Report from Dagstuhl Seminar 21351

Universals of Linguistic Idiosyncrasy in Multilingual Computational Linguistics

Edited by

Timothy Baldwin[1], William Croft[2], Joakim Nivre[3], and Agata Savary[4]

1 **The University of Melbourne, AU**
2 **University of New Mexico – Albuquerque, US**
3 **Uppsala University, SE**
4 **Université de Tours – Blois, FR**

Abstract

Computational linguistics builds models that can usefully process and produce language and that can increase our understanding of linguistic phenomena. From the computational perspective, language data are particularly challenging notably due to their variable degree of **idiosyncrasy** (unexpected properties shared by few peer objects), and the pervasiveness of non-compositional phenomena such as **multiword expressions** (whose meaning cannot be straightforwardly deduced from the meanings of their components, e.g. red tape, by and large, to pay a visit and to pull one's leg) and constructions (conventional associations of forms and meanings). Additionally, if models and methods are to be consistent and valid across languages, they have to face specificities inherent either to particular languages, or to various linguistic traditions.

These challenges were addressed by the Dagstuhl Seminar 21351 entitled "Universals of Linguistic Idiosyncrasy in Multilingual Computational Linguistics", which took place on 30-31 August 2021. Its main goal was to create synergies between three distinct though partly overlapping communities: experts in typology, in cross-lingual morphosyntactic annotation and in multiword expressions. This report documents the program and the outcomes of the seminar. We present the executive summary of the event, reports from the 3 Working Groups and abstracts of individual talks and open problems presented by the participants.

Seminar August 30–31, 2021 – http://www.dagstuhl.de/21351
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.11.7.89
Edited in cooperation with Artur Kulmizev, Meagan Vigus

Universals of Linguistic Idiosyncrasy in Multilingual Computational Linguistics, *Dagstuhl Reports*, Vol. 11, Issue 07, pp. 89–138
Editors: Timothy Baldwin, William Croft, Joakim Nivre, and Agata Savary
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Executive Summary

Timothy Baldwin (The University of Melbourne, Australia, tbaldwin@unimelb.edu.au)
William Croft (niversity of New Mexico, Albuquerque, USA, wcroft@unm.edu)
Joakim Nivre (Uppsala University, Sweden, joakim.nivre@lingfil.uu.se)
Agata Savary (University of Tours, France, agata.savary@univ-tours.fr)

This Dagstuhl Seminar was initially planned as a 1-week event in June 2020 (with number 20261) with the following objectives:

- **Theoretical**: To deepen the understanding of language universals, and of how they apply to linguistic idiosyncrasy, so as to further promote unified modelling while preserving diversity.
- **Practical**: To improve the treatment of idiosyncrasy in treebanking frameworks, in computationally tractable ways and, thus, to foster high quality NLP tools for more languages with greater typological diversity.
- **Networking**: To promote a higher degree of convergence across typology-driven initiatives, while focusing on three main aspects of language modelling: morphology, syntax, and semantics.

Due to the COVID-19 pandemic, the event was first rescheduled and finally reduced to a 2-day online event on 30-31 August 2021, with two 3-hour sessions, repeated for better inclusiveness of various time zones (which corresponds to about 20% of the initially planned duration).

Prior to the event, participants submitted discussion issues, based on which working groups and the program were formed, as described in our Wiki space[1].

More precisely, the program of the event followed the Dagstuhl model:

- A list of recommended **readings** was published prior to the event
- **Introductory talks**, given by the 4 organizers, ensured common understanding of the scope and challenges to address.
- **Personal introductions** of all participants helped achieve a community building effect, despite the online setting.
- **Working groups** (WGs) were built on the basis of the discussion issues submitted by the participants. Each WG had 4 co-leaders, at least one of which could attend repeated sessions, so as to ensure consistency between the 2 time-zone sub-groups. The following WGs were created:
  - WG1: What counts as a word?
  - WG2: What counts as a MWE and as a construction?
  - WG3: Syntax vs. semantics
- **Discussion issues** were addressed in WGs by the proposers' short introductions followed by brainstorming.
- Plenary **reporting** sessions from WGs took place twice for every time zone.

The event attracted 51 participants, who judged it successful and expressed the need for a full-size onsite follow-up event. All the organizational details and outcomes of the seminar are gathered in our Wiki space[2].

[1] https://gitlab.com/unlid/dagstuhl-seminar/-/wikis/home
[2] `https://gitlab.com/unlid/dagstuhl-seminar/-/wikis`

Despite its very reduced and fully online format, the seminar achieved part of its objectives, stressed the importance of some initially-defined research questions, gave rise to new questions, and showed the efficiency of some instruments.

