior to the event.
- Recordings from the **introductory talks**, given by the 4 organizers at the 2021 online seminar, ensured common understanding of the terminology, scope and challenges to address.
- **Personal introductions** of all participants helped achieve a community building effect.
- Six outstanding speakers were invited to give plenary **inspirational talks**.
- **Working groups** (WGs) were built in a bottom-up manner on the basis of discussion issues submitted by the participants. WGs ran in parallel, were coordinated and minuted by two co-leaders each, and were organized in the following settings.
- For **days 1 and 2** (Monday-Tuesday) the discussion issues were submitted by the participants prior to the event. On this basis 5 WGs were formed:
 - WG1 – *Below and beyond word boundaries* (co-leaders: Daniel Zeman and Reut Tsarfaty)
 - WG2 – *Annotation of particular kinds of constructions* (co-leaders: Manfred Sailer and Nathan Schneider)
 - WG3 – *Representing the semantics of MWEs* (co-leaders: Dag Haug and Nianwen Xue)
 - WG4 – *Finding idiosyncrasy in corpora* (co-leaders: Francis Bond and Nurit Melnik)
 - WG5 – *Methodological Issues and community interactions* (co-leaders: Amir Zeldes and Gosse Bouma)

- Day 3 was dedicated to **reporting**, collecting new issues and re-designing the WGs.
- As a result, 5 other WGs were formed for **days 4 and 5** and reported on on day 5:
 - WG6 – *Below and beyond word boundaries* (co-leaders: David Yarowsky and Omer Goldman), continuation of WG1
 - WG7 – *Construction grammar meets Universal Dependencies* (co-leaders: Lori Levin and Peter Ljunglöf), continuation of WG2 and WG4
 - WG8 – *To semantics and beyond!* (co-leaders: Archana Bhatia and Kilian Evang)
 - WG9 – *Cross community/formalism discussions (big, hairy problems)* (co-leaders: Chris Manning and Laura Kallmeyer)
 - WG10 – *Large language models (and other NLP tools)* (co-leaders: Francis Tyers and Mathieu Constant)¹
- Wednesday afternoon featured a **hike** in the surrounding countryside.
- The evenings were dedicated to **socializing**. This included a piano-violin duet, a guitar duet, a jazz improvisation, a swing dancing duet, and a choir singing songs suggested by the participants, in English, Georgian, German, and Latin (for the sake of language diversity!).

All the inputs and instantaneously produced outcomes (minutes, slides, useful links) are downloadable from our Wiki space.²

The event attracted **37 participants**. Their feedback during and after the event was mostly enthusiastic. At least one group formed at the event continues online meetings to further discuss the scientific challenges (representation of constructions in the Universal Dependencies framework).

Based on the reports submitted by the WG co-leaders and by individual proposers of discussion issues, we can estimate the extent to which the event achieved its initial objectives.

- On the **networking** side, the seminar brought together several pre-existing communities and allowed them to achieve synergies:
 - Linguistic experts specialized in analyzing constructions and collecting them in so-called constructicons, intensely collaborated with NLP experts, notably over the problem of how to represent constructions formally and query them in corpora.
 - While the community of typology experts was unfortunately under-represented (despite the best efforts of the organizers), the few attending experts were frequently consulted, which yielded several enlightening discussions.
 - The communities of Universal Dependencies (UD) and Universal Morphology (UniMorph) converged, even further than initially expected, around the problems of annotating subword units.
 - The communities of UD and PARSEME, which had started aligning objectives prior to the seminar, further strengthened coordination.
 - New links were established between PARSEME and the Universal Meaning Representation (UMR) community. This effect is important since the former models lexical and morpho-syntactic properties of MWEs, while the latter offers a framework for representing their semantics.
 - An unplanned networking effect also occurred between our seminar and Dagstuhl Seminar 23192 on “Topological Data Analysis and Applications”, running in parallel on the same site. Bei Wang Phillips (University of Utah – Salt Lake City, US) gave an evening invited talk to our invitees on the applications of topological methods to interpretability of word embeddings in distributional semantics.

¹ No report was provided for this group, which only met for a short session before splitting into other groups.

² <https://gitlab.com/unlid-dagstuhl-seminar/unlid-2023/-/wikis/home>

- On the **theoretical** side, the seminar focused even more than expected on the notion of construction, which is broader and harder to capture than multiword expressions, and has been defined in wildly divergent ways across different communities. The confluence of different communities led to theoretical results including the following:
 - Steps were taken towards a formal definition of construction, as an expression in a formal graph language (similar to the one supported by the Grew-match corpus browser)
 - Advances in formalizing the notion of an “interesting” construction, which relates to the notion of idiosyncrasy, a core concept in a narrower guise in the multiword expression community
 - Formalizing the task of searching for “a similar but different construction” as an instance of the theoretical problem of approximate tree/graph matching
 - Progress towards understanding the notion of idiosyncrasy as an instance of rule breaking which is “creative” and “has a purpose”, as opposed to, for instance, plain grammar/spelling errors (rule breaking with no purpose)
 - Understanding idiosyncrasy via cross-linguistic triangulation – what is seen as idiosyncratic in one language can be systematic across many languages/language families (e.g. kinship terms)
 - Progress towards formalizing the annotation of semantics of UD and multiword expressions, especially for temporal and negation expressions

We also addressed a major challenge in language technology, which is a universal definition of the notion of a word. Namely, proposals emerging from WG1 and WG6 suggest that the difficult challenges for defining wordhood across languages should be alleviated by lifting the constraint of a rigid segmentation of a sentence into words prior to linguistic analysis. Instead, proposals of formats allowing different granularity of description items (below and beyond the word level) were suggested and discussed.

- On the **practical** side, discussions at the seminar led to a number of proposals for tools, procedures or practices to support interdisciplinary research. Some of these were tested out already in Dagstuhl, while others are being realized in follow-up activities to the seminar. The following is a non-exhaustive list of examples:
 - Practical steps were taken towards improved UD guidelines for multiword expressions, which will facilitate interfacing UD and PARSEME in the future.
 - Concrete guidelines were drafted for representing subword units in UD, which will facilitate the integration of resources from UD and UniMorph.
 - Discussions of construction-oriented UD guidelines (based on “a Swadesh list for morphosyntax”) resulted in a prototype implementation with links to annotation examples in different languages.
 - Discussions of future extensions of UD explored concrete proposals for new feature mechanisms to incorporate notions of constituency.
 - Practical exercises demonstrated how the grew-match system (and other search tools) can be used to search for constructions in linguistic corpora.
 - Participants discussed concrete proposals for automatically identifying idiosyncratic phenomena in corpora.

The survey organized by the Dagstuhl Officers shortly after the event shows very encouraging results (in most categories it was ranked higher than the average of the Dagstuhl Seminars from the past 60 days). The major drawbacks noticed by the participants were the insufficient number of experts in typology (less than 5%),³ and of young researchers (about 32%).

References

- 1 Timothy Baldwin, William Croft, Joakim Nivre, and Agata Savary. 2021. Universals of Linguistic Idiosyncrasy in Multilingual Computational Linguistics (Dagstuhl Seminar 21351). *Dagstuhl Reports*, 11(7), pages 89–138.

³ This was notably due to the last minute cancellation, for personal reasons, by William Croft, one of the 4 co-organizers of the event.

2 Table of Contents

Executive Summary

Agata Savary, Timothy Baldwin, Joakim Nivre, William Croft 23

Overview of Talks

Adrian’s Fish Tail: Compounds and Adnominal Possession Across Languages
Maria Koptjevskaja-Tamm 29

What Kinds of Parts do Multi-part Expressions Have?
Lori Levin 30

Universal Dependencies: Its Multilingual NLP Successes and Other Surprising Impacts
Christopher Manning 30

Indigenous Voices from the Past: Opening up the Florentine Codex to Modern Digital Scholarship
Francis M. Tyers 30

Working groups

WG1: Above and Below Word Level
Daniel Zeman and Reut Tsarfaty 31

WG2: Annotation of Particular Constructions
Nathan Schneider and Manfred Sailer 35

WG3: Semantics of Multi-Word Expressions
Dag Haug and Nianwen Xue 37

WG4: Finding Idiosyncrasy in Corpora
Nurit Melnik and Francis Bond 39

WG5: Methodological Issues and Community Interactions
Gosse Bouma and Amir Zeldes 44

WG6: Above and Below Word Level
David Yarowsky and Omer Goldman 46

WG7: UniCoDeX (Universal Construction Dependency Xrammar)
Peter Ljunglöf and Lori Levin 47

WG8: To Semantics and Beyond
Archna Bhatia and Kilian Evang 53

WG9: Fostering Corpus-based Typology [“Big Hairy Problems”]
Laura Kallmeyer and Christopher Manning 57

Open Problems

Semantic Parsing and Sense Tagging the Princeton WordNet Gloss Corpus
Alexandre Rademaker, Francis Bond, and Daniel Flickinger 60

NLP-based Study of Universals of Linguistic Idiosyncrasy
Agata Savary 64


Subword Relations, Superword Features
Daniel Zeman 67

Participants 70

3 Overview of Talks

3.1 Adrian’s Fish Tail: Compounds and Adnominal Possession Across Languages

Maria Koptjevskaja-Tamm (Stockholm University, SE)

License  Creative Commons BY 4.0 International license
© Maria Koptjevskaja-Tamm

As is well known, adnominal possession is not restricted to possession *stricto sensu*, but can cover many other relations, e.g., *Adrian’s house / sister / finger / school* etc. Typical possessors act as *anchors* or *reference point entities* for identification of the head, and the whole construction can therefore be said to denote *anchoring relations*. In many languages these are clearly distinguished from expressions used for *typifying relations*, i.e., for *qualifying* classes of entities via their relations to other entities. To give a couple of examples, typifying relations are expressed by noun phrases with adjectives derived from nouns in Russian (e.g., *kofe-jn-aja čaška* “coffee-ADJ-F.SG.NOM cup” = “a coffee cup” and *ryb-ij xvost* “fish-ADJ.M.SG.NOM tail” = “a fish tail”) and by noun-noun compounding in Swedish (*fisk+stjärt* “fish+tail” and *kaffe+kopp* “coffee+cup”), whereas the standard possessive construction in both languages contains the possessor in the genitive case. Other languages, however, utilize identical or, at least, very similar constructions for both anchoring and typifying relations. This is the case with adnominal dependents in the genitive case in Lithuanian, e.g., *Adrian-o namas* “Adrian-GEN house” = “Adrian’s house” and *kavos puodelis* “coffee:GEN cup” = “a coffee cup”. The Lithuanian phrase *žuvies uodega* “fish:GEN tail” may therefore refer to a tail of a particular fish, but also denote a class of tails that share certain properties without necessarily being a part of a fish (as those belonging to mermaids). The cross-linguistic variation exemplified by Russian, Swedish and Lithuanian is not surprising. The rationale for a similar treatment of anchoring and typifying relations is obvious – both types of adnominal dependents characterize entities via their relations to other entities. On the other hand, typifying adnominals differ in that 1. the dependent is not individualized; 2. the dependent-head combination refers to a subclass of a broader class and often functions as a classificatory label for it, suggesting that the dependent and the head together correspond to one concept; 3. the head cannot be identified via its relation to the dependent. In my talk I present a typology of the formal ways in which European languages deal with the distinction between anchoring and typifying relations and suggest several generalizations on the form-function correlations in this area. The insights gained from the talk may have consequences for syntactic and semantic annotation of multilingual language resources and tools, including the perennial issue of the border between words, multi-word expressions and regular syntactic phrases.

References

- 1 Maria Koptjevskaja-Tamm. 2005. Maria’s ring of gold: adnominal possession and non-anchoring relations in the European languages. In Kim, Ji-yung, Yu. Lander, and B. H. Partee (Eds.), *Possessives and Beyond: Semantics and Syntax*, pages 155–181. Amherst, MA: GLSA Publications.
- 2 Maria Koptjevskaja-Tamm. 2002. Adnominal possession in the European languages: form and function. *Sprachtypologie und Universalienforschung (STUF)*, 55(2), pages 141–172.

3.2 What Kinds of Parts do Multi-part Expressions Have?


Lori Levin (Carnegie Mellon University – Pittsburgh, US)

License  Creative Commons BY 4.0 International license
© Lori Levin

This talk distinguished multi-word expressions from multi-part expressions, where the parts are not necessarily words. The English causal excess construction (e.g., It was so big that it fell over) was presented as an example of a multi-part expression where the parts are words, parts of speech, morphological features, and abstract syntactic processes. The talk also addressed morphosyntactic strategies, specifically the issue of representing the meaning of constructions in strategy-neutral semantic frames.

3.3 Universal Dependencies: Its Multilingual NLP Successes and Other Surprising Impacts

Christopher Manning (Stanford University, US)

License  Creative Commons BY 4.0 International license
© Christopher Manning

Joint work of Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, Daniel Zeman
Main reference Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, Daniel Zeman: “Universal Dependencies”, *Computational Linguistics*, Vol. 47(2), pp. 255–308, MIT Press, 2021.
URL http://dx.doi.org/10.1162/coli_a_00402

This talk outlines some of the design goals of the morphosyntactic annotation framework Universal Dependencies (UD), emphasizing how they differ from previous NLP practice and the natural predilections of linguists. In particular, for broad usage, “less is more”. In NLP, UD enables new tools that work well from raw text on dozens of languages and these tools are still improving and used, despite the dominance of Large Language Models in the field this decade. But a new arc of work is using UD resources for typological linguistic and psycholinguistic research. This work has somewhat different design goals, motivating a more surface structure oriented representation, as has been explored in Surface Syntactic Universal Dependencies (SUD), and there is room to do more here. But, overall, the success of UD and its wide adoption represents a success for linguistics – and a win for non-linguists who want simple, useful language processing tools.

3.4 Indigenous Voices from the Past: Opening up the Florentine Codex to Modern Digital Scholarship

Francis M. Tyers (Indiana University – Bloomington, US)

License  Creative Commons BY 4.0 International license
© Francis M. Tyers

Joint work of Francis M. Tyers, Robert Pugh, Valery A. Berthoud F.

The Florentine Codex is a bilingual text produced in Mexico in the 16th century. It describes the lives and beliefs of the Indigenous people who were living in the Valley of Mexico before the arrival of the Spanish. The text is in Nahuatl, the lingua franca of the area, and is accompanied by a translation or summary in Spanish. In this talk we describe the processing of the Nahuatl text and some linguistic issues that will need to be addressed in the annotation,

including relational nouns, functional incorporation and subordination and clause structure. We also describe efforts to translate from the Nahuatl of the era to modern varieties of Nahuatl spoken in the Sierra of Puebla.

4 Working groups

4.1 WG1: Above and Below Word Level

Daniel Zeman (Charles University – Prague, CZ) and Reut Tsarfaty (Bar-Ilan University – Ramat Gan, IL)

License © Creative Commons BY 4.0 International license
© Daniel Zeman and Reut Tsarfaty

Joint work of Daniel Zeman, Reut Tsarfaty, Omer Goldman, Sylvain Kahane, Sara Stymne, Francis Tyers, Ekaterina Vylomova, David Yarowsky

Goals

- To extend the UD representation in a way that can accommodate complex morphosyntactic phenomena (further discussed in WG6)
- To provide a proof of concept annotation for languages with no 1:1 mapping between segments and morphosyntactic nodes
- To improve parallel representation of argument structure between languages with radically different realization mechanisms (polysynthetic vs isolating)
- To discuss the role / contribution of derivational morphology (further discussed in WG6)
- To discuss how morphological marking complements compound and MWE
- To help field linguists who want to represent segmentation down to morphs
- To ultimately propose guidelines for cross-linguistically consistent annotation of polysynthetic (and in particular polypersonal agreement, noun-incorporating) languages

Representing phrase-level information in phrase-level features in UD, somehow comparable to representations for languages in which the same information is expressed in a single word (or a sublexical morph) was also discussed in WG4 and WG9.

Mechanisms that have been proposed

- “Empty nodes” (= abstract nodes) for indicating pronominal feature bundles
- “Empty nodes” carrying the lemma of an incorporated noun (not necessarily a linear segment)
- A layered representation of phrase-level features for each (lexical) node which is distinct from the word-level features that the lexical item contributes. (“Layered” as in layered features, which exist as language-specific in UD, and have also been introduced in UniMorph 4.0.)
- Linear segmentation (of clitics etc) is optional, not mandatory

Implication

To allow for the free use of these mechanisms to express non-explicit morphological phenomena, and still stay faithful to the UD guidelines, it has been proposed that the kind of representation we develop is a part of the enhanced, rather than basic, UD trees.

Empty (abstract) nodes are part of the enhanced UD graph (but not of the basic UD tree). (Note: They are called *empty* nodes in the current UD documentation but it is a misnomer because they often are not really empty, they may have a word form, UPOS tag etc. So we propose that UD should switch to a different term, such as *abstract* nodes.) Enhanced UD already contains abstract nodes that represent elided predicates in gapping constructions. If we now add abstract nodes for a different purpose, namely to represent segments of surface words, we need a way of distinguishing different types of abstract nodes. We also need to identify the surface word to which the segment (abstract node) belongs. Therefore, the segment-abstract nodes could be identified by `PartOf=ID` in the MISC column, where `ID` is the ID of a regular node in the basic UD tree.

Enhanced UD seems to be a suitable area where segments and their relations could be represented. It is still part of the UD specification (as opposed to add-ons built on top UD, using the CoNLL-U Plus format), meaning that such data could be part of official UD releases. At the same time, the Enhanced UD guidelines are less developed and frozen than the Basic UD guidelines, so it should be easier to add new guidelines here. Enhanced annotation is considered optional in UD corpora and can be easily separated from the basic annotation, hence the additional complexity can be completely transparent for users who are not interested in it. These are the main reasons why we propose to use the enhanced representation for subword relations, as opposed to the multi-word token mechanism (MWT), which would lead to the new annotation being visible also in the basic representation. (Moreover, the assumption about MWT is that an orthographic word is split into multiple morphosyntactic words. If we also use it to further split morphosyntactic words into morphs, we will have to solve the problem of distinguishing the different levels of granularity and different types of units.)

We have not come to a conclusion about the labels of the relations between subword units. We have identified two levels of granularity that might be of interest:

- Decomposition to lexical units: compounds and incorporation
- Complete segmentation to morphs

Technical details of the proposal

- Phrase-level features are put to the MISC column of the node whose subtree contains all nodes that belong to the phrase. Not all nodes in the subtree belong necessarily to the phrase, so the phrase has to be specified using node ids. It can be discontinuous. A possible representation is like this: `Phrase=1,3-5,7`
- The features have to be distinguished from other things that may be present in the same MISC cell. For UD-style features, prefixing the feature name with “Phrase” might work: `PhraseAspect=Prog|PhraseTense=Pres`. Datasets that use UniMorph-style features instead might need just one string: `PhraseUniMorph=V;PROG;PRS`.
- UD documentation should retire the term “empty node” and switch to “abstract node”, as these nodes often are not empty in the sense of not having any lexical or morphological value.
- If the dataset contains segmentation of syntactic words to smaller units that are not syntactic words, the abstract nodes should be used. Consequently, the segmentation is only visible in the enhanced representation while the basic tree stays reasonably simple.
- As abstract nodes are already used for other purposes than morphological segmentation, the nodes resulting from segmentation should be distinguished from other abstract nodes. Specifically, an abstract node (of the old kind) is not considered to correspond to any part of any surface token. The rule for the abstract nodes resulting from segmentation would

be that they appear between the node corresponding to the surface token they are part of, and the next node (abstract or regular). Each abstract node resulting from segmentation would have in MISC a reference to its corresponding surface token or syntactic word: `PartOf=2`.

- While an abstract segment-node knows to which surface token it belongs, it does not have to declare precisely which substring of the surface token it represents. Consequently, the order of the abstract nodes is not prescribed, although annotators are encouraged to follow the ordering of the corresponding morphs where it is observable.
- The abstract segment-nodes have to be connected in the enhanced graph, if for nothing else, then to maintain compatibility with Enhanced UD. It is yet to be seen whether and to what extent it is useful to define “syntactic” relations between the segments. As a minimum, the main lexical root has to be declared as the head and the other segments can be attached directly as its dependents. The relation labels (deprels) have to be taken from the UD repository. In some cases `compound` could be used. Cases with no better option could use a subtype of `dep`, such as `dep:infl`.
 - Languages with incorporation will attach the incorporated segments as core arguments of the verb: `obj`.
 - Another possible usage of the relations between segments is to show the order of derivation. CCG-like categories might then be added to MISC to signal that this morph combines with a `VERB` and once combined, the result is an `ADJ`.
 - Also note that some languages seem to have examples where another word would modify just one segment of the current word (Turkish *mavi arabadakiler* “those in the blue car”; lit. “blue car-in-those”). Here the enhanced graph would have an `amod` relation between *araba* and *mavi*, while in the basic tree it would go directly between *arabadakiler* and *mavi*.
- Two possible levels of segmentation are envisioned (both are optional, so people do not have to segment if they do not want to): 1. just split compounds (including incorporated nouns); 2. segment all the way down to morphs. Splitting compounds is useful for cross-linguistic parallelism. Complete segmentation is useful for field linguists who want to represent it, including features and glosses. And of course, there is a third level of segmentation, which already exists in UD: multi-word orthographic tokens are split to syntactic words; this one is done in basic UD and does not require abstract nodes.
- The `FORM` of an abstract segment can be empty (underscore), but it can be non-empty where it makes sense. The `LEMMA` should have a canonical form of that segment (e.g. English prefixes *in-* and *im-* would share the canonical form *in*).
- It is possible to say that some forms in a paradigm table are segmentable while others are not. For instance, one could say that the English verb *closed* is segmented to *close + d*, where the suffix is the bearer of the feature `Tense=Past`, but if the verb is irregular like *made*, it can stay unsegmented and bear the feature `Tense=Past` as a whole.
- If a user/application only wants to work with basic trees, the segmentation will be transparent for them. However, it is also possible that they want to work with the enhanced graph for other reasons but they still do not want to see the segmentation (or they want to see the compound level but not the complete decomposition to morphs). For that purpose we need an algorithm that will only extract the enhanced graph over full syntactic words. This has to be worked out. (Also, both the head of the subtree of segments and the original unsegmented word need to be attached somewhere in the enhanced graph.)

- In general it would be useful if one can say what is the backbone tree in the enhanced graph, and which edges are extra (creating reentrancies and cycles). This is currently not possible; one would have to modify the labeling schema for enhanced relations. Note that the backbone enhanced tree is not necessarily identical to the basic UD tree, as it can contain abstract nodes. The current proposal does not (yet) say how this should be done.
- Note that an abstract node may be also needed to represent a participant of an event (subject, object, oblique) which is not overtly represented as a word. This may correspond to an abstract segment-node under the verb, if the participant is referenced by the verbal morphology, but it is also possible that there is no trace of it in morphology and we still need to represent it e.g. to annotate coreference. (Conversely, in some languages a participant may be overtly referenced multiple times in the same clause: clitic doubling, full noun phrase, and verbal morphology.)

References to related UD issues

- <https://github.com/UniversalDependencies/docs/issues/701>
- <https://github.com/UniversalDependencies/docs/issues/703>
- <https://github.com/UniversalDependencies/docs/issues/704>

Future work

- Do a pilot segmentation of compounds in Parallel UD treebanks (PUD, 1000 news and Wikipedia sentences per language). Try a subset of PUD languages, especially those with lots of compounding, such as German and Swedish. Examine parallelism in word (morph) alignment.
- Convert UD morphological features in UD treebanks to UniMorph (update the conversion procedure, originally tested with UniMorph 2.0, to the latest set of UD features and the latest version of UniMorph (4.0). For each word form, compare the UD annotation with the corresponding entry in UniMorph word lists, if available. Possibly improve UD and/or UniMorph data based on the comparison. Prepare a new version of UniMorph where a new column will say for each word form its frequency in UD treebanks.
 - Expand the UD-UniMorph comparison to phrase-level morphology, such as periphrastic tense, aspect or voice.

References

- 1 Timothy Baldwin, William Croft, Joakim Nivre, and Agata Savary. 2021. Universals of Linguistic Idiosyncrasy in Multilingual Computational Linguistics (Dagstuhl Seminar 21351). *Dagstuhl Reports*, 11(7), pages 89 – 138.
- 2 Martin Haspelmath. 2022. Draft. Defining the Word.
- 3 Sylvain Kahane, Martine Vanhove, Rayan Ziane, and Bruno Guillaume. 2021. A morph-based and a word-based treebank for Beja. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, pages 48–60, Sofia, Bulgaria. Association for Computational Linguistics.
- 4 Büşra Marşan, Salih Furkan Akkurt, Muhammet Şen, Merve Gürbüz, Onur Güngör, Şaziye Betül özateş, Suzan üsküdarlı, Arzucan özgür, Tunga Güngör, and Balkız öztürk. 2022. *Enhancements to the BOUN treebank reflecting the agglutinative nature of Turkish*. arXiv preprint arXiv:2207.11782.
- 5 Francis Tyers and Karina Mishchenkova. 2020. Dependency annotation of noun incorporation in polysynthetic languages. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 195–204, Barcelona, Spain (Online). Association for Computational Linguistics.

4.2 WG2: Annotation of Particular Constructions

Nathan Schneider (Georgetown University – Washington, DC, US) and Manfred Sailer (Goethe-Universität Frankfurt am Main, DE)

License © Creative Commons BY 4.0 International license

© Nathan Schneider and Manfred Sailer

Joint work of Nathan Schneider, Manfred Sailer, Christopher Manning, Maria Koptjevskaja-Tamm, Jörg Bucker, Gosse Bouma, Lori Levin, Timothy Baldwin, Amir Zeldes

WG2 discussed challenges in syntactic annotation, with particular attention to the Universal Dependencies (UD) framework. The discussion featured issues with current UD annotation policies/guidelines as well as some broader challenges. The UD guidelines issues were (i) two possible changes to the relation inventory, and (ii) mischievous nominal constructions. The other main issue (not specific to UD) was (iii) negation and idiomaticity. discussion also touched briefly on the relationship between Construction Grammar and UD, but this was deemed more appropriate for a new working group. For this topic see the report from WG7.

Possible Changes to the UD Relation Inventory

Two specific objections to the current (UDv2) relations were discussed. First, there is a top-level **indirect object** relation (`iobj`), but it has been a source of confusion, owing in part to the diverse range of uses of the term “indirect object” in different linguistic traditions. We concluded that `iobj` should be considered for removal in UDv3, with multiple `obj` for double object constructions (different phrases meeting the UD criteria for `obj` could optionally distinguished via subtypes). Second, UD’s broad interpretation of the **adverbial clause** relation (`advcl`) lumps together adjunct clauses and complement clauses with oblique-like marking.⁴ The group agreed to endorse the idea of a new `cobl` relation in UDv3 that would target oblique subordinate clauses [1].

Mischievous Nominal Constructions

Various specialized patterns in names, numbers, dates, and measurements are known to be challenging for syntactic annotation [2, 3]. We discussed cases like “Eminent linguist Mr. Bill Croft”, where the phrase “eminent linguist” and the title “Mr.” are arguably modifiers as they can be freely omitted. We agreed that the UD guidelines should be changed to treat these as modifiers, adopting the suggested label `nmod:desc` (for “descriptor”) to distinguish these from other kinds of `nmod` [2].

More controversial were date expressions (“*July 30, 1980*” – should this be considered headed?), numbered entities (“*Room S108*”), and other names with an entity type (“*Michigan Street*”, “*Lake Michigan*”) or suffix (“*Richard III*”, “*BMW Inc.*”). Some of these cases are currently listed as `flat` in the guidelines. It emerged from the discussion that there were two different interpretations of `flat` held by members of the group. One interpretation is that `flat` is an “escape hatch” for tricky cases related to names (etc.) which do not easily fit general-purpose syntactic constructions. This is the expansive view of `flat`. Another interpretation is that `flat` should be reserved for expressions that are truly **headless**, defying all possible attempts to identify one part as the syntactically most important element. We leave these issues to further discussion.

⁴ Consider the sentence “As it was raining, we worked on improving annotation.” The second subordinate clause, “on improving annotation”, is not *adverbial* in the ordinary use of the term, but rather is a complement clause with prepositional marking. “As it was raining” is a true adverbial clause adjunct, and should remain `advcl`.

Negation and Idiomaticity

Negation is a grammatical category in all human languages. However, there is no explicit uniform encoding of negation in treebanks or in UD. Negation is also often expressed through a combination of different morphosyntactic elements, but, being a grammatical category, it is usually not considered a multiword expression or a phraseme. The purpose of the discussion point was to raise awareness for this phenomenon and to explore where to locate the representation of negation in annotated corpora. The following data can serve to illustrate the phenomenon.

Examples of morphosyntactic strategies for clausal negation include:

- In German, a clause can be negated by simply adding the negative adverb *nicht* “not”.
- In standard French, clausal negation is expressed by a combination of a pre-verbal particle, *ne*, and post-verbal *pas*, as in *Il ne pleut pas*. “It isn’t raining.” In this example, negation is expressed through two elements, both of which only express the category of negation.
- We also find so-called neg-words such as English *nothing*, i.e., indefinites that fulfill another dependency in a clause but, in addition, also mark negation, as in English *Alex said nothing*.

It is common that languages can use more than one of these elements to express a single negation, as in the French example *Personne ne fait confiance à personne*. (gloss: nobody NE makes confidence to nobody) ‘Nobody trusts anybody.’