- On the **networking** side, the intended convergence effect was clearly apparent. While the initial proposal and invitee list was dominated by NLP-oriented members of the UD and PARSEME communities, strong contributions came notably from the less numerous typology and UniMorph experts. The four communities interacted actively, and reinforcing these interactions is intended for the near future. Notably, steps were taken towards:
  - integrating typology experts in the PARSEME core group
  - accompanying a seminal work in typology (Croft, to appear) with a "companion volume" about practical implementation of morphosyntactic concepts in UD.
- On the **theoretical** side, the event showed:
  - The importance of the research question *How to identify words across languages?* (item I.A in the seminar proposal), to which the whole of Working Group 1 was dedicated. In particular, new insights from lesser-studied languages, brought by typology experts, allowed us to broaden the perspective on this issue.
  - The need for capturing the relationship between the two fundamental notions in this proposal: a multiword expression and a construction, studied by Working Group 2. From the linguistic and typology perspective, a MWE is a special case of a construction, which is rarely made explicit in current NLP models. But the notion of a construction needs a more formal definition to be implementable in NLP, notably as far as the type-token opposition is concerned (question II.B in the seminar proposal). Thus, the typology-NLP interactions are essential in the quest for an optimal model.
  - The scope of the syntax-semantics interface issues (question II in the proposal) addressed by Working Group 3. On the one hand, the interests of the community in this respect exceeded the scope intended by the event organizers. Namely corpus-lexicon interlinking for all language units, not only for MWEs, was targeted. On the other hand, MWEs are exemplars of condensed syntax-semantic interface issues, and as such provide good case studies in this domain.
- On the **practical** side, some initial proposals emerged as to harmonizing UD treebank annotation guidelines with: (i) modelling morphological properties at the subword level (heavily studied by UniMorph), (ii) labelling MWEs (core activity of PARSEME).

Each multidisciplinary approach like ours bears heavy risks of intractability. This is because different communities often have different objectives and points of view on the same phenomena, and they may fail to agree on a unified approach, or even on the usefulness of working towards such a unification. In our case, there is a tension between:

- diversity and descriptive detail required in linguistics,
- necessary simplifications for the sake of robustness in NLP.

In other words, it is legitimate to question the usefulness of universality-driven initiatives (in NLP) if idiosyncrasy and diversity are basic properties of language data. Yet even typologists seek language universals which abstract away from the idiosyncrasy.

We feel that the event allowed us to mitigate this tension. Namely, even if a universality-based treebank fails to render the diversity of possible analyses of a language phenomenon, it is still useful not only for NLP applications but also for linguistic and typological analyses. This is because relevant examples are easy to extract (and to further re-interpret), as long as the annotation is consistent and well-documented.

Another barrier-lifting effect of the event concerned the relation between UD and PARSEME. It seems that the MWE categories defined by UD and PARSEME are less incompatible than initially expected, simply because the definition of an MWE in itself is different in UD and PARSEME. This could have been a source of major incompatibility but since a MWE does not really have a status in the UD annotation process, the discrepancies could (at least in some cases) be overcome relatively easily.

In conclusion, the event provided, in our opinion, a proof of concept for the framing objectives set up in the original Dagstuhl seminar proposal. However, since the effective framework and duration was severely reduced as compared to the initially intended setting, only part of these objectives could be achieved. Thus, we are currently putting efforts to ensure follow-up events. In particular, a new Dagstuhl seminar with roughly the same objectives has been submitted.

2 Table of Contents

Executive Summary

Overview of Talks

Working groups

Open problems

3 Overview of Talks

3.1 Multiword Expressions

Timothy Baldwin (The University of Melbourne, AU)

A multiword expression ("MWE") satisfies the following two conditions: (1) it is decomposable into multiple simplex words; and (2) it is lexically, phonetically, phonologically, morphosyntactically, semantically, and/or pragmatically idiosyncratic. Given the focus of this workshop on MWEs, Universal Dependencies, and Linguistic Typology, we primarily focus on lexical, morphosyntactic, and semantic idiosyncrasy in this talk, in addition to noting that our definition relies crucially on the concept of "simplex word". Lexical idiosyncrasy occurs when an MWE has one or more elements which do not have a usage outside of MWEs, such as *ad* in *ad hoc*, or *fro* in *to and fro*. Morphosyntactic idiosyncrasy occurs when the morphosyntax of the MWE differs from that of its components, as happens in the case of the transitive *wine and dine [SOMEONE]* "entertain [SOMEONE] with wine and food", as distinct from the intransitive *dine* (and also *wine*, although the simplex verbal usage of *wine* is, in itself, uncommon). Semantic idiosynrasy occurs when the meaning of the MWE is not simply the sum of its parts, either due to there being a mismatch in semantics (e.g. *blow hot and cold* "alternate between two polar-different moods or attitudes", which is completely divorced semantically from its component words) or there being extra semantics encoded in the MWE not found in the component words (e.g. *designated driver* being associated with the specific situation of a group going out to drink alcohol, with the designated driver making sure they drink in such a way as to be able to legally drive the group home).

Particular complications with MWEs as pertain to this workshop include: (a) What is a "word" in our definition, noting complications with non-segmenting languages, and also languages without a pre-existing writing system (Walpiri, Mohawk, ...)? (b) How to proceduralise/text for the different forms of idiosyncrasy in a way which generalises across different languages? (c) What is an MWE and what is (purely) constructional? (d) How should MWEs be represented to capture their (cross-linguistic) idiosyncrasies (but also their compositionality)?

Determinerless prepositional phrases (i.e. PPs where the head noun is a singular count noun, and lacks a determiner, such as *in gaol* or *per student*) are an excellent case study of the complexities of determining whether a given expression is an MWE or not. Two properties of particular interest with determinerless PPs are: productivity (i.e. how productive is a given preposition in combing with different nouns), and modifiability (i.e. can the noun in the determinerless PP be pre- or post-modified). In English, determinerless PPs populate the full spectrum from non-productive, non-modifiable constructions such as *ex cathedra* to fully-productive, fully-modifiable constructions such as *per*, e.g. *per recruited student that finishes the project.* Most determinerless PPs, however, lie between these extremes, and have limited productivity, and also idiosyncratic modifiability properties.