Negation itself is not a uniform phenomenon on the meaning and usage side either, i.e., morphosyntactic negation marking can encode clausal negation, but also constituent negation, meta-linguistic negation, or expletive negation. A negative meaning can be explicit, or can be inferred. In addition, we find idiomatic combinations that express negation of various types: The German expression *einen Dreck* (lit.: “a dirt”) “nothing” marks a sentence as morphosyntactically negative. The English expression *I’ll be damned if . . .* has the effect that the *if* clause is semantically, though not morphosyntactically, negative. Finally, we find items that mark only pragmatically inferable negativity like sarcasm or irony. An example is German clause final, intonationally separated *. . . – also nicht!* (lit.: “dots thus not”) or *. . . und ich bin der Kaiser von China* “. . . and I am the emperor of China,” as in *Alex ist echt schlau – also nicht/und ich bin der Kaiser von China!* “Alex is really clever – certainly not!”

The various types of semantic negativity are also relevant for the treatment of MWEs, as we find lexical expressions that are restricted to occur in clauses with a particular type of negation, so-called *negative polarity items*. In the simplest case, there is a fixed expression with a particular negator, such as the German bound word *Unterlass* “stop”, which only occurs in the combination *ohne Unterlass* “without stop”. The German modal verb *brauchen* “need” requires a semantically negative clause, but the choice of the negation strategy is irrelevant. From a collocational or MWE perspective, *brauchen* would not be a single word, but rather form a collocation or MWE together with the semantic category of negation. Finally, NPIs such as English *lift a finger* or *fine* in cases with an explicit semantic negation, but also in some cases of (conventionalized) pragmatic negation, such as in denial (*But I DID lift a finger*). However, sarcasm and irony (i.e., only conversational pragmatic negation) do not seem to license any known NPIs.

Turning back to the seminar theme: Multiple exponence of negation is not marked as a dependency in UD, nor as a MWE in PARSEME. If Negation is a morphosyntactic category, shouldn’t it be marked? Yes, but probably as a grammatical category at the clausal level. As for the other phenomena, negation and negation-related phenomena might be a good motivation and testing ground for adding constructional information (as proposed by WG7) in addition to syntactic dependencies and classical MWEs to corpus annotation.

References

- 1 Adam Przepiórkowski and Agnieszka Patejuk. 2018. Arguments and adjuncts in Universal Dependencies. In *Proceedings of COLING*, pages 3837–3852. Santa Fe, New Mexico, USA.
- 2 Nathan Schneider and Amir Zeldes. 2021. Mischievous nominal constructions in Universal Dependencies. In *Proceedings of UDW*, pages 160–172. Sofia, Bulgaria.
- 3 Daniel Zeman. 2021. Date and time in Universal Dependencies. In *Proceedings of UDW*, pages 173–193, Sofia, Bulgaria.

4.3 WG3: Semantics of Multi-Word Expressions

Dag Haug (*University of Oslo, NO*) and Nianwen Xue (*Brandeis University – Waltham, US*)

License © Creative Commons BY 4.0 International license
© Dag Haug and Nianwen Xue

Joint work of Dag Haug, Emily B. Bender, Archana Bhatia, Kilian Evang, Jan Hajic, Laura Kallmeyer, Carlos Ramisch, Nianwen Xue

- In compositional frameworks MWEs are defined by non-compositionality.
- Also the approach taken by PARSEME: “Probably the most salient property of MWEs is semantic non-compositionality.”
- In PARSEME the distinction remains intuitive, because not tied to a particular mapping theory, while (non-)compositionality presupposes to a mapping from syntax to semantics (but not a specific syntactic or semantic framework).
- Even in settings where you have compositionality, there will be borderline cases like “*white wine*”, “*dry wine*”, “*red hair*”, etc.
- No clear way to distinguish a non-compositional analysis from one where *white*, *dry*, *red* has a special meaning only used in certain domains (subsecutive adjective)
- UMR annotates meaning directly, so MWEs cannot be defined by non-compositionality. Instead it is cases where several lexical items map to a single node in the semantic graph. It becomes an alignment issue between the syntactic structure and the UMR annotation.
- But arguably there are cases, where you would want to represent the idiom with several concepts, because it is internally modifiable.
- We decided to go through the categories of MWEs in PARSEME with regard to how they should (in principle) be represented semantically
- Two main questions:
 1. whether they are decomposable, i.e., which morphosyntactic components contribute to (distinguishable) meaning components;
 2. what these meaning components look like.

PARSEME has contributed considerably to MWE identification. In this context, semantic properties of MWEs have also been discussed. But a number of issues have been left open concerning the semantics of MWEs. In particular the actual semantic representations of MWEs in the context of annotating and processing MWEs have not been tackled so far.

We started from the MWE typology in the PARSEME annotation guidelines⁵ and discussed how reasonable semantic representations for the different types could be built.

⁵ https://parseme.fr/lis-lab.fr/parseme-st-guidelines/1.3/?page=030_Categories_of_VMWEs

LVC.full

In this type, the noun denotes an event and the verb contributes TAM features (in the sense of UMR). Argument structure is not modified, but “shared” between the LVC and the eventuality described by the second argument.

LVC.cause

The noun denotes an event or state, sometimes figuratively. Its first argument is not the first argument of the light verb. The noun does not provide a nominalization of the eventuality described by the entire sentence.

We discussed the following cases from the PARSEME guidelines:

- “*give a headache*” is actually at the same time LVC.cause and idiomatic MWE, at least in many cases. Question: How to annotate this? “*headache*” is an metaphor, which is embedded in a LVC.cause.
- “*grant rights*”: it is not clear whether this is a LVC. Maybe not, since the meaning of “*grant*” is relatively rich: transfer, causing somebody to have something ...
- “*provoke the destruction*”: rather not an LVC since the verb is a full verb with a specific meaning.
- “*give a bath*” might be a better example, compared to “*take a bath*”, which is not causative, but “*give*” still seems to carry some real meaning here?

It is an open question whether we consider “*take a bath*” and “*give a bath*” to be both instances of “*bath*” or do we want to say that the latter is “cause to take a bath”?

In sum, we are a little skeptical of the LVC.cause type, since in many of the examples, the verb carries too much meaning to be a light verb. The discussion was then partially continued in WG8.

VIDs (verbal idioms)

A big challenge to find an appropriate concept. Do we combine concepts, or do we make new concepts? For example for the German idiom “*ein kleines Vöglein hat mir gezwitschert*”: Is the concept “little-bird-tweeting” or “being-told-in-secret”?

A second important question is how to deal with modification of idioms? E.g. “*jump on the bandwagon*”, “*jump on the AI bandwagon*”, “*jump on the latest AI bandwagon*”. UMR creates a new concept that takes arguments like “AI”, “latest”, but this doesn’t work well if the process is truly recursive. The Düsseldorf group decomposes such idioms so that “jump” means “join”, “bandwagon” means “fad”. Other similar examples are “*pull strings*”, “*pull family strings*”, “*take the project under its federally funded wing*”, “*who made them kick their respective buckets*”.

Here are more modification examples, from Riehemann 2001 (found in the North American News Corpus):

- Meanwhile modern navigation and transport ensured that **no significant stone on the planet was left unturned**, no nation or tribe undiscovered or undocumented.
- King and Alexander, who sued each other after their bitter 1992 divorce, **buried the legal hatchet** in May.
- Russia cannot be allowed to **call NATO’s shots**.

The problem in such cases is to make it possible for the modifier to access the slot that it modifies. But it can also be problematic to construct the correct semantics when we have cases where a modifier appears relatively low in the structure of the idiom, but actually modifies the idiom as a whole (external modification):

- “leave a (very, extremely, horridly) bad taste in someone’s mouth”: Is the position of the modifier wrt the whole idiom interesting or difficult here?
- “they want to have their political cake and eat it too”

WG4 also looked at internal modification of verbal idioms.

IRVs: inherently reflexive verbs

The semantics of this group seems clear: the reflexive (by definition of inherently reflexive) does not contribute a participant.

More interesting is the case of reciprocals such as “*sich treffen*” vs. English “*meet*”, where you can have both “We met” and “We met each other”. Should all of these have the same argument structure, i.e. should there be an implicit second argument in the cases like “We met”? The consensus of the group was yes.

We also discussed examples like “*find oneself in a difficult situation*”. These are syntactically idiosyncratic, but not semantically not idiomatic.

In sum, we saw little need to deal with inherently reflexive verbs as MWEs.

Verb particle constructions

“*run over*”. In MRS, the verb is the concept and the particle is selected. In UMR, the verb and the particle are concatenated and an aspect feature is (possibly) added. In DRT, the particle has empty semantics and there is just one semantic unit.

Multi-verb constructions

“*make do*”. These are very rare, typically opaque and form one unit with just one semantic contribution which cannot be internally modified.

Inherently adpositional construction

“*rely on*”. In both UMR and DRT, the verb provides the concept and the selected adposition mediates a thematic role which is governed by the verb.

4.4 WG4: Finding Idiosyncrasy in Corpora

Nurit Melnik (*The Open University of Israel – Raanana, IL*) and Francis Bond (*Palacký University Olomouc, CZ*)

License © Creative Commons BY 4.0 International license
© Nurit Melnik and Francis Bond

Joint work of Timothy Baldwin, Archana Bhatia, Nina Böbel, Francis Bond, Mathieu Constant, Daniel Flickinger, Maria Koptjevskaja-Tamm, Peter Ljunglöf, Nurit Melnik, Alexandre Rademaker, Agata Savary, Leonie Weissweiler

Introduction

The working group was formed to address three main discussion topics:

- Discovering linguistic idiosyncrasy (Nurit Melnik)
- Identifying non-compositional MWEs in text (Francis Bond)
- NLP-based study of universals of linguistic idiosyncrasy (Agata Savary)

Over the course of six sessions, the group engaged in intense discussions which included but were not limited to these issues. This document aims to provide a comprehensive overview of our discussions and suggest potential directions for future research.

Discussion topics

The central theme of our discussions revolved around the concept of idiosyncrasy. We were fortunate to have a diverse group of participants with expertise in various fields such as NLP, computational linguistics, theoretical linguistics (particularly Head-driven Phrase Structure Grammar and Construction Grammar), typology, grammar engineering, and more. This diverse range of perspectives greatly enriched our discussions.

Our conversations were structured around three interrelated topics.

Idiosyncratic vs. regular. Although we did not want to delve into the question of what is the precise definition of idiosyncrasy the issue hovered over our discussions.

How to find idiosyncrasy. We brainstormed different methods, automatic and manual, for finding cases of idiosyncrasy within a single language as well as across different languages.

Accounting for mismatches between levels. We mostly discussed challenging cases of mismatches involving multi-word expressions (MWEs).

Idiosyncratic vs. regular

Our initial strategy was not to spend time trying to come up with an precise definition of idiosyncrasy. Instead, we began by looking for phenomena which we would intuitively identify as idiosyncratic. We realized that such phenomena generally stood in opposition to what is “regular”.

One type of idiosyncratic phenomena are syntactic structures which are not part of a regular core. These structures are often discussed in the Construction Grammar literature. For example, the English *the X-er the Y-er* construction has a very unique morphosyntactic pattern.

- (1) The more the merrier. [en]

Another example is the *do-be* construction, which is subject to various idiosyncratic constraints that are not derived from general properties of the language.

- (2) What you have to do is *(to) get ready. [en]

Some constructions have a regular syntactic structure, but they can host lexical items which are not expected to appear in them.

- (3) I sneezed the foam off my cappuccino. [en]

Regular syntactic structures may also have idiosyncratic meanings. Thus, for example, in the following exchange, the meaning of the coordination of the two identical Swedish adjectives meaning ‘happy’ is ‘not so good’.

- (4) Are you happy? [en] Happy and happy [sw]

Another type of idiosyncrasy involves exception to rules. For example, although adjectives in English precede the nouns that they modify, the adjective *enough* can only appear post-nominally.

- (5) a. That sounds good enough. [en]
 b. *That sounds enough good.

We also considered the notion of idiosyncrasy from a typological cross-linguistic perspective and asked whether there are particular domains which are more susceptible to idiosyncrasy. One phenomenon that we focused on was what is referred to as *pro*-drop, namely the ability to “drop” pronominal subjects, which is found in Japanese but not in English.

- (6) a. 着いた tuita “ ϕ arrived” [ja]
 b. I arrived. [en]

The common wisdom is that *pro*-drop depends on the richness of the morphology. However as the Japanese example indicates, this is not necessarily the case and this property is more idiosyncratic than is believed.

Other domains which were mentioned as potentially exhibiting idiosyncrasy were: temporals, kinship terms, negation, existentials, possessives, inherently reflexive verbs. This topic was later developed in the discussions of WG7, where they proposed to compile a typologically informed collection of “meta-constructions” (or domains) and the “morphosyntactic strategies” which languages employ to realize them.

How to find idiosyncrasy

Written sources. The most obvious resource for finding out about language are grammar books and language teaching materials. This type of literature often expands on the distinction between phenomena which can be accounted for by rules and phenomena which constitutes exception to these rules, i.e., idiosyncrasy. In addition to language teaching material, errors in learners’ output may also indicate idiosyncratic phenomena.

Published papers in linguistics, mainly in Construction Grammar and related frameworks, present and analyze constructions (e.g., the English *the X-er the Y-er, do-be* and *way* constructions). Naturally, these papers can be used as resources for finding idiosyncratic constructions. Moreover, it was suggested that it could also be possible to automatically parse and mine linguists’ papers for examples of such phenomena.

Albeit not written per se, an additional resource for finding idiosyncrasies are online constructicons which are developed for various languages.⁶ Some examples are:

- Brazilian Portuguese constructicon: <https://webtool.framenetbr.ufjf.br/index.php/webtool/report/cxn/main>
- Swedish constructicon: (SweCcn): <https://spraakbanken.gu.se/eng/sweccn>
- Russian constructicon:
<https://spraakbanken.gu.se/karp/#?mode=konstruktikon-rus&lang=swe&advanced=false&searchTab=special&hpp=25&extended=and%7Crus-construction%7Cequals%7C&page=1>
- English constructicon:
<http://sato.fm.senshu-u.ac.jp/frameSQL/cxn/CxNeng/cxn00/21colorTag/index.html>

⁶ This topic was further discussed in WG7 meetings.

Parsing corpora. Written sources are useful for finding *known* cases of idiosyncrasy. A bigger challenge is discovering new ones. For this purpose we discussed methods of using NLP tools to agnostically explore corpora and identify language use which cannot be accounted for by “regular” grammar rules.