3.2 Multiword expressions: constructional and typological perspectives

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

Multiword expressions have played a large role in construction grammar (Goldberg 1995, 2006; Croft 2001), and construction grammar provides a different and possibly useful perspective on the treatment of MWEs in computational linguistics. In computational linguistics, MWEs are often described as "words with spaces", that is, MWEs are assimilated to the lexicon. In construction grammar, one can describe constructions as "MWEs without words", that is, syntax is assimilated to MWEs. Syntactic constructions are organized in a lattice, in which the highest nodes are schematic and general syntactic structures, and the lowest nodes are the most specific and restricted constructions – that is, prototypical MWEs (Croft and Cruse 2004). The real issue in analyzing MWEs from a constructional perspective is: how general is the construction/MWE and its parts?

The seminal paper in construction grammar (Fillmore et al. 1988) classifies idioms in a way that demonstrates the continuum of generality. Idioms are made up of unfamiliar pieces: words that occur nowhere else, or familiar pieces: words that occur in other constructions. The pieces are unfamiliarly arranged: a syntactic pattern occurring nowhere else, or familiarly arranged: a syntactic pattern found in other constructions. Both occur to varying degrees.

The most MWE-like are fully substantive: every part of the construction is a specific word. An MWE like the rhetorical question *Who's gonna make me?* consists of familiar pieces familiarly arranged. An idiom like *all of a sudden* consists of familiar pieces but they are familiarly arranged. The idiom *kith and kin* includes an unfamiliar piece, *kith*, but in a familiar NP coordination construction. For a truly unfamiliar arrangement of unfamiliar pieces, one must turn to borrowed phrases such as *joie de vivre.*

The problematic cases for MWE analysis are constructions which are partly general ("familiar") in some way or another. In the *The Xer, the Yer* construction, as in *The bigger, the better*, the form *the* is unfamiliar – it comes from an Old English oblique demonstrative form, not the definite article – and so is the parallel paratactic construction for comparatives. The unfamiliar piece *heed* occurs only in *pay heed to* and *take heed of*, which are familiar constructions (Nunberg, Sag and Wasow 1994). Likewise, comparative *than* occurs in only the comparative construction, which fluctuates between the older elliptical subordinate clause construction *She is taller than I am* and the innovative oblique phrase construction *She is taller than me.* With respect to familiar pieces, the construction *Nth cousin M times removed* represents an otherwise unfamiliar syntactic pattern. So do the English Auxiliaries, which have unique syntax in negative and interrogative constructions, as in *Isn't she nice?* (Bybee and Thompson 1997).

Finally, familiar pieces familiarly arranged may also have unfamiliar semantics and differing degrees of flexibility (generality) in syntax. The idiomatically combining expression *pull strings* is a substantive argument structure construction, consisting of only the verb *pull* and the noun *strings* as well as a schematic Subject role. It occurs in a variety of other constructions, including the Passive and the Object Relativization constructions: *Strings were pulled for me; the strings that she pulled for me….* Other substantive constructions such as the idiomatic phrase *kick(ed) the bucket* are more restricted, occurring only in different tense-aspect constructions.

These constructions raise the question: where do we draw the line for MWEs in this continuum of syntactic and semantic specificity? Should we draw the line at all? Even the most general, most "compositional" constructions, the modifier-noun construction and argument structure constructions have semantic idiosyncrasies. *Red pen* could refer to the color of the surface of the pen, stripes on the pen, the ink, and so on. The hue described by *red* varies with wine, hair, beans and so on. Argument structure constructions have verb-specific meanings for Subject, Object and so on; compare *I dried the dishes, I saw the hawk, She entered the room, This switch calibrates the temperature.*

Typologically, MWEs display certain patterns. Some MWEs evolve into single words, for example Proto-Basque **gu-re kide-a-n* 'in the company of us' > Basque *gu-rekin* 'with us' (Trask 1996:115-16) or Old English *ear wicga* 'ear one_that_moves' > *earwig* (Brinton and Traugott 2005:50). As a result, all of the idiomatic patterns described above occur with morphemes in words as well as in multiword expressions. This raises the question of how important is it to draw the "word level" line, not to mention how difficult is it? (Zingler 2020)

The syntax of MWEs displays some common patterns across languages, based on their diachronic origin. MWEs that grammaticalize to inflections are typically syntactically fixed phrases, or become so. The commonest of these are flags (case markers), TAMP (tense-aspect-modality-polarity) markers, and conjunctions. MWEs that lexicalize to referring phrases use the same strategies as modification constructions (Pepper 2020), and tend to be syntactically fixed, though often morphologically flexible. MWEs that evolve to complex predicates are the most varied in origin. They include light verbs (*They had a drink; You paid attention to me!*), serial/compound verbs (*Go fetch the paper*), copulas (*She is a teacher*), verb-argument phrases (*Strings were pulled for me, Butter wouldn't melt in Pat's mouth*) and verb-particle constructions (*She cut it up, The pond froze over*). Secondary predicates are also MWE-like (*The pond froze solid, They shot him dead/to death*). Complex predicates of all types tend to be syntactically flexible, and sometimes partly general.