Broad-coverage precision grammars as rule-based computational grammars are a useful tool for identifying cases of grammatical “rule breaking” phenomena. We discussed an experiment performed by Baldwin [1], who ran the English Resource Grammar over a sample of the British National Corpus and conducted a thorough error analysis. One example of an idiosyncratic construction that was identified by virtue of its rejection by the grammar is the *do-be* construction, e.g., “the thing we should do is buy a new car”.⁷

Recent neural/deep learning approaches for NLP do not share this characteristic; they are capable of parsing everything – from the most regular to the most idiosyncratic, as well as ungrammatical. However, there may be a way to probe the “black box” and to look at probabilities, surprisal, entropy, perplexity scores or confidence levels in order to identify instances that challenge models which are based on statistic regularity. It may be possible to use an incremental parser and look for spikes which indicate unexpected semantic or syntactic co-occurrences. Some group members expressed an interest in further exploring these methods.

Accounting for mismatches between levels

One general type of idiosyncrasy that we discussed is mismatches between levels (phonology, morphology, semantics, syntax), particularly in the domain of MWEs. Following are some more specific cases that were presented as particularly challenging for NLP.

MWEs can encode single predications. In other words, MWEs can appear in the sense hierarchy in the same way as a single word. For example, the meaning of the English MWE *look up* is similar to *phseek*, yet it is subject to idiosyncratic syntactic constraints.

- (7) a. I looked up the word. [en]
 b. I looked the word up.
 c. *I looked up it.
 d. I looked it up.

Moreover, some MWEs may look like “regular” phrases but as MWEs their parts do not exhibit “regular” syntactic behavior. For example, the relationship between the noun and adjective in the MWE *hot dog* is not intersective modification; *hot dog* is not a $\text{dog}_{n:1}$ that is $\text{hot}_{a:1}$, but a kind of sausage ($\text{hot_dog}_{n:1}$). This effects the syntax:

- (8) a. # I ate a very hot dog. [en]
 b. I ate a very hot pizza.

However, even if we think of a MWE as a single predicate, bits of it can still be accessed and modified. In some cases, internal syntactic modification becomes semantically external. For example, although *Texan* is modifying the *dust* part of the MWE in (9a), it is interpreted as modifying the entire meaning of the MWE (9b).

- (9) a. He bit the Texan dust. [en]
 b. He died in Texas.

⁷ See discussion in [2].

It should be noted that internal modification being interpreted externally is not only a feature of MWEs. Consider the following example.

- (10) a. I have an occasional drink. [en]
 b. I drink occasionally.

Syntactically, the adjective *occasional* is modifying the noun *drink*, but semantically, what is occasional is not the drink but rather the entire drinking event. Accounting for mismatches between levels was also discussed in WG1, WG3, WG9.

Future work

Following our WG discussions several of the participants joined forces with members of WG2 (*Annotation of particular kinds of constructions*) to discuss the relation between Construction Grammar and Universal Dependencies, thus forming WG7, initially named “CxG meets UD”. Some of the topics which we identified as ideas for future work were discussed in WG7 sessions.

- Create a cross-linguistic idiosyncrasycon/constructicons even the ‘rare’ constructions are often cross-linguistic
- Sense-tag corpora

Other ideas for future work were related to the challenge of automatic discovery of idiosyncrasy.

- Investigate the feasibility of detecting surprisal in automatic parsing
- Do 10-fold cross-validation and mine the errors
 (But how do we distinguish errors, creativity and idiosyncrasy?)

Finally, one issue that came up and prompted ideas for future research addressed the notion of idiosyncrasy with respect to large language models (LLMs): How do LLM-generated texts compare to natural texts in terms of the frequency and distribution of various idiosyncratic phenomena such as idioms, MWEs and *that*-less relative clauses?

References

- 1 Timothy Baldwin, John Beavers, Emily M Bender, Dan Flickinger, Ara Kim, and Stephan Oepen. 2005. Beauty and the beast: What running a broad-coverage precision grammar over the BNC taught us about the grammar—and the corpus. In *Linguistic evidence: Empirical, theoretical, and computational perspectives*, pages 49–70.
- 2 Dan Flickinger and Thomas Wasow. A corpus-driven analysis of the do-be construction. 2013. In Philip Hofmeister and Elisabeth Norcliffe, editors. *The core and the periphery: Data-driven perspectives on syntax inspired by Ivan A. Sag*, pages 35–63. Centre for the Study of Language and Information.

4.5 WG5: Methodological Issues and Community Interactions

Gosse Bouma (University of Groningen, NL) and Amir Zeldes (Georgetown University – Washington, DC, US)

License © Creative Commons BY 4.0 International license

© Gosse Bouma and Amir Zeldes

Joint work of Gosse Bouma, Amir Zeldes, Carlos Ramisch, Agata Savary, Emily Bender, Joakim Nivre, Lori Levin, Sara Stymne, Teresa Lynn

The group discussed several issues which roughly fit into the following topics:

- **UD Maintenance:** Current UD treebank maintenance is approaching a crisis, as more and more resources are neglected, and active developers are in charge of amounts of data they cannot update by themselves when guidelines are revised. The group discussed strategies to recruit and teach new UD annotators, with several recommendations emerging:
 - We must motivate newcomers to contribute, for example by:
 - * creating live public leaderboards reflecting committed contributions
 - * highlighting the status of up-to-date resources on the main UD page
 - * finding venues (workshops, special issues) to publish papers about maintenance efforts (Findings of UD?)
 - * creating a designation for UD contributors that can easily be put on CVs (“UD editor” or similar)
 - * approaching motivated potential contributors, incl. native speaker linguists, retired linguists, Master’s program students looking for projects and others
 - * organizing tutorials at venues like ESSLI or the Linguistic Institute, advertising at the Linguistic Olympiad, or possibly leveraging networks like Unidive
 - Use GitHub issues more clearly to recruit maintenance workers (“help wanted”)
 - More challenging possibilities with new developments:
 - * set up easy-to-use interfaces where UD repo contents can be easily imported for editing, so that willing contributors only need to receive a login
 - * figure out a gamified environment where multiple contributors compete for inter-annotator agreement or other score metrics
- **A Swadesh list of constructions:** A list of abstract typologically widespread constructions with examples of UD annotations in multiple languages would be useful for a variety of purposes, including didactic venues, documentation, validation and consolidation of practices for new languages joining UD (this was also outlined in WG7 and WG9). Such constructions could include “predicative possession” (x had y) or “property comparison” (x is ADJ_{er} than y), and the examples would span strategies from across languages. The group discussed several aspects of the creation of such a resource:
 - Data collection should be simplified as much as possible to lower the barrier for entry and increase likelihood of contributions by asking contributors to simply supply a Grew Match query for each construction and to assert that, e.g. the first 3 hits are correct examples (otherwise, they should specify in a separate column the match numbers of some correct hits)
 - Problems raised included the likelihood that examples would omit interesting variants in each language which exceed 3 basic examples, or cases about which the list simply does not ask (Lori pointed out that for some questionnaires, e.g. the Lingua checklist used by Comrie and Smith, possibly thousands of permutations of morphological categories would need to be considered)

- The group converged on the idea that asking for maximal diversity in the example types was desirable, with the understanding that more variants will inevitably be missed, but we should still focus on getting some, rather than no information
- Strategies to select the constructions were also discussed:
 - * A grammar description framework, such as the Grammar Matrix (<https://matrix.ling.washington.edu/>) could be used to create an outline of a language type, and each distinct type implicates distinct constructions of interest
 - * The UD Cairo corpus of 20 example sentences potentially contains a good list of candidate constructions
 - * The numbered list of constructions in Croft’s book can be used as a starting point as well
- **NLP for Typology:** There is a growing interest in using UD treebanks for typology (witness work by Levshina [2], token-based typology, and others). This raised various discussion points:
 - UD treebanks are often limited in size, and may not be comparable in genre and register across languages. Can we use NLP (i.e. automatic annotation or methods for selecting and annotating comparable fragments across languages)?
 - UD annotation was originally not (or not exclusively) designed with this application in mind, and it misses some dimensions that could be very valuable for typology. Treebanks could be made more valuable for typologists by adding construction level annotation (ie explicit annotation of clause types such as questions or passives, even if this is sometimes implicitly encoded) and/or annotation beyond syntax (semantic dimensions, information packaging)
 - Using UD for cross-lingual comparison does presuppose that we know when phenomena are comparable across languages (see Haspelmath’s discussion of comparative concepts [4]), but this may not always be the case for decisions made in UD annotation.
- **Unifying UD and PARSEME:** This topic is discussed in some detail in Savary [3], PARSEME meets Universal Dependencies. Even though PARSEME and UD are two orthogonal annotation layers, they can be merged, e.g. by adding a ParseM column to UD CONLL-U format. Discussion points:
 - The notion “MWE” is used somewhat sloppy in UD to group compound, fixed, and flat relations, where compound definitely should not be under the rubric MWE, and fixed and flat may be better represented as “head-less” rather than MWE.
 - PARSEME covers some MWE types that do not fall under one of the UD relations compound, fixed, or flat.
 - In the future, PARSEME aims to include nominal constructions as well. This could be coordinated with proposals in WG2 for the analysis of mischievous nominal constructions [1].

References

- 1 Nathan Schneider and Amir Zeldes. 2021. Mischievous nominal constructions in Universal Dependencies. In *Proceedings of the Fifth Workshop on Universal Dependencies (UDW, SyntaxFest 2021)*
- 2 Natalia Levshina. 2019. Token-based typology and word order entropy: A study based on Universal Dependencies. *Linguistic Typology*, 23, pages 533–572
- 3 Agata Savary, Sara Stymne, Verginica Barbu Mititelu, Nathan Schneider, Carlos Ramisch, and Joakim Nivre. 2023. PARSEME Meets Universal Dependencies: Getting on the Same Page in Representing Multiword Expressions. *Northern European Journal of Language Technology*, 9, 1
- 4 Martin Haspelmath. 2010. Comparative concepts and descriptive categories in crosslinguistic studies. *Language*, v.86, pages 663–687

4.6 WG6: Above and Below Word Level

David Yarowsky (Johns Hopkins University – Baltimore, US) and Omer Goldman (Bar-Ilan University – Ramat Gan, IL)

License © Creative Commons BY 4.0 International license

© David Yarowsky and Omer Goldman

Joint work of Goldman, Omer; Yarowsky, David; Zeman, Daniel; Stymne, Sara; Vylomova, Ekaterina; Kahane, Sylvain

WG6 was a direct continuation of WG1 from the 1st half of the seminar and was tasked with proposing enrichment for Universal Dependencies (UD) above and below the word level. While WG1 discussed both addition of phrase level features above the word level and morphological segmentation below it, WG6 focused solely on the morphological segmentation, hashing out the different possible implementations and discussing the concrete steps needed to be taken to achieve it. In essence, the participants proposed a two-step process of alignment between UniMorph and UD that will allow usage of UniMorph data in UD trees.

UD-UniMorph alignment

In the current state of affairs, both UD and UniMorph contain morphological data but they tag it according to different although generally compatible annotation schemas. Therefore we suggest an effort to align these two resources. This effort will probably not include any changes needed to be done in UD.

As a first step we propose to map the differences between UD and UniMorph, both in terms of the features used and the structure of the features. The mapping will be done as part of an effort to extract morphological data from UD to UniMorph. Once it'll be done, we could produce a list of the changes we believe are required and approve through the relevant committees in the management of both datasets.

Having both projects using the same annotation schema for morphological data will make it the cooperation between the project organizers smoother.

Morphological Segmentation in UD

Currently, morphological information in UD is confined to the “feats” column where features are attributed to the entire node, and it makes languages seem different depending on the frequency of white space usage. Decomposing words into morphemes and marking the relations between morphemes with dependency arcs as done between words, will equate the structure of different languages and will make it easier to typologically compare languages.

The group proposed segmenting words into morphemes, mostly using the segmentation files from UniMorph, currently existing for about a dozen of languages. The lemma of each morpheme will be its canonical form and the features be associated only with the relevant morpheme rather than with the entire word.

There were 2 main implementation options discussed: one where the content of each “morphological node” is a morpheme, and one where it is a truncated word. For example, a word like “*industrialization*” will be decomposed into “*industry*”, “*-al*”, “*-ize*” and “*-ation*” according to the first option, and to “*industry*”, “*industrial*”, “*industrialize*” and “*industrialization*” according to the second. The benefits of the latter is that it does not require the annotators to decide on the “canonical form” and whether it even exists and it is more closely aligned with the UniMorph derivational morphology data, but on the other hand this option is less intuitive and diverges to some extent from the structure of UD for words.

This discussion should be revisited after the completion of the first phase.

4.7 WG7: UniCoDeX (Universal Construction Dependency Xrgrammar)

Peter Ljunglöf (University of Gothenburg, SE) and Lori Levin (Carnegie Mellon University – Pittsburgh, US)

License © Creative Commons BY 4.0 International license

© Peter Ljunglöf and Lori Levin

Joint work of Baldwin, Timothy; Bhatia, Archana; Böbel, Nina; Bond, Francis; Bucker, Jörg; Constant, Mathieu; Flickinger, Daniel; Kahane, Sylvain; Levin, Lori; Ljunglöf, Peter; Lynn, Teresa; Melnik, Nurit; Nivre, Joakim; Rademaker, Alexandre; Sailer, Manfred; Savary, Agata; Schneider, Nathan; Weissweiler, Leonie; Zeldes, Amir

Introduction

This is a summary of the discussions that took place in Working Group 7 during the Dagstuhl Seminar 23191 *Universals of Linguistic Idiosyncrasy in Multilingual Computational Linguistics*.

WG7 was formed from the initial working groups 2 (*Annotation of particular kinds of constructions*) and 4 (*Finding idiosyncrasy in corpora*), who independently of each other realised that they wanted a group that was focused on the relation between Construction Grammar and Universal Dependencies. So for the second half of the seminar we formed WG7 with the initial name “CxG meets UD”, but we changed the name of the group to the more catchy UniCoDeX (*Universal Construction Dependency Xrgrammar*).

The group met for four sessions and had intense discussions which resulted in the formation of three interconnected tasks. This document tries to reflect the results of our discussions and the way forward.

Overview of discussion topics

The overall question of WG7 was how to connect Construction Grammar (CxG) with Universal Dependencies (UD) in a way that both Construction Grammar projects and Computational Linguistics projects can benefit from. Our discussions were organized around three interrelated topics:

A. Documenting morphosyntactic diversity in UD.⁸ How can UD annotations and guidelines be improved to better reflect typological differences between languages (also discussed in WG1)?

B. Standard for a construction annotation layer in UD. How could construction annotations be added to augment UD treebanks? What annotation standard would be needed (also discussed in WG5 and WG9)?

C. Searching for constructions in UD treebanks. How could UD treebanks be queried for interesting examples of a given construction (the topic was also discussed in WG4, WG5, WG9)?

Topic A. Documenting morphosyntactic diversity in UD

The general goal of this topic is to verify the typological coverage of UD, increase its consistency, and advise users on how to analyze constructions in their languages. There are two important concepts: *meta-constructions* framed in comparative terms based on [1]; and

⁸ Previous name: “Typologically valid UD annotation guidelines”

morphosyntactic strategies that specific languages may employ. The group working on this topic will create a collection of annotation examples and guidelines for different languages and meta-constructions:⁹

- This collection corresponds to the “Swadesh list” for morphosyntax from WG2 – we dub it the *Nivre list* since it was proposed by Joakim Nivre. Using this checklist, treebankers can determine which morphosyntactic strategies are used in their languages for each meta-construction and how to annotate them.
- One component of this collection is a spreadsheet with languages in the rows and meta-constructions in the columns, where the cells contain links to queries yielding annotation examples and notes indicating which morphosyntactic strategy is illustrated in each example.
- Another component is a table of meta-constructions and morpho-syntactic strategies, with examples from different languages and language families.

This collection will be used to promote consistent and typologically informed coverage of morphosyntactic strategies in UD and update the general UD annotation guidelines from a typological perspective, including morphosyntactic strategies as defined by [1].

- This will help UD to find areas where we can improve current explanations and analyses to be more typologically oriented.
- It will also be of help when extending the guidelines to improve the coverage of explanations.
- It will ensure that it is possible to represent all (or at least most) known typological diversity in UD.
- It will facilitate using UD for research in language typology.

Example(s). Object predication is a cross-linguistic meta-construction which uses different strategies in different languages. In this meta-construction a semantic object is information-packaged as a predicate. (This is conventionally called a predicate nominal.)

- English: verb copula strategy (“Dani is a student”)
- Russian: zero strategy (“Dani student”)
- Hebrew: pronoun copula strategy (“Dani hu student”)
- Classical Nahuatl: inflect the noun as a verb: (“ni-ticitl”, 1sg-doctor)

We want annotation guidelines for each language/strategy so that an annotator will not necessarily go to an English treebank by default. The table of morphosyntactic constructions and strategies should be able to guide a treebanker to the right strategy. The spreadsheet should guide the treebanker to examples from languages that use that strategy, which illustrate how to make dependency trees for that strategy.

Future work. People who are interested in working on this topic after Dagstuhl:

- Lori and Joakim (group leaders)
- Alexandre, Amir, Archana, Jörg, Leonie, Nathan, Nurit, Sylvain

As a concrete first step the group agreed to do the following in the near future:

- add at least 10 languages and 10 meta-constructions to the Nivre list.
 - We will start with basic meta-constructions like clausal possession, comparison, and argument alignment (accusative or ergative).
 - We will write guidelines for each meta-construction and each strategy.
- when this has been finished, there will be an internal review within the group to decide about future steps.

⁹ [1] calls them “constructions”, but we refer to these as meta-constructions to differentiate them from language-specific constructions in the Construction Grammar sense.

Topic B. Standard for a construction annotation layer in UD

The general goal of this topic is to develop recommendations for how to annotate UD treebanks with constructions. They should be useful for several different use cases, workflows and granularities, such as:

- Use cases: we could be building a new constructicon from the ground up and want to come up with good definitions of constructions, or we might already have an existing constructicon which we want to use for annotation or extend with new constructions
- Different annotation granularities: from the coarsest level (to just annotate the head of a construction with its name), to the most fine-grained (to also annotate all the construction elements with their names and spans within the sentence)
- Different workflows: people might want to annotate one construction at a time, or several at once – or they might want to use an iterative approach where they start with coarse-level annotation and then refine them

Future work. People who are interested in working with this topic after Dagstuhl:

- Leonie (group leader)
- Alexandre, Amir, Archana, Francis, Lori, Manfred, Nathan, Nina, Nurit, Peter, Sylvain

As concrete first steps the group agreed to annotate a limited family of constructions in different languages, with the hope of writing a joint paper during autumn.¹⁰ Some initial ideas of constructions that could be interesting to annotate were:

- age constructions, rates (mph etc), comparatives, resultatives, ...
- idiosyncratic, lexicalised constructions, such as X-and-X (Swedish), N-über-N (German), N-after-N (English)
- cross-linguistically common constructions such as types of conditionals, possession, comparison etc. (i.e. exponents of meta-constructions, see above)

The group will continue discussing topics such as:

- naming convention for constructions
- integration with existing annotation tools
- annotating/markings candidates that have been checked and are not a certain construction
- which token in the UD tree should be annotated with the construction?
 - the natural choice is to annotate the token that is highest in the UD tree – but it is unclear what to do if the construction covers disconnected parts of the UD tree

The group agreed to postpone more complicated questions, such as:

- how to handle cross-sentential constructions
- how to handle nesting and composition of constructions
- how to handle constructions on different levels of granularity (more specific vs. more general constructions)

Topic C. Searching for constructions in UD treebanks

The general goal of this topic is how to formulate search queries that can locate interesting examples of a given construction. This is very closely related to the previous two topics, as they all depend on being able to search for constructions in treebanks.

¹⁰Possibly targeting LREC-COLING 2024, with submission deadline October, or ICCG 2024 with deadline in spring.

The group discussed some issues that arise when it comes to formulating search queries, such as:

- we want guidelines that help people with writing and refining queries
- we want (semi-)automatic techniques for extracting relevant search queries from an existing Constructicon entry
- precision/recall tradeoff: it is probably more important to have a good recall than good precision, but it is usually easier to improve the precision by modifying a query
- possible strategies to increase the recall can be to use approximate tree matching or to loosen some constraints in the query

Future work. People who are interested in working with this topic after Dagstuhl: the same as for topic B, with Leonie as group leader.

In the beginning this topic will be closely related with topic B. To be able to annotate the treebanks we will have to formulate search queries that can find potential candidates. While doing this iterative process for a diverse set of constructions in different languages we hope to come up with more general guidelines on how to write construction queries.

Related work/links

The following are the existing Constructicons that we are aware of:

- English: Berkeley FrameNet Constructicon: <http://sato.fm.senshu-u.ac.jp/frameSQL/cxn/CxNeng/cxn00/21colorTag/>
- English: Birmingham English Constructicon: <https://englishconstructicon.bham.ac.uk/>
- English: CASA (FAU Erlangen-Nürnberg): <https://constructicon.de/>
- German: FrameNet-Konstruktikon (HHU Düsseldorf): <http://framenet-constructicon.hhu.de/>
- Swedish: Svenskt konstruktikon (Univ. of Gothenburg): <https://spraakbanken.gu.se/karp/#?mode=konstruktikon>
- Brazilian Portuguese: FrameNet Brasil (FU Juiz de Fora): <https://www2.ufjf.br/framenetbr-en/>
- Japanese: Japanese FrameNet (Keio University): <https://jfn.st.hc.keio.ac.jp/>
- Russian: Russian Constructicon (UiT Arctic University of Norway): <https://constructicon.github.io/russian/>
- Most of the different constructions were presented at the Constructicon Alignment Workshop (CAW, December 2022), and video recordings are available here: <https://www.globalframenet.org/caw2022>

Croft’s “Morphosyntax: Constructions of the World’s Languages” [1] contains a glossary of different *comparative concepts* (meta-constructions, strategies, information packaging, etc.), and this glossary is available online here:

- Interactive interface: <https://spraakbanken.github.io/ComparativeConcepts/>
- GitHub repo: <https://github.com/spraakbanken/ComparativeConcepts>

Finally, here is a list of different search engines for corpora, tools and treebanks, that can be used to find constructions:

- Grew-match: <https://match.grew.fr/>
- SPIKE (query-by-example): <https://spike.apps.allenai.org/>
- DepEdit: <https://gucorpling.org/depedit/>
- UDAPI: <https://udapi.github.io>

- Korap (IDS-Mannheim: <https://korap.ids-mannheim.de/> and <https://github.com/KorAP/>)
- Corpus workbench (CWB, useful for larger corpora) – several sites use CWB, such as:
 - Språkbanken Korp (Univ. of Gothenburg): <https://spraakbanken.gu.se/korp/>
 - CQPWeb (Lancaster Univ.): <https://cqpweb.lancs.ac.uk/>
 - CWB source code can be found here: <https://cwb.sourceforge.io/>

Proposed standard for construction annotation in UD

We propose a new layer for selectively annotating constructions on top of UD trees. This is intended for constructions (in the sense of Construction Grammar) whose form and meaning/function is not already captured well by the UD tree. Construction instances receive a type name (possibly from a construction resource) and may contain relations to construction elements. The elements of the construction are not constrained by the UD tree: e.g., a construction element may cut across multiple UD subtrees. For now, we envision that they would be marked in the MISC column of .conllu files, though in principle they could be moved to a separate extension column.

The annotation layer does not have the goal of directly indicating the elements of form or meaning that are characteristic of or required by the construction, beyond indicating the construction evoker and spans of construction elements. Aspects of the UD analysis (tags, deprels, morphological features) that are characteristic of a construction's form should be described as such in a type-level construction entry. The precise contents of such an entry are not part of this proposal, but constructions incorporating UD information in some way already exist (e.g., the Russian Constructicon).

Full

Showing three overlapping constructions for completeness:

```

1 Sam    CxnEltOf=5:predicative-age.Individual,5:property-predication.Subj
2 is     CxnEltOf=property-predication.Cop
3 three  CxnEltOf=4:num-mod.Quantity,5:predicative-age.Value
4 years  Cxn=num-mod|CxnEltOf=4:num-mod.Counted,5:predicative-age.Units
5 old    Cxn=predicative-age,property-predication|CxnEltOf=5:property-predication.Pred

```

This effectively encodes construction-element relationships as dependencies (*offset:relation* notation echoes DEPS column), which would allow for straightforward graph querying. A common query might be to list the UD deprels associated with a construction element.

Note that i) a word may evoke multiple constructions, ii) a word may be both the evoker and an element of an evoked construction, iii) a word may participate in multiple elements of the same evoked construction.

Comma-separated lists should be sorted primarily by head node (where present), secondarily by construction name, thirdly by construction element name.

Full-consolidated

```

1 Sam    _
2 is     _
3 three  _
4 years  Cxn=num-mod(3:Quantity,4:Counted)
5 old    Cxn=predicative-age(1:Individual,3:Value,4:Units),\
        property-predication(1:Subj,2:Cop,3-5:Pred)

```

This is equivalent to the Full representation but consolidates all parts of an evoked construction on one line. It might be suitable for human annotation, to be automatically expanded to the Full representation with a script.

Comma-separated construction elements should be listed in node sort order. Constructions should be sorted alphabetically by name.

Simple

A partial representation may be useful in certain stages of an annotation workflow, e.g. before the full description of the construction is known, or before applying semiautomatic methods to identify construction elements.

The Simple notation includes the name of a construction, omitting any construction elements. A span may optionally be included for rendering purposes, but this span does not necessarily have any theoretical status.

```
1 Sam _
2 is _
3 three _
4 years Cxn=3-4:num-mod
5 old Cxn=1-5:predicative-age,property-predication
```

Exclusions

When manually reviewing forms that are candidate matches of a construction, it may be helpful to indicate that one of them is a non-match (a false positive). This can be done with the ExcludeCxn feature:

```
1 Sam _
2 is _
3 three _
4 years Cxn=3-4:num-mod
5 old Cxn=1-5:predicative-age,property-predication|ExcludeCxn=object-predication
```

Though we suggest the name ExcludeCxn in this standard, it should be regarded as a tool for development. Ideally, a corpus will be systematically reviewed for candidates of a construction, and excluded candidates discarded in the final version of the data.

Linking to a constructicon

If a constructicon resource exists, it should be declared in a metadata line in the file, and names of constructions from the resource should be prefixed with a namespace.

TBD issues

- Where are spans vs. heads used? Is a construction-evoking element allowed to be a span? Allow discontinuous spans (and change existing commas to semicolons)?
- A status field to indicate auto rather than gold matches?
- Allow question marks to indicate uncertainty during development?

Example annotations

(1) “The more you post the more money you make”. Let’s assume that the first comparative word (“more”) is the head. Then that word will be annotated like this in the simple format (the span 1–8 is optional):


```

1 the _
2 more Cxn=1-8:comparative-correlative
3 money _
...

```

And like this in the Full notation:

```

1 the _
2 more Cxn=1-8:comparative-correlative(1-4:Condition,6-10:Result,\
    2:ConditionDegree,7:ResultDegree)
3 money _
...

```

(2) “Sam is so glad that you are here that he baked a cake”. Advanced example (consolidated notation), showing two candidate matches of the same construction type on the same construction evoker, one of which is correct and one of which is incorrect (indicated by an excluded span):

```

...
3 so _
4 glad Cxn=causal-excess(1:Predicand,3:Degree,3-8:Cause,9-13:Result)\
    |ExcludeCxn=causal-excess(5-8:Result)
5 that _
...

```

Or in the simple form:

```

...
3 so _
4 glad Cxn=1-13:causal-excess|ExcludeCxn=1-8:causal-excess
5 that _
...

```

References

- 1 William Croft. 2022. *Morphosyntax: Constructions of the World’s Languages*. Cambridge: Cambridge University Press.

4.8 WG8: To Semantics and Beyond

Archna Bhatia (*Florida IHMC – Ocala, US*) and Kilian Evang (*Universität Düsseldorf, DE*)

License © Creative Commons BY 4.0 International license
© Archna Bhatia and Kilian Evang

Joint work of Timothy Baldwin, Emily B. Bender, Archna Bhatia, Francis Bond, Kilian Evang, Jan Hajič, Laura Kallmeyer, Carlos Ramisch, Nianwen Xue, Amir Zeldes

Introduction

Somewhat continuing the discussion started in WG3, Working Group 8 (WG8) *To Semantics and Beyond* at the Dagstuhl Seminar 23191 on *Universals of Linguistic Idiosyncrasy in Multilingual Computational Linguistics* identified topics related to semantics that have presented to be problematic from various angles, such as semantic representations, inconsistencies or differences in representations/handling of notions across frameworks, cross-lingual handling of constructions or semantic notions, and coverage, e.g., both for arguments and modifiers. Specifically, the following four topics were identified, and we focused on the first two topics during the WG8 discussion sessions:

1. Serial verb constructions and light verb constructions from a cross-lingual perspective in terms of representation of their semantics, validity of statements such as “Predicates describe one event”, representations across different frameworks, leveraging other annotations or resources (e.g., the notion of scope tree in MRS), The topic was also discussed in WG3.
2. Inventory of semantic relations to describe states, processes and events cross-linguistically and for both arguments and adjuncts (also discussed in WG3 and WG5)
3. Inconsistent handling of entities and coreference across resources
4. Aspects of lexical semantics such as linking with wordnets, internal semantic structure, relation between internal semantic structure and compositionality, representation in existing resources/corpora

In the next two sections, we summarize the discussions corresponding to the two focused topics.¹¹

The Semantics of Serial Verb Constructions and Light Verb Constructions

Verbal constructions can present interesting phenomena involving a wide range of semantic considerations, e.g., in terms of determining whether a single event is involved or multiple events are involved, while at the same time each of these constructions themselves may also present a broad continuum to make it harder to demarcate the boundaries of the construction. This can present issues for describing such constructions or developing proper theoretical accounts of them. We discussed two verbal constructions, the serial verb constructions (SVCs) and light verb constructions (LVCs), that illustrate this issue.