The typology of multiword expressions remains to be explored in greater detail. They raise their own questions: Are there "universals of idiosyncrasy"? How do we find and formulate them?

References

1 Laurel J. Brinton and Elizabeth Closs Traugott. *Lexicalization and Language Change*. Cambridge University Press, Cambridge, UK, 2005

2 Joan L. Bybee and Sandra A. Thompson. *Three frequency effects in syntax*. Proceedings of the 23rd Annual Meeting of the Berkeley Linguistics Society, ed. Matthew L. Juge and Jeri O. Moxley, 378-88. Berkeley Linguistics Society, Berkeley, CA, 1997

3 William Croft. *Radical Construction Grammar: Syntactic Theory in Typological Perspective*. Oxford University Press, Oxford, UK, 2001

4 Croft, William and D. Alan Cruse. *Cognitive Linguistics*. Cambridge University Press, Cambridge, UK, 2004

5 Charles J. Fillmore, Paul Kay and Mary Catherine O'Connor. *Regularity and idiomaticity in grammatical constructions: the case of* let alone. Language 64:501-538, 1988

6 Adele E. Goldberg. *Constructions: A Construction Grammar Approach to Argument Structure*. University of Chicago Press, Chicago, IL, 1995

7 Adele E. Goldberg. *Constructions At Work: The Nature of Generalization in Language*. Oxford University Press, Oxford, 2006

8 Geoffrey Nunberg, Ivan A. Sag and Thomas Wasow.*Idioms*. Language 70:491-538, 1994

9 Steve Pepper. *The Typology and Semantics of Binominal Lexemes: Noun-Noun Compounds and their Functional Equivalents*. PhD thesis, University of Oslo, Oslo, NO, 2020

10 R. L. Trask. *Historical linguistics*. Arnold, London, 1996

11 Tim Zingler. *Wordhood Issues: Typology and Grammaticalization*. PhD dissertation, University of New Mexico, Albuquerque, NM, 2020

3.3 Principles of the UD Annotation Framework

Joakim Nivre (Uppsala University, SE)

Universal Dependencies is a framework for cross-linguistically consistent morphosyntactic annotation and a project to create annotated corpora in this framework for as many languages as possible. The project started in 2014 when version 1 of the annotation guidelines was released together with 10 treebanks. Since then, data sets have been released roughly every six months, with the most recent release (v2.8) featuring 202 treebanks representing 114 languages. A major milestone was the release of version 2 of the guidelines in 2016. For more information about the guidelines and resources, we refer to [5] for version 1 and to [6] for version 2. The linguistic theory underlying the annotation framework is laid out in [4].

The goal of UD is to enable cross-linguistically consistent morphosyntactic annotation to support multilingual research in natural language processing and linguistics. Ideally, UD should therefore facilitate meaningful linguistic analysis within and across languages and support morphosyntactic processing in monolingual and cross-lingual settings. To facilitate adoption of the framework, it is based on pre-existing de facto standards and common usage, in particular an evolution of (universal) Stanford dependencies [1, 2, 3], Google universal part-of-speech tags [7], and the Interset interlingua for morphosyntactic tagsets [8]. It is important to emphasize that UD is meant to complement – not replace – language-specific annotation schemes. For researchers interested in the finer details of a single language, UD may not be ideal since it is designed for cross-linguistic comparison and therefore by necessity has to abstract over some of these details.

A fundamental design principle of UD is a commitment to lexicalism, which in this context essentially means that the fundamental annotation units are words. Words have internal morphological properties and enter into syntactic relations with other words. Both of these aspects should be reflected in the annotation, but the principles are different and the annotation is therefore organized into two layers: a morphological layer and a syntactic layer. It is important to note, however, that the relevant notion of word here is that of a syntactic word – not a phonological or orthographical word – and UD therefore permits a two-level segmentation in order to recognize syntactic words over and above basic tokens.

In the morphological annotation layer, each word is assigned a lemma, a part-of-speech tag and (optionally) a set of features. Part-of-speech tags are taken from a revised and extended version of the Google universal part-of-speech tag set containing a fixed inventory of 17 categories. Morphological features are taken from a larger inventory that can be extended as the need arises, but where the names of features and feature values are standardized across languages.

In the syntactic annotation layer, words are connected by grammatical relations into a dependency tree, normally rooted in the main predicate of a sentence. The backbone of this structure consists of direct relations between predicates, arguments and modifiers, which are normally realized as content words. Function words are attached to the content word that they specify. This means, for example, that determiners are attached to nouns and that auxiliaries are attached to main verbs. Even adpositions are essentially treated as case markers of nominals rather than heads of prepositional phrases. The rationale for this conception of syntactic structure is to maximize parallelism across structurally different languages. By and large, relations between content words are more likely to be parallel across languages, while function words in one language often correspond to morphological inflection or nothing at all in other languages. UD provides a taxonomy of 37 universal relations for the classification of syntactic relations, with optional language-specific subtypes. The taxonomy is organized by two main principles, a distinction between core arguments and oblique modifiers at the clause level, and a distinction between three main types of linguistic structures: clauses, nominals and modifiers.

The annotation of multiword expressions (MWEs) is a challenge for UD, since they transcend the traditional morphology-syntax distinction assumed in UD. The current policy is to give a special treatment of MWEs only when they are morphosyntactically irregular. The clearest example is the special relation *fixed*, which is used to connect the components of a completely fixed grammaticalized MWEs such as *in spite of* or *by and large*. In addition, the relations *compound*, for any kind of word-level compounding, and *flat* for any kind of headless construction, can be used to annotate certain types of MWEs. However, the guidelines for handling multiword expressions in UD, and for relating the morphosyntactic UD annotation to specialized MWE annotation, is in need of further elaboration.

References

1 Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006).*

2 Marie-Catherine de Marneffe and Christopher D. Manning. 2008. The Stanford Typed Dependencies Representation. In *Proceedings of the COLING Workshop on Cross-framework and Cross-domain Parser Evaluation.*

3 Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. 2014. Universal Stanford Dependencies: A Cross-Linguistic Typology. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014).*

4 Marie-Catherine de Marneffe, Christopher Manning, Joakim Nivre, Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2), 255–308.

5 Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, Daniel Zeman. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016).*

6 Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, Daniel Zeman. 2020. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020).*

7 Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A Universal Part-of-Speech Tagset. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012).*

8 Daniel Zeman. 2008. Reusable Tagset Conversion Using Tagset Drivers. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008).*

3.4 Multilingual modelling of verbal multiword expressions in the PARSEME framework

Agata Savary (Université de Tours – Blois, FR)

Joint work of the PARSEME community
Main reference Agata Savary, Marie Candito, Verginica Barbu Mititelu, Eduard Bejček, Fabienne Cap, Slavomír Čéplö, Silvio Ricardo Cordeiro, Gülşen Eryiğit, Voula Giouli, Maarten van Gompel, Yaakov HaCohen-Kerner, Jolanta Kovalevskaitė, Simon Krek, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Lonneke van der Plas, Behrang QasemiZadeh, Carlos Ramisch, Federico Sangati, Ivelina Stoyanova, Veronika Vincze: "PARSEME multilingual corpus of verbal multiword expressions", Language Science Press, 2018.
URL http://dx.doi.org/10.5281/zenodo.1471591

PARSEME is a community which emerged from the European COST Action on *Parsing and Multiword Expressions* funded by the European Commission in 2013–2017 [4, 3]. It gathered 31 countries, 30 languages and 6 dialects from 10 language genera. It had a number of outcomes (publications, language resources, tutorials, methodologies and a book series[3]). Notably, a collaborative effort of 25 language teams resulted in annotation guidelines for verbal multiword expressions, unified across 25 languages, as well as in a manually annotated corpus for these languages. This resource has seen 3 releases so far, and is distributed under open licenses.

By the PARSEME definition, a multiword expression (MWE) is a continuous or discontinuous sequence of words which: (i) contains at least two *lexicalized components*, including a head word and at least one other syntactically related word, (ii) displays some degree of idiosyncrasy. A component of a MWE is said to be lexicalized[4] if replacing it by semantically related words results in a meaning shift which goes beyond what is expected from the replacement (as in ***turn the tables*** "change from a weaker to a stronger position" vs. *turn the chairs*, *rotate the tables*). Thus, by contrast with Croft's introductory talk in this workshop, the notion of lexicalization applies not only to phrases but to their components as well. PARSEME admits lexical, morphological, syntactic and/or semantic idiosyncrasy as a defining criterion for MWEs. Thus, by contrast with [1], collocations, i.e. expressions exhibiting statistical idiosyncrasy only, are not included in the scope of MWEs.

In the PARSEME guidelines and corpus, focus is on verbal MWEs (VMWEs). A VMWE is a MWE such that: (i) its *canonical form* builds a (weakly) connected graph whose head is a verb, (ii) it passes the idiosyncrasy tests from the PARSEME guidelines. A canonical form is the least syntactically marked syntactic variant which preserves the idiomatic reading, e.g. a finite verb, active voice, non-negated form and a form with no extraction are considered less syntactically marked than infinitive/participle, passive voice, a negated form, and a form with an extraction, respectively. For instance, if we come across the occurrences on the left-hand side of Fig. 1, we first transform it into a canonical form (like the one on the right-hand side), and this is the forms which we test against the guidelines.

The PARSEME guidelines were conceived with 3 main objectives in mind: (i) to formalise idiomaticity in a cross-linguistically unified and computationally tractable way, (ii) to unify what is truly similar, thus emphasizing what is language-specific, (iii) to make the annotation reproducible. These objectives imply a series of principles and constraints. Namely,

[3] *Phraseology and Multiword Expressions* at Language Science Press: `https://langsci-press.org/catalog/series/pmwe`

[4] In examples, hexicalized components are highlightes in bold.

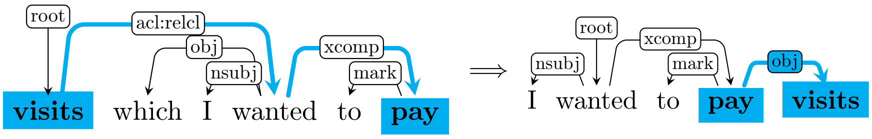


Figure 1 Transforming candidate VMWEs into their canonical forms.

annotation follows a decision diagram (with a unique starting point) and atomic decisions are binary (although non-compositionality is a matter of scale). Semantic non-compositionality is considered the major property to capture but is hard to test directly. Therefore it is approximated by lexical and morpho-syntactic inflexibility. For instance in French, *la porte s'ouvre* (lit. "the door opens itself") "the door opens" contains a combination of a verb (*ouvre* "opens") and of a reflexive clitic (*s'* "itself"), which might be idiomatic. However, this expression means roughly the same as *quelqu'un ouvre la porte* "someone opens the door", which proves that this is a regular middle passive (or inchoative) use of the reflexive clitic, rather than a VMWE. Such inflexibility tests are driven by the syntactic structure, which creates a strong dependence on the underlying syntactic theory. For the sake of cross-lingual validity, PARSEME annotation largely relies on the morpho-syntactic annotation provided by Universal Dependencies [2].