SVCs, e.g., *persuade X to take a hike*, or *try to run*, are argued to involve control structures where one of the verbs is considered to be incorporated into the other verb. In such constructions, multiple independent lexical verbs are combined to indicate a single event (or two sub-events of a single event connected temporally indicating either simultaneous or consecutive temporality). LVCs, e.g., *take a bath*, *give a bath*, or *give a speech*, involve a verb which does not indicate the event itself but provides some aspectual or causal information about the event and the semantic core of the event is expressed by the nominal element.¹²

These constructions raise important questions about how one can determine semantic uniqueness of an event, whether the event is expressed by a verb, and also what an event is. We discussed tests for semantic uniqueness of an event and noted that neither of these tests were sufficient by themselves in indicating whether a single event was expressed or multiple events were expressed, but they might indicate tendencies that multiple tests together could help confirm to determine if a single event was involved. These included tests such as modification, argument sharing, and coordination.

In regards to modification, we identified examples in Thai SVCs indicating that negation (as modification) could be used but either in only one place (i.e., with one of the verbs) or it may mean the same irrespective of the position of negation. Cross-lingual examples such as these and the English, *I persuaded Francis not to take a hike* are interesting. Note that the use of negation in this position could be used to indicate persuading Francis to not take a particular hike or any hike, but in this position, it does not negate the act

¹¹ The document with more detailed notes for the discussions can be found at: <https://docs.google.com/document/d/13-J0kaCKAshDShRFE9NN81Yysc7jmLMlswmrfneRRo>

¹² In some languages, e.g., in Hindi, as the Hindi PARSEME annotated data indicate, the event may also be determined based on a combination of other elements such as an adjective with a verb.

of persuading Francis to take a particular/any hike. Quantification (as a modification strategy) can also be useful. In terms of semantic representation, “one event” may be represented semantically through, for example, the same position in the (Minimal Recursion Semantics/MRS representation) quantifier scope tree which can also entail a test of semantic uniqueness. However, quantifiers based test also does not always work. It is not always clear how to create the quantifier examples.

The argument sharing test is also not diagnostic. A vast amount of literature involving language specific approaches to SVCs discusses argument sharing as a criterion for a single event. While it may be a necessary criterion for some languages, it is not sufficient in determining semantic uniqueness of event. Also, while English SVC *I tried to run* involves subject sharing, a Thai construction equivalent to “*pound X flat*”, does not involve “subject sharing”¹³ but such resultative constructions in Thai are also considered to be SVCs. To take another example, the argument sharing test does not help resolve whether the Japanese construction equivalent to “*jump-rise the stairs*” involves one event with manner information or two events. In such verb-verb constructions, one verb (“jump”) is a hyponym of the other (“rise”).

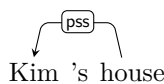
Coordination can also be considered as a test for semantic uniqueness but is again found to not be a strong diagnostic. For example, in the case of LVCs, since the N+V combination expresses an event, one can expect coordination of nouns in such constructions to not be possible. But PARSEME corpus shows there are quite a few of such cases. For example, it is possible to say *He took a bath and a shower*. Coordination also seems possible where the nominal part of the LVC is modified and is coordinated with a non-LVC noun, e.g., *the bath that I took and the bathtub were wonderful*.

The group found such examples with LVCs involving these diagnostics, and their flexibility when testing their boundaries particularly interesting and ended up focusing a large part of the WG8 sessions on discussing LVCs further. We are continuing our discussions involving LVCs, their characteristics and behavior across languages to arrive at a more general and typologically informed semantic representation of these constructions beyond Dagstuhl with many participants from WG8 and other groups (the current participants’ list continuing these discussions includes: Emily M. Bender, Archana Bhatia, Kilian Evang, Dan Flickinger, Jan Hajič, Dag Haug, Laura Kallmeyer, Carlos Ramisch, Nianwen Xue). We plan to continue these discussions to also include other MWEs to develop their semantic representations while taking into account the observed cross-lingual patterns as well as the idiosyncratic behaviors they demonstrate.

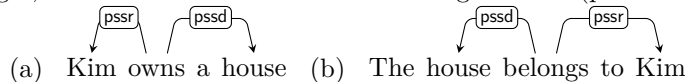
Towards a Unified Inventory of Semantic Dependency Labels

There is currently no commonly accepted inventory of semantic roles and relations that can be applied across languages, domains – something like Universal Dependencies, with similarly low barriers to use in applications, but for semantic roles and relations. Kilian Evang started the discussion by formulating some desiderata for such a scheme: it should be usable without reference to a lexicon, it should cover at minimum all dependencies between content words (thus, both arguments and modifiers) with a unified vocabulary, and handle states and events in a unified way. Semantic relations should not be too fine-grained, He illustrated what such a scheme might look like using a possession/control relation, labeled *pss*. This label could be used directly for modifier relations:

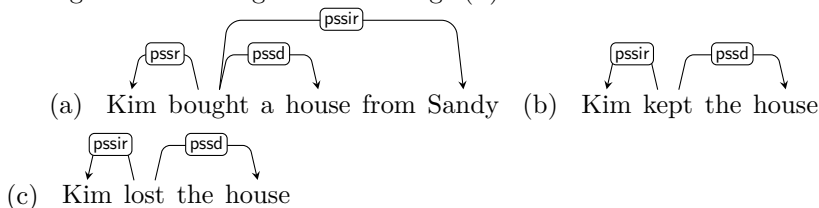
¹³ although an argument is shared



For arguments of predicates denoting states, the same relation is split over two dependency edges, dubbed the “domain” and the “range” of the (possession) relation:



For arguments of predicates denoting events, there is the additional possibility of having an argument denoting an *initial* range (ir):



Differences to existing schemes were discussed: MRS first assigns semantically unspecified argument labels (ARG0 = event, then ARG1, 2, 3 in order of decreasing obliqueness). Separate, language-specific lexicons à la FrameNet assign specific roles to these argument labels, per predicate (e.g., *own*, *belong* may be in the same frame, but for *belong* ARG2 is the owner, and for *own*, ARG1). In PropBank, all arguments are verb-specific. SynSemClass is an event type ontology under development that defines coarser, but still hierarchically ordered classes of events (and states), with e.g. the “ownership” class containing verbs like *own*, *possess*, *hold*. “Buying” would be a separate class, e.g., it is not clear that the seller is also the initial owner in all cases. This links to the general question of how much annotators should be allowed to infer that is not strictly entailed by the predicate, e.g. *I visited my mother at 4 o'clock yesterday* – should the annotator include the information that it was in the afternoon? Verbs in the Prague Dependency Treebank are linked to FrameNet, VerbNet, OntoNotes, PropBank, English WordNet, and SynSemClass. In the TRIPS ontology, there are events of change, events of state, and within those further classes.

Potential issues and next steps with the proposed schemes were discussed. Compared to UD, additional semantic edges need to be annotated compared to UD, e.g., in control constructions. It was therefore suggested that annotation start atop Enhanced UD, even if automatically predicted and imperfect. Similarly, existing PARSEME data could be used to automatically pre-annotate semantic edges in VMWEs, e.g. *kick* → *bucket* with special labels. The next steps should focus on annotating a bunch of data to evaluate the proposal, and also on lexical mapping to other resources like SynSemClass, PropBank, Universal Proposition Banks, VerbAtlas, VerbNet, FrameNet, SNACS, UMR to study synergies and enable comparisons.

4.9 WG9: Fostering Corpus-based Typology [“Big Hairy Problems”]

Laura Kallmeyer (*Universität Düsseldorf, DE*) and Christopher Manning (*Stanford University, US*)

License © Creative Commons BY 4.0 International license

© Laura Kallmeyer and Christopher Manning

Joint work of Laura Kallmeyer, Christopher Manning, Joakim Nivre, Reut Tsarfaty, Gosse Bouma, Agata Savary, Dag Haug

We started with two questions:

1. How can we make corpus-based typology easier, in particular with respect to investigating interesting phenomena beyond word order (also discussed in WG5)?
2. How can we bring some of the constituency information typologists might want to encode into an easy-to-use annotation format that connects UD and constituency (also discussed in WG4, WG5, WG7)?

First, we looked a little into RRG [Role & Reference Grammar] and its distinctions of nucleus, core, and clause for different levels of tightness of binding of elements and whether that can and should be captured in UD. We could represent this by adding something to the dependency label.

Then we dealt with polysynthetic languages with noun incorporation.

Question discussed after the break: Do we want to have constituency representations in UD or can we make do with phrase-level features?

Joakim says things to consider:

- Phrase-level features: We may need them.
- Do we want segmentation in the basic representation for incorporating or polysynthetic languages? Do we need complex feature structures, i.e., where features can take a content word as a feature value.
- One thing maybe missing is when there are multiple realizations of the same argument: *moi je pense que*: There are sort of 3 realizations of 1SG.

Chris:

- UG really has no treatment of complex predicates/light verbs/serial verbs/monoclausality.

Reut:

- Using a featural analysis we could represent something like a periphrastic passive the same as a single word morphological passive.

Gosse:

- Do we need more bracketing, starting to head in the direction of constituency?

Joakim:

- How can we do this with a manageable amount of complexity?

Agata:

- Field linguist thinks UD isn't linguistic enough for him. How? Is SUD better?

Chris:

- We don't need to provide a linguistic theory, if we're just providing enough that people can search for examples of interest, then we've won.

Reut:

- What's more important: Ease of annotation or ease of searching?

Note that the flat relation does not imply a head. We have to educate people that there are links in UD because everything is made into a tree, but not all of the arrows are head dependent relations.

Clause-level phrasal features are a good way to capture interesting types of constructions and simultaneously capture things that are above the level of a single token.

We compiled a list of phrasal features that one might want to have:

- Clause level features:
- Clause types:
 - Interrogative
 - * Presuppose: yes, no
 - * Wh-question: yes, no
 - * Tag-question: yes, no
 - Declarative
 - Imperative
- Voice/Valency changing constructions:
 - Active
 - Passive
 - Antipassive
 - Middle voice
 - Causative
 - Applicative
- Polarity:
 - Positive
 - Negative
- Information structure:
 - Cleft?
 - Pseudocleft?
 - Extraposition?
- Tense/Aspect/Mood
- Evidentiality
- Long-distance dependency?
- Nominal phrase level features:
- Nominal types:
 - Definiteness
 - Case
 - Construct state
 - Agreement
 - Deverbal noun

We then came back to the Nahuatl example. Our proposal: In the enhanced dependencies, have abstract nodes for those arguments that follow from the morphologically encoded information. Besides that, encode all other properties in the features and project these also to the clause level (on the same node).

Dag:

- We need to add in the edeps nodes for subj, obj shown by agreement (but pro-drop) or else we can't show control relationships that include them in the enhanced dependencies. UD is not available in tools used in fieldwork.

Is UD a theoretical framework or a pre-theoretical annotation framework?

- But it does make some theoretical choices like being content-word head
We should test this with some concrete constructions!
- Serial verb constructions?

Friday: Phrase-level features

- Work more on the feature list: There seems to be a need for phrase-level features and we have a first cut at a list of them.
- Put a survey together on how to change things: Need to get more buy-in from a big community and understand more about more language families.
- Easiest place to put them immediately is in the FEATS column and just have new names.
- We are wanting to capture periphrastic tense and aspect, not just morphological form of individual verbs
- We also have a mechanism of layered features for agreement with multiple things: Hebrew: **אהבה** Gender[subj]=neut, number[subj]=sg, person[subj]=1, gender[obj]=fem, number[obj]=sg, person[obj]=3rd

Let's extend this to phrase level features: Definite[phrase]=Yes

- If we used something like Conx[Phrasal]=Periphrastic or Conx[Phrasal]=Light then we could capture this.

UD Governance

Can we have open zoom discussions of changes not just dictates in emails? [“This is an information session rather than an active discussion” – manage it.] There are going to be UD and PARSEME introductions in UniDive but they're not really community discussions. Can we go back to having real UD workshops where things are discussed and worked out? Perhaps together with the next combined Coling/LREC. For different language groupings, whether language families or subgroups such as user-generated content or ancient languages, can we explicitly have more organization and subcommunity organizers? Would that help? We decided to break early so that we could get our coffee ahead of time and be on time to the final session.

5 Open Problems

5.1 Semantic Parsing and Sense Tagging the Princeton WordNet Gloss Corpus

Alexandre Rademaker (IBM Research – Sao Paulo, BR), Francis Bond (Palacký University Olomouc, CZ), and Daniel Flickinger (North Newton, US)

License © Creative Commons BY 4.0 International license

© Alexandre Rademaker, Francis Bond, and Daniel Flickinger

Main reference Alexandre Rademaker, Abhishek Basu, Rajkiran Veluri: “Semantic Parsing and Sense Tagging the Princeton WordNet Gloss Corpus”, in Proc. of the Global Wordnet Conference, 2023.

In 2008, the Princeton team released the last version of the “Princeton Annotated Gloss Corpus”. In this corpus, the word forms from the definitions and examples (glosses) of Princeton WordNet are manually linked to the context-appropriate sense in WordNet. However, the annotation was incomplete, and the dataset was never officially released as part of WordNet 3.0, remaining as one of the standoff files available for download. Eleven years later, in 2019, one of the authors of this abstract restarted the project aiming to complete the sense annotation of the approximately 200 thousand word forms not yet annotated. Intending to provide an extra level of consistency in the sense annotation and a deep semantic representation of the definitions and examples promoting WordNet from a lexical resource to a lightweight ontology, we now employ the English Resource Grammar (ERG), a broad-coverage HPSG grammar of English to parse the sentences and project the sense annotations from the surface words to the ERG predicates.

The disambiguation of words in the glosses can also improve WordNet and provide completeness and consistency. For instance, the initial versions of WordNet do not contain relations that indicate how words like “racquet”, “ball”, and “net”, and the concepts behind them, are part of another concept that can be expressed by “court game” [12]. In WordNet 3.0 the “domain relations” between synsets were introduced to alleviate this so-called “tennis problem” of WordNet [21], but the disambiguated gloss of the synset *tennis, lawn_tennis* would already enrich the connections among the concepts. Another desired property is that all words used in the definitions are defined in this same resource. Hopefully, this completeness could also help us ensure quality in our long-term endeavor during the expansion of WordNet to highly technical domains. Once more concepts are added or redefined, the glosses would be refined and disambiguated, forcing us to use the newly added senses in a productive cycle of editing, testing, and correcting.

Regarding the multiword expressions such as “military formation”, “geological formation”, “reticular formation”, and “reaction formation”. The expression “military formation” stands out in many glosses. The expression exists as a MWE but a similar expression, “naval formation” does not, with both appearing in the gloss “the side of military or naval formation”. We discussed whether “naval formation” should be considered a MWE or whether “military formation” should not be considered one.