The PARSEME guidelines put forward a typology of VMWEs with categories of 3 kinds. Firstly, *universal categories* (i.e. occurring in all 25 languages under study) consist of: (i) *verbal idioms* (VIDs), like *to* ***call it a day***, and (ii) *light verb constructions* (LVCs) with two subtypes: LVC.full (*to* ***give*** *a* ***lecture***) and LVC.semi (***grant rights***). Secondly, *quasi-universal categories* occur in many languages but not all: (iii) *inherently reflexive verbs* (IRVs), like as in *to* ***help oneself*** "to take something freely", (iv) *verb-particle constructions* (VPCs) have two subtypes: VPC.full (*to* ***do in*** "to kill") and VPC.semi (*to* ***eat up*** "to eat completely", (v) *multi-verb constructions* (MVCs) like ***copy-paste***. Finally, *language-specific categories* are allowed and one has emerged so far: *inherently clitic verbs* (LS.ICV) in Itialian, like ***prenderle*** (lit. "to take it") "to be beaten".

The PARSEME quest for universality-oriented modeling of VMWEs opens many questions to which typology experts could greatly contribute. For instance, we would like to know if the two VMWE categories identified as univeral (VIDs and LVCs) truly occur beyond the 25 languages which we have studied. The annotation guidelines for MVCs also need more insight. There, Indo-European verb-verb constructions like ***make do*** or ***copy-paste*** in English are classified along with serial verbs like ***kar le-na*** (lit. "do take") "do something for one's own benefit" in Hindi. It remains unclear if this reflects true similarities rather than "false friends". Moreover, the statistics of the PARSEME corpus also reveal intriguing distributional phenomena. Notably, combinations with a verb and a reflexive clitic (as *I wash myself, she bought herself a present*) are frequent in many languages. Still, IRVs are frequent in Slavic and Romance, as well as in German, but rare or non-existent in other languages (e.g. English).

Future work in PARSEME also calls for stronger synergies with Universal Dependencies (UD). We already benefit from the UD definition of a word (a pre-requisite for defining a MWE) and from the UD morpho-syntactic annotations (pre-requisites for syntax-driven inflexibility tests). Thus, the universality of the UD categories and tags enables the universality of the PARSEME guidelines. However, joint challenges still need to be addressed. Firstly, the UD and the PARSEME definitions of a MWE are partly redundant and competing. For instance ***let alone*** in the following example is annotated as a MWE both at the level of UD dependencies (with the `fixed` label) and in the PARSEME layer (as a VID): *they never gave him a present, **let alone** a cake.* Moreover, some UD relations only partly overlap with the PARSEME categories. For instance, UD defines *inherently reflexive verbs* (marked with the `expl:pv` label) as those which never occur without the reflexive clitic. This corresponds to only part of the PARSEME criteria for IRVs.

Future work in the PARSEME corpora initiative includes: (i) extending the annotation guideines to new MWE categories (nominal, adjectival, adverbial, functional, . . .), (ii) unifying PARSEME and UD annotation guidelines, (iii) validating them by experts in typology, (iv) including new languages and language families, (v) continuous corpus enhancements with regular releases.

References

1 Timothy Baldwin and Su Nam Kim. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing*, pages 267–292. CRC Press, Taylor and Francis Group, Boca Raton, FL, USA, 2 edition, 2010.

2 Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. Universal Dependencies. *Computational Linguistics*, 47(2):255–308, 07 2021.

3 Carlos Ramisch, Agata Savary, Bruno Guillaume, Jakub Waszczuk, Marie Candito, Ashwini Vaidya, Verginica Barbu Mititelu, Archna Bhatia, Uxoa Iñurrieta, Voula Giouli, Tunga Güngör, Menghan Jiang, Timm Lichte, Chaya Liebeskind, Johanna Monti, Renata Ramisch, Sara Stymne, Abigail Walsh, and Hongzhi Xu. Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 107–118, online, December 2020. Association for Computational Linguistics.

4 Agata Savary, Manfred Sailer, Yannick Parmentier, Michael Rosner, Victoria Rosén, Adam Przepiórkowski, Cvetana Krstev, Veronika Vincze, Beata Wójtowicz, Gyri Smørdal Losnegaard, Carla Parra Escartín, Jakub Waszczuk, Matthieu Constant, Petya Osenova, and Federico Sangati. PARSEME – PARSing and Multiword Expressions within a European multilingual network. In *7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2015)*, Poznań, Poland, November 2015.