A familiarity with a particular domain also plays a role in the annotation process, affecting both the senses assigned and the decisions regarding which collocations should be considered MWEs. For instance, the expression “rock formation” is not part of PWN, but it appears many times in the corpus.

1. a national park in Utah having colorful *rock formations* and desert plants and wildlife (08603525-n)
2. the gradual movement and formation of continents (11434448-n)

Although some of us believe the expression should be added to PWN, it is not in the lexicon yet, and so, some annotators chose the sense “(geology) the geological features of the earth” for the word “formation” in all occurrences of the expression. This decision was understandable if we consider that the word “rock”, in one of its senses, naturally evokes the domain “geology”. The same can also be said for the word “continent” in Example 2. But one annotator, a geology expert, consistently took the sense “a particular spatial arrangement” for the word “formation” in this expression. His decision was based on the strict interpretation of “geological formation” as a domain-specific concept and reinforced by the fact that “geological formation” in PWN has “physical object” as its hyperonym, not “formation” (as a process).

Some cases of multiword expressions (MWE) seem to support our belief that sense annotation and PWN maintenance should be joint work. First, we need to define and enforce heuristics to determine when a given word sequence is a multiword expression (being sense annotated as a single entity), and when its component tokens should be annotated individually. The compositionality and conventionality criteria from [11] may help, however these criteria are not as clear-cut as we would like them to be. Take the case of “first degree” and the example “all of the terms in a linear equation are of the *first degree*” in its definition (synset 05861716-n); we can annotate it as “first degree” (this same sense being defined in the synset where the example is given); but there is no sense for “second degree”, or “third degree”, which are equally valid. This leads us to consider that it should be annotated individually, and that the “first degree” sense should be removed from PWN.

Another possible approach we are considering is to follow the criteria adopted in [31] and its implementation in the ERG grammar. The alignment of the sense annotation (in the tokens) with the predicates (from the MRS) reveals a lot of cases to consider going far beyond the tokenization mismatches. For some idiomatic expressions we can really on ERG lexicon information (e.g. “the one where digestion **takes place**”). We also have to consider the coordinations of expressions sharing tokens. In the definition “a semisolid mass of coagulated **red** and **white blood cells**”, our sense annotation explicit reuse the tokens “blood cells” and annotate two senses “red blood cell” and “white blood cell”. In the MRS representation, it is yet not clear in which predicate (or set of predicates) to attach the sense identifiers. The easier cases are the adjective-noun constructions, sharing handlers in the MRS representation (e.g. “big toe”), this is not too different from the case of multiple adjectives modifying the same noun (e.g. “uric acid”). But what about the noun-noun compounds? ERG has a special abstract predicate called “compound” to represent the underspecified relation between two nouns. For instance, in “a **blood disease** characterized by an abnormal multiplication of macrophages“, the **blood** noun is related to **disease** with this abstract predicate (consider it an underspecified preposition connecting this two nouns). The sense tagging annotate the two words as a glob associated to the sense “blooded disease” but not all compounds may represent a single sense (examples above). Verbal phrases such as “**coughing up** blood from the respiratory tract” is another frequent case where tokenization differ. In the manual sense annotation, the two tokens (“coughing” and “up”) forms a glob annotated with one sense, the ERG analyses produces a single predicate.

All of the cases we highlighted above are being considering in the on going work of expand the ERG lexicon with all WordNet lexical entries, an exploratory work that aims to support a more semantic driven parsing selection and raking model.

References

- 1 Eneko Agirre, Oier Lopez De Lacalle, Aitor Soroa, and Informatika Fakultatea. 2009. Knowledge-based wsd and specific domains: Performing better than generic supervised wsd. In *IJCAI*, pages 1501–1506.
- 2 Eneko Agirre, Oier López de Lacalle, and Aitor Soroa. 2018. The risk of sub-optimal use of open source nlp software: Ukb is inadvertently state-of-the-art in knowledge-based wsd.
- 3 Eneko Agirre and Aitor Soroa. 2009. Personalizing pagerank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 33–41. Association for Computational Linguistics.
- 4 Satanjeev Banerjee and Ted Pedersen. 2002. An adapted lesk algorithm for word sense disambiguation using wordnet. In *Computational Linguistics and Intelligent Text Processing*, pages 136–145, Berlin, Heidelberg. Springer Berlin Heidelberg.
- 5 Pierpaolo Basile, Marco de Gemmis, Anna Lisa Gentile, Pasquale Lops, and Giovanni Semeraro. 2007. Uniba: Jigsaw algorithm for word sense disambiguation. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 398–401, Prague, Czech Republic. Association for Computational Linguistics.
- 6 Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, Roberto Navigli, et al. 2021. Recent trends in word sense disambiguation: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. International Joint Conference on Artificial Intelligence, Inc.
- 7 Peter Clark, Christiane Fellbaum, and Jerry Hobbs. 2008a. Using and extending wordnet to support question-answering. In *Proceedings of the 4th Global Wordnet Conference*, pages 111–119, Hungary.
- 8 Peter Clark, Christiane Fellbaum, Jerry R Hobbs, Phil Harrison, William R Murray, and John Thompson. 2008b. Augmenting wordnet for deep understanding of text. In *Semantics in Text Processing. STEP 2008 Conference Proceedings*, pages 45–57.
- 9 Ann Copestake. 2002. *Implementing typed feature structure grammars*, volume 110. CSLI publications Stanford.
- 10 Ann Copestake, Dan Flickinger, Carl Pollard, and Ivan A Sag. 2005. Minimal recursion semantics: An introduction. *Research on language and computation*, 3(2):281–332.
- 11 Meghdad Farahmand, Aaron Smith, and Joakim Nivre. 2015. A multiword expression data set: Annotating non-compositionality and conventionalization for english noun compounds. In *Proceedings of the 11th Workshop on Multiword Expressions*, pages 29–33.
- 12 Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.
- 13 Dan Flickinger. 2000. On building a more efficient grammar by exploiting types. *Natural Language Engineering*, 6(1), pages 15–28.
- 14 Dan Flickinger. 2011. Accuracy v. robustness in grammar engineering. In Emily M. Bender and Jennifer E. Arnold, editors, *Language from a Cognitive Perspective: Grammar, Usage and Processing*, pages 31–50. CSLI Publications, Stanford, CA.
- 15 Dan Flickinger, Yi Zhang, and Valia Kordoni. 2012. Deepbank. a dynamically annotated treebank of the wall street journal. In *Proceedings of the 11th International Workshop on Treebanks and Linguistic Theories*, pages 85–96.
- 16 Michael Wayne Goodman. 2019. A python library for deep linguistic resources. In *2019 Pacific Neighborhood Consortium Annual Conference and Joint Meetings (PNC)*, pages 1–7. IEEE.
- 17 Sanda M. Harabagiu, George A. Miller, and Dan I. Moldovan. 1999. Wordnet 2: a morphologically and semantically enhanced resource. In *Proceedings of SIGLEX99: Standardizing Lexical Resources*, pages 1–8.

- 18 Michael C McCord. 2004. Word sense disambiguation in a slot grammar framework. Technical Report RC23397, IBM.
- 19 John Philip McCrae, Alexandre Rademaker, Ewa Rudnicka, and Francis Bond. 2020. English WordNet 2020: Improving and extending a WordNet for English using an open-source methodology. In *Proceedings of the LREC 2020 Workshop on Multimodal Wordnets (MMW2020)*, pages 14–19, Marseille, France. The European Language Resources Association (ELRA).
- 20 Rada Mihalcea and Dan I. Moldovan. 2001. extended wordnet: progress report. In *Proceedings of NAACL Workshop on WordNet and Other Lexical Resources*, pages 95–100.
- 21 George A Miller. 1993. The association of ideas. *The General Psychologist*, 29:69–74.
- 22 George A Miller and Christiane Fellbaum. 2007. WordNet then and now. *Language Resources and Evaluation*, 41(2):209–214.
- 23 George A Miller, Claudia Leacock, Randee Teng, and Ross T Bunker. 1993. A semantic concordance. In *Proceedings of the workshop on Human Language Technology*, pages 303–308. Association for Computational Linguistics.
- 24 Dan Moldovan and Adrian Novischi. 2004. Word sense disambiguation of wordnet glosses. *Computer Speech & Language*, 18(3):301–317.
- 25 Ian Niles and Adam Pease. 2003. Linking lexicons and ontologies: Mapping wordnet to the suggested upper merged ontology. In *Ike*, pages 412–416.
- 26 Stephan Oepen. 2001. [incr tsdb()] – competence and performance laboratory. User manual. Technical report, Computational Linguistics, Saarland University, Saarbrücken, Germany. In preparation.
- 27 Stephan Oepen, Kristina Toutanova, Stuart M Shieber, Christopher D Manning, Dan Flickinger, and Thorsten Brants. 2002. The lingo redwoods treebank: Motivation and preliminary applications. In *COLING 2002: The 17th International Conference on Computational Linguistics: Project Notes*.
- 28 Woodley Packard. 2015. *Full forest treebanking*. Ph.D. thesis, University of Washington.
- 29 Adam Pease and Andrew Cheung. 2018. Toward a semantic concordancer. In *Proceedings of the 9th Global Wordnet Conference*, pages 97–104, Nanyang Technological University (NTU), Singapore. Global Wordnet Association.
- 30 Alexandre Rademaker, Bruno Cuconato, Alessandra Cid, Alexandre Tesseracto, and Henrique Andrade. 2019. Completing the Princeton annotated gloss corpus project. In *Proceedings of the 10th Global Wordnet Conference*, pages 378–386, Wrocław, Poland. Global Wordnet Association.
- 31 Ivan A Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Computational Linguistics and Intelligent Text Processing: Third International Conference, CICLing 2002 Mexico City, Mexico, February 17–23, 2002 Proceedings 3*, pages 1–15. Springer.
- 32 Karin Kipper Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, University of Pennsylvania.

5.2 NLP-based Study of Universals of Linguistic Idiosyncrasy

Agata Savary (University Paris-Saclay, CNRS – Orsay, FR)

License  Creative Commons BY 4.0 International license
 © Agata Savary

According to William Croft, the universals of linguistic idiosyncrasy established so far include the following:¹⁴

1. Meaning becomes idiosyncratic before morphosyntactic form does (which is why most MWEs are syntactically mostly regular)
2. Idiosyncrasy leading to lexicalization is most likely to develop in:
 - a. (typifying) modifier-referent constructions (*women’s magazine*, *table leg*)
 - b. verb-object constructions (*to pull one’s legs*, *to pay a visit*)
3. Idiosyncrasy leading to grammaticalization is most common in:
 - a. adposition path (*in light of N*)
 - b. tense-aspect-modality-polarity path (*be go-ing to V*)
 - c. discourse marker → sentence connective path (*all the same*, *oh by the way*).

The PARSEME framework, with its cross-linguistically unified guidelines for verbal multiword expressions (VMWEs), and its corpus annotated for VMWEs in 26 languages, could help corroborate these hypotheses. In particular, it could provide some evidence about the frequency of lexicalization in verb-object constructions (2b above).

Namely, in the latest 1.3 release of the corpus [5], the global statistics for all 26 languages show that among the 127,500 annotated VMWEs:¹⁵

- 44,171 are labeled as light verb constructions (LVCs)
- 29,062 are inherently reflexive verbs (IRVS)
- 26,214 are verbal idioms (VIDs)

These 3 categories contain large percentages of verb-object combinations.

More precisely, according to the PARSEME guidelines:

- LVCs are combinations of semantically light verbs and predicative nouns expressing the semantics of the action or state. Two subcategories are defined. In LVC.full the verb’s subject is the noun’s semantic argument as in (sl) *imeti predavanje* ‘give a lecture’. In LVC.cause the verb’s subject is the cause or source of the noun, as in (en) *to grant right*. Most LVCs in their so-called *canonical forms* (the least syntactically marked syntactic variants which preserve the idiomatic reading), consist of verbs heading direct objects.¹⁶
- IRVs are combinations of a verb *V* and a reflexive clitic *R* such that (i) *V* never occurs without *R*, as in (sv) *gifta sig* (lit. *get-married oneself*) ‘get married’, or (ii) *R* distinctly changes the meaning or valency of *V*, as in (es) *recogerse* (lit. *to gather oneself*) ‘to go home’.
- VIDs gather cases not covered by other categories. The verb’s dependents are unrestricted, including subjects, as in (en) *a little bird told me*, direct objects, as in (ro) *a întoarce foaia* (lit. *to turn the sheet*) ‘to become harsher’, etc. The verb can have several

¹⁴ <https://gitlab.com/unlid-dagstuhl-seminar/unlid-2023/-/wikis/Universals-of-linguistic-idiosyncrasy-established-so-far>

¹⁵ <https://parseme.grew.fr/tables/?data=parseme/labels@1.3>

¹⁶ They can also be verbs with oblique complements, whether introduced by a preposition, as in (pl) *występować w obronie uciekinierów* (lit. *to stand out in the defense of refugees*) ‘to defend refugees’, or by a non-accusative case, as in (pl) *obdarzać kogoś zaufaniem* ‘to endow sb trust.INS’. These verb-oblique combinations are more rare than verb-object combinations.

dependents, as in (en) *cut a long story short*, or combine features from other VMWE categories, as in (sv) *sätta sig upp mot någon* (lit. *sit oneself up against someone*) ‘defy someone’.

The precise frequency and distribution of verb-object combinations in VMWEs in various languages can be estimated e.g. with the Grew-match corpus browser [2],¹⁷ however, care must be taken with the design of the queries. For instance query (1), which seems to straightforwardly correspond to hypothesis 2b, results in only 1,296 matches out of all 7,313 VMWEs annotated in Polish,¹⁸ i.e. only 17%.

```
(1) pattern {
      MWE [label]; %A MWE is a new node with a "label" feature
      MWE -> V;    %The MWE has at least two nodes, marked V
      MWE -> O;    %...and O
      V[upos=VERB]; %The universal POS of V is VERB
      V -[obj]-> O; %The dependency between V and O is obj
    }
```

At first sight, this seems to invalidate hypothesis 2b. However, a finer study of encoding and variants of VMWEs provides more insight into verb-object combinations they contain.

Firstly, the syntactic dependencies underlying the PARSEME VMWE annotations rely on the Universal Dependencies standards, corpora and tools [1, 3, 6]. In UD IRVs would often contain the *expl* relation (rather than *obj*) to indicate that the reflexive pronoun cannot be mapped on any semantic argument of the verb, even if, strictly syntactically speaking, the reflexive clitic is most often truly a direct object. On the other hand, all IRVs in Polish cannot be considered verb-object combinations since in some of them the reflexive is in dative or instrumental case, indicating an oblique rather than a direct object, as in (pl) *wyobrażać sobie* (lit. *to imagine oneself.dat*) ‘to imagine’. Therefore, a more precise query finding IRVs with the reflexive clitic really playing the role of a direct object might be as in (2).

```
(2) pattern {
      MWE [label="IRV"]; %the MWE is an IRV
      MWE -> R;        %It has a node...
      R[form="się"];   %...whose surface form is "się"
    }
```

Secondly, VMWEs frequently occur in syntactic variants which violate the prototypical verb-object structure with the verb dominating the noun via the *obj* dependency. Namely, in Polish, and likely in many other Slavic languages, objects often require *structural case*, i.e. their case depends on the presence of negation and on the form of the head [4]. When the verb is not negated or nominalized, as in (pl) *bić pianę* (lit. *to whip foam.acc*) ‘to speak a lot without adding much to the discussion’, the object is in accusative case. Otherwise it is in genitive, as in (pl) *nie bić piany* (lit. *not to whip foam.gen*) and (pl) *bicie piany* (lit. *whipping foam.gen*). The object also takes genitive case when preceded by some quantifiers, as in (pl) *bić dużo piany* (lit. *to whip a lot of foam.acc*). These case fluctuations may provoke the assignment of the *obl* rather than *obj* dependency between the verb and the noun, notably when the syntactic data in the PARSEME corpus stem from automatic parsing.

¹⁷<https://parseme.grew.fr/>

¹⁸<https://parseme.grew.fr/?corpus=PARSEME-PL@1.3>

Note also that when a verb in a VMWE is nominalised, it can be tagged with the NOUN part-of-speech rather than VERB. Other syntactic variants e.g. extractions, as in (pl) *kara, którą wymierzili* ‘the penalty which they imposed’ invert the direction of the dependency and change its label (here `acl:recl` from the noun *kara* ‘penalty’ to the verb *wymierzili* ‘imposed’). These subtleties suggest that query (1) should be more relaxed, as in (3), and even then it might not have enough coverage.

```

pattern {
  MWE [label="LVC.full"|"LVC.cause"|"VID"];
(3)  MWE -> V;
     MWE -> O;
     V -[obj]-> O;
}

```

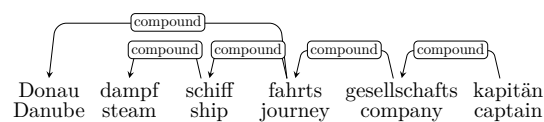
When these finer queries (2) and (3) are run on the Polish PARSEME corpus they indicate that VMWEs with direct objects are contained in at least:

- 1564 out of 3625 LVC and VID occurrences (43%)
- 3642 out of 3688 IRV occurrences (99%)
- 5206 out of all 7313 annotated VMWEs occurrences, i.e. **71%**

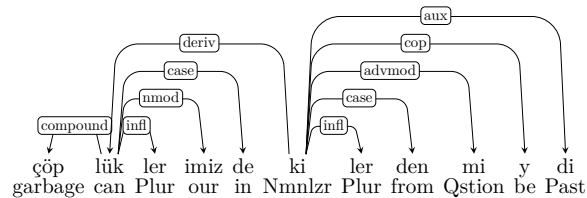
This is a much more encouraging quantification of idiosyncrasy in verb-object pairs, in relation to hypothesis (2b). Of course for this statistic tendency to be considered universal, similar sets of queries should be designed in order to appropriately cover the verb-object occurrences in other languages. It is also worth noting that hypothesis 2b and the others from the beginning of this section do not mention if the likelihood of particular types of idiosyncrasies is suggested with respect to occurrences or types (unique constructions). Here, we only dealt with the former but studying the latter is equally interesting.

References

- 1 Marie-Catherine De Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, **47**(2), pages 255–308.
- 2 Bruno Guillaume. 2021. Graph matching and graph rewriting: GREW tools for corpus exploration, maintenance and conversion. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pp. 168–175, Online. Association for Computational Linguistics.
- 3 Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.
- 4 Agnieszka Patejuk, Adam Przepiórkowski. 2014. Structural case assignment to objects in Polish. In Miriam Butt and Tracy Holloway King, editors, *The Proceedings of the LFG'14 Conference*, pages 429–447, Stanford, CA. CSLI Publications.
- 5 Agata Savary, Chérifa Ben Khelil, Carlos Ramisch, Voula Giouli, Verginica Barbu Mititelu, Najet Hadj Mohamed, Cvetana Krstev et al. 2023. PARSEME corpus release 1.3. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 24–35, Dubrovnik, Croatia. Association for Computational Linguistics.
- 6 Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.



■ **Figure 1** *Donaudampfschiffahrtsgesellschaftskapitän* “Danube steamship company captain”.



■ **Figure 2** *çöplüklerimizdekilerdenmiydi* “was it from those that were in our garbage cans?”.

5.3 Subword Relations, Superword Features

Daniel Zeman (Charles University – Prague, CZ)

License © Creative Commons BY 4.0 International license
© Daniel Zeman

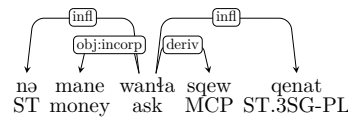
Introduction

Universal Dependencies (UD) subscribes to the lexicalist principle, claiming that dependency relations connect *words*, while the process of constructing words by combining smaller units (*morphemes*) is substantially different. Consequently, word-internal structure is normally not shown in UD.¹⁹ Nevertheless, there seems to be some demand [1] for a UD extension that would allow for showing word-internal structure in a way similar to how inter-word relations are represented. Here are some motivational examples:

- German compounds are written as one word and represented by one tree node in UD. English compounds may be perfectly parallel to the German ones, yet they are typically written as multiple orthographic words. In UD, they are multiple nodes connected by **compound** relations. The parallel structure is not visible in German UD but it could if the compounds were split into multiple tree nodes (Fig. 1).²⁰ Moreover, other annotation may pertain just to one part of a compound: We may want to annotate the MWE *Rolle spielen* “to play a role” in the compound *Hauptrolle spielen* “to play the main role”.
- Turkish words may combine several derivational and inflectional steps. Traditional analysis would break them up to *inflection groups* but in UD they are mostly kept together and the internal structure is not visible (unlike Fig. 2). See also [4].
- Chukchi transitive verbs may incorporate their objects and switch to intransitive inflection (Fig. 3) [5].
- Fieldworkers may prefer morpheme-based analysis when documenting a language; a UD example is the treebank of Beja [3].

¹⁹ Except for the optional **MSeg** and **MGloss** attributes in the MISC column of some treebanks, which can at least hint at the morphemic composition of a word.

²⁰ In fact, compounds are a gray zone. While most UD languages do not split them, they are split in Sanskrit UD, as such analysis is traditional in Sanskrit linguistics.



■ **Figure 3** *nəmanewan¹asqewqena* “they constantly asked for money”.

```
# text = Er spielt die Hauptrolle im Haus.
# text_en = He plays the main role in the house.
1  Er      er      PRON  _  Case=Nom|PronType=Prs 2  nsubj  _  _
2  spielt  spielen VERB  _  Mood=Ind|VerbForm=Fin 0  root   _  _
3  die     der     DET   _  Case=Acc|PronType=Art 4  det    _  _
4-6 Hauptrolle _      _      _      _      _      _
4  Hauptrolle Hauptrolle NOUN  _  Case=Acc|Number=Sing 2  obj    _  _
5  haupt    haupt   ADJ   _  Degree=Pos              6  amod   _  _
6  Rolle    Rolle   NOUN  _  Case=Acc|Number=Sing 4  wroot  _  _
7-8 im      _      _      _      _      _
7  in       in      ADP   _  _                        9  case   _  _
8  dem     der     DET   _  Case=Dat|PronType=Art 9  det    _  _
9  Haus    Haus    NOUN  _  Case=Dat|Number=Sing 2  obl    _  SpaceAfter=No
10 .       .       PUNCT _  _                        2  punct  _  _
```

■ **Figure 4** CoNLL-U with subword relations.

Precisely defining a *word* (even a *syntactic word*) cross-linguistically is a difficult task [2]. However, it matters less if we can annotate inter-word and intra-word relations in a similar manner. We propose to work within WG1 (and partially WG2) on an extension of UD that would support such annotation.

Subword Relations

As relations between subword units violate the lexicalist principle, they cannot be part of a regular UD treebank under the current guidelines; they have to be an extension that stands outside UD proper. Nevertheless, the file format should retain low-level compatibility with CoNLL-U so that existing tools can still process it. So, while new relation labels are conceivable, there should be no new line types beyond the existing 5 (comment, multiword token, node, empty node, empty line). There may be extra columns for readability (CoNLL-U Plus²¹) but it should be possible to collapse them into MISC attributes if needed.

Ideally, the format should accommodate normal UD treebank plus additional subword annotation and there should be a script that throws away the extra relations and extracts the regular UD treebank. If a word is decomposed, the relations between its parts should probably form a tree (\Rightarrow single root). The annotation of the root morpheme will differ from the annotation of the whole word, so we need nodes for both.²² Multiword token lines must be used to indicate the mapping of the nodes to surface tokens (Fig. 4).

²¹ <https://universaldependencies.org/ext-format.html>

²² As one of the reviewers noted, this has drawbacks, too. Parallelism between languages will be somewhat spoiled, as German *Hauptrolle* will now have three nodes, while English *main role* will have only two. Alternatively, the word-level morphological annotation could be stored for the morphemes spanning the word in a similar manner to what we propose for superword features in the next section.

Superword Features

Conversely, we may want to assign word-like annotation to a multiword expression. For example, a MWE functions like an adverb although its member words are not adverbs. Some treebanks already mark this with `MWEPOS=ADV` (or `ExtPos`) in MISC. Similarly, for German verbs with separable prefixes (e.g. *ein/steigen* “get on”), we may want to indicate the lemma that describes the two parts together. We may also want to add morphological features to sets of words, e.g., `Tense=Fut` for periphrastic future (composed of words that are not future themselves).

The MWE does not have to be linearly contiguous, so we cannot abuse multiword token lines for this purpose. MWEs tend to be catenas,²³ suggesting that the MISC column of the head node could hold such annotations. They are not complete subtrees though: in *I have come home*, the head of the periphrastic verb form *have come* is *come*, but we want to exclude the other dependents (*I* and *home*) from the annotation of the verbal features. We thus need a MISC attribute with the IDs of the nodes that are included in the MWE, e.g., `MWSpan=1-3,5`.

Multiple MWEs could have their annotation placed at the same head node, meaning that we have to use numeric ids to mark MISC attributes that pertain to the same MWE. For example, in *He has played the main role in the process*, we could annotate `MWSpan[1]=2-3 | MWLemma[1]=play | MWUPOS[1]=VERB | MWAspect[1]=Perf` and `MWSpan[2]=2-3,6 | MWLemma[2]=play role | MWUPOS[2]=VERB | MWAspect[2]=Perf`. Essentially, what we are looking at is a constituent-oriented analysis combined with dependencies, although ‘constituents’ in this sense are not linearly contiguous spans of words.

Acknowledgements

This work was supported by the grants 20-16819X (LUSyD) of the Czech Science Foundation; and LM2023062 (LINDAT/CLARIAH-CZ) of the Ministry of Education, Youth, and Sports of the Czech Republic.

References

- 1 Timothy Baldwin, William Croft, Joakim Nivre, and Agata Savary. 2021. Universals of Linguistic Idiosyncrasy in Multilingual Computational Linguistics (Dagstuhl Seminar 21351). *Dagstuhl Reports*, 11(7):89 – 138.
- 2 Martin Haspelmath. 2022 Draft. Defining the Word.
- 3 Sylvain Kahane, Martine Vanhove, Rayan Ziane, and Bruno Guillaume. 2021. A morph-based and a word-based treebank for Beja. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, pages 48 – 60, Sofia, Bulgaria. Association for Computational Linguistics.
- 4 Büşra Marşan, Salih Furkan Akkurt, Muhammet Şen, Merve Gürbüz, Onur Güngör, Şaziye Betül Özateş, Suzan Üsküdarlı, Arzucan Özgür, Tunga Güngör, and Balkız Öztürk. 2022. Enhancements to the BOUN treebank reflecting the agglutinative nature of Turkish. *arXiv preprint*, arXiv:2207.11782.
- 5 Francis Tyers and Karina Mishchenkova. 2020. Dependency annotation of noun incorporation in polysynthetic languages. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 195 – 204, Barcelona, Spain (Online). Association for Computational Linguistics.

²³ Even catena is probably not always granted. Grouping auxiliaries without the main verb would be a problem, although one may argue that this can be left for SUD to deal with. But coordination may complicate things. In *The food has been cooked and eaten*, one may want to combine the auxiliaries not only with *cooked* but also with *eaten*. Maybe we can say that this would be a catena in the enhanced dependency graph.

Participants

- Timothy Baldwin
MBZUAI – Abu Dhabi, AE
- Emily M. Bender
University of Washington –
Seattle, US
- Archana Bhatia
Florida IHMC – Ocala, US
- Nina Böbel
Universität Düsseldorf, DE
- Francis Bond
Palacký University Olomouc, CZ
- Gosse Bouma
University of Groningen, NL
- Jörg Bücker
Universität Düsseldorf, DE
- Mathieu Constant
ATILF – Nancy, FR
- Marie-Catherine de Marneffe
FNRS – UC Louvain, BE & Ohio
State University – Columbus, US
- Kilian Evang
Universität Düsseldorf, DE
- Daniel Flickinger
North Newton, US
- Omer Goldman
Bar-Ilan University –
Ramat Gan, IL
- Jan Hajic
Charles University – Prague, CZ
- Dag Haug
University of Oslo, NO
- Sylvain Kahane
University Paris Nanterre, FR
- Laura Kallmeyer
Universität Düsseldorf, DE
- Maria Koptjevskaja-Tamm
Stockholm University, SE
- Lori Levin
Carnegie Mellon University –
Pittsburgh, US
- Peter Ljunglöf
University of Gothenburg, SE
- Teresa Lynn
MBZUAI – Abu Dhabi, AE
- Christopher Manning
Stanford University, US
- Nurit Melnik
The Open University of Israel –
Raanana, IL
- Joakim Nivre
Uppsala University, SE
- Alexandre Rademaker
IBM Research – Sao Paulo, BR
- Carlos Ramisch
Aix-Marseille University, FR
- Manfred Sailer
Goethe-Universität Frankfurt am
Main, DE
- Agata Savary
University Paris-Saclay, CNRS –
Orsay, FR
- Nathan Schneider
Georgetown University –
Washington, DC, US
- Sara Stymne
Uppsala University, SE
- Reut Tsarfaty
Bar-Ilan University –
Ramat Gan, IL
- Francis M. Tyers
Indiana University –
Bloomington, US
- Ekaterina Vylomova
The University of Melbourne, AU
- Leonie Weissweiler
LMU München, DE
- Nianwen Xue
Brandeis University –
Waltham, US
- David Yarowsky
Johns Hopkins University –
Baltimore, US
- Amir Zeldes
Georgetown University –
Washington, DC, US
- Daniel Zeman
Charles University – Prague, CZ