4 Working groups

4.1 Working Group 1 (What counts as a word?)

Francis Tyers, Ekaterina Vylomova, Daniel Zeman, Tim Zingler

Identifying words cross-linguistically

Tim presented a typological view (Zingler 2020). There is no single definition of word, applicable to all phenomena in all languages. However, there are clues that can help with a vast majority of cases. Phonological word vs. morphological word. Phonological clues: word = domain of stress/tone assignment, vowel harmony, phonologically conditioned allomorphy. Morphological clues: fixed order (relative position of each morpheme within a word is fixed), non-selectivity (words can co-occur with different word classes, affixes cannot). Example where phonological and morphological clues yield different results: English genitive *'s*. It is not a phonological word because it is prosodically dependent on the previous segment. However, it is syntagmatically independent (non-selective), i.e., a morphological word:

(1) a. [*The boy*]*'s dog*

b. [*The boy who ran away*]*'s dog*

Wordhood issues in Universal Dependencies

Dan presented several practical issues that emerged in Universal Dependencies. In UD, the surface segments identifiable by spaces and other special characters are orthographic words, while the nodes in dependency trees are syntactic words (essentially corresponding to morphological words mentioned above). The UD issues included the following (we only give more details about the first issue in this report):

- Diverging views of word boundaries in Japanese and Korean
- Incompatible treatment of definite markers in Arabic and Hebrew
- Incompatible treatment of subject agreement morphemes in Arabic and Amharic

Japanese vs. Korean

Japanese has no orthographic words, hence there is more freedom in deciding what should count as a word-level unit for annotation and language processing. There are several different traditions how to define words in Japanese, and the one used in UD is very fine grained, making Japanese look essentially as an analytical language.

Korean has space-delimited orthographic units, and these are treated as words in UD. However, these words are rather coarse-grained. In result, Korean looks very different from Japanese, although the two languages are usually considered typologically similar. (On the other hand, Korean UD bears some similarities with Turkic languages, which are also supposed to be typologically similar to Korean and Japanese.)

The following three examples are three possible segmentations of the same Japanese phrase, meaning “I went to the beauty salon of/in Kyodo”. The segmentation (2a), where verbal morphemes are treated as auxiliary words, is currently used by the Japanese UD team. (2b) would be advocated by some to be a better fit, but it is currently not used. In (2c) the postpositions are treated as case suffixes; this approach is not advocated for Japanese at all, yet it is probably the closest one to the current approach taken in Korean UD.

(2) a. *Kyōdō no miyōshitsu ni it te ki mashi ta*

b. *Kyōdō no miyōshitsu ni itte kimashita*
Kyodo of beauty.salon to going come

c. *Kyōdōno miyōshitsuni itte kimashita*

UniMorph issues related to clitics

Ekaterina presented some open problems in UniMorph, which has mostly data automatically dug out of Wiktionary (currently just the English edition of Wiktionary). The paradigm tables in Wiktionary often include analytical forms, such as the Polish conditional in (3a), or even extra paradigms for reflexive verbs, such as *podróżować się* in (3b).

(3) a. *byłybyście podróżowały*

"you would have dyed pink"

b. *byłybyście się podróżowały*

"you would have turned pink" ("you would have dyed yourself pink")

Open question: Should we keep such multi-word units in the inflectional database of UniMorph? (Related issue: The data are not consistent with respect to this. Different languages are treated differently in different language editions of Wiktionary.)

Reut: Often the morphological features of individual words participating in a periphrastic verb form (main verb + auxiliaries) do not straightforwardly show the features of the resulting form. For example, the German future *wir werden sehen* "we will see" contains a present auxiliary form and an infinitive of the main verb, but none of them alone can be described as `Tense=Fut`. So it would be actually useful to have some kind of "phrase-level features" where the future could be annotated.

Impact on studies carried on the data

Natalia presented two typological studies with UD-annotated (automatically parsed) data. In both studies, she ran the same experiment several times, each time with a slightly different definition of word (derivable automatically from the UD annotation). While the results did not vary too much for most of the languages, for some of them the difference was significant. Recommendation: manually annotated UD data should employ transparent and well documented tokenization decisions. In the ideal case, the user could switch between different levels of "word granularity". (Levshina 2020, 2021)

Artur presented another study his group did to see whether contextual neural representations have internal preference for a particular dependency scheme. The results were often in line with typological assumptions about the languages, however, some anomalies occurred, which may have been caused by inconsistent approaches to tokenization in various UD treebanks. This leads to a similar recommendation as in the study presented by Natalia. (Kulmizev et al. 2020)

Defining the word in little-known languages, with morphological singularities

Emmanuel: When describing languages that are lesser-known (and rarely written, e.g., creoles), the situation is similar to languages whose writing system does not delimit words overtly. It is often the case that various authors use various ad hoc conventions, which are not mutually compatible. Peculiar morphosyntactic properties of the language may further complicate the situation, for example, when a morphological agreement morpheme is separated from the verb whose agreement with an actant the morpheme encodes.

Dependency analysis of noun incorporation in polysynthetic languages

Fran: In polysynthetic languages, a large part of the interesting structure is hidden inside words, at the sub-word level. Example (4) is from the Chukchi UD treebank.

(4) *Qonpə nəwiswetsəqiwqinetʔəm nəmanewanɬasqewqenat*

“They (children) constantly went to play, constantly asked for money.”

The colored word is an example of a nominal object incorporated in a verb. Linguists agree that this should be one word. There are phonological clues such as vowel harmony, and also morphological clues: the yellow morphemes are verbal inflectional affixes. The blue morpheme is the verbal stem “to ask”, while the red mane is the incorporated object “money”. The verb uses intransitive inflection; if the object were not incorporated and appeared instead as an independent word (which is also possible in Chukchi), the verb would use its transitive inflection pattern. UD sticks to the word as the basic unit, corresponding to a node in a dependency tree. Consequently, the tree of (4) does not reveal that “money” is the object of “to ask”.

Dependency structure vs. word-internal morphology

David: UniMorph is currently expanding to derivational/compound morphology. Here it might be useful to annotate word-internal structure similarly to how relations between words are modeled in Universal Dependencies. Then the German compounds (5), which are typically one word in UD, could have a tree similar to what UD does with English compounds, which are typically written as multiple words.

(5) *Donau|dampf|schiff|fahrts|gesellschafts|kapitän*
lit. Donau steam ship journey company captain

Even in English, some compounds may be written with or without space (steam ship vs. steamship, or white space vs. white-space vs. whitespace), and these somewhat arbitrary orthographic choices would create asymmetric analyses in UD. Even if the space slightly changes the meaning of the compound, we want to have similar analyses:

(6) a. *We hired a dish washer*
b. *We hired a dishwasher*

Agata: Splitting German compounds is also important for annotation of multi-word expressions. For example, *Rolle spielen* “to play a role” is considered a MWE, and “role” can be modified by a word that is not part of the MWE. In English, the modifier is likely to be a separate word, thus it is easy to exclude it when annotating the MWE. However, in German the modification is likely to be realized as a compound: *Hauptrolle spielen* “to play the main role”.

Possible solutions to some of the issues

Defining word cross-linguistically

There might be "good enough" criteria that work 95% (or more) of the time. One such criterion is "fixed order": If no morpheme within a string of morphemes S can move without changing the meaning of S, then S is a word. This criterion should work most of the time, although there is at least one known exception from Huallaga Huánuco Quechua (Weber 1989: 221), where *huknayllamannaw* and *huknayllanawman* have the same meaning in (7a) and (7b):

(7) a.

Ishka-n	*tikra-sha*	*huknaylla-man-naw*
lit.	two-3P	turn-3PERF

"The two of them have become as though one."

b.

Ishka-n	*tikra-sha*	*huknaylla-naw-man*
lit.	two-3P	turn-3PERF

"The two of them have become as though one."

Annotating word-internal structure

A new layer of stand-off annotation over UD trees could be defined. This new layer would annotate word-internal structure in a fashion as similar to UD trees as possible. (Yet it would still sit clearly outside the principles of basic UD annotation, therefore it has to be a separate layer.) Such a layer could help both with annotating compounds and with the incorporated nouns in polysynthetic languages. The exact nature and taxonomy of the word-internal relations has yet to be discussed. Potential future collaboration between UD and UniMorph is foreseen here. Many of the word-internal dependency annotations could be precomputed and stored in the lexicon, with minimal context-sensitive ambiguity.

In a similar spirit, one could also think of a layer above the orthographic words, which would enable defining morphological features of periphrastic forms, or of multi-word expressions.

References

1 Artur Kulmizev, Vinit Ravishankar, Mostafa Abdou, Joakim Nivre (2020). Do Neural Language Models Show Preferences for Syntactic Formalisms? In Proceedings of ACL. https://aclanthology.org/2020.acl-main.375.pdf

2 Natalia Levshina (2020). How tight is your language? A semantic typology based on Mutual Information. In Proceedings of TLT. https://aclanthology.org/2020.tlt-1.7.pdf

3 Natalia Levshina (2021). Corpus-based typology: applications, challenges and some solutions. In: Linguistic Typology. https://doi.org/10.1515/lingty-2020-0118

4 Francis M. Tyers, Karina Mishchenkova (2020). Dependency annotation of noun incorporation in polysynthetic languages. In Proceedings of UDW. https://universaldependencies.org/udw20/papers/2020.udw2020-1.22.pdf

5 Tim Zingler (2020). Wordhood issues: Typology and grammaticalization (PhD dissertation). University of New Mexico. https://digitalrepository.unm.edu/ling_etds/71

4.2 Working Group 2 (MWEs and Constructions)

Steve Pepper, Lori Levin, Aline Villavicencio

What counts as an MWE and how are they classified?

How to define MWE

The 'standard' definition, "lexical items that can be decomposed into multiple lexemes, and display lexical, syntactic, semantic, pragmatic and/or statistical idiomaticity" (Baldwin & Kim 2010), was broadly accepted by the group, but the following questions were raised:

- Is statistical idiomaticity alone sufficient to qualify as an MWE? Tim now says no, others disagree.
- Should "lexical items" be replaced with the term 'complex constructions' (or "complex units"), in order to avoid a strict separation between lexicon and grammar?
- Why "syntactic" and not "morphosyntactic"?
- There are interdependencies between different kinds of idiosyncrasy, e.g. syntactic/semantic, etc. What are the others?
- How to determine the size and core components of an MWE?
  - Do we consider *Customer service 101* to be an MWE?
- Per Croft: Should we draw the line [between MWEs and 'vanilla' constructions]?
- Proposed revised definition: **"Multiword expressions are complex constructions that display significant lexical, morphosyntactic, semantic, pragmatic and/or statistical idiosyncrasy."**

How to classify MWEs

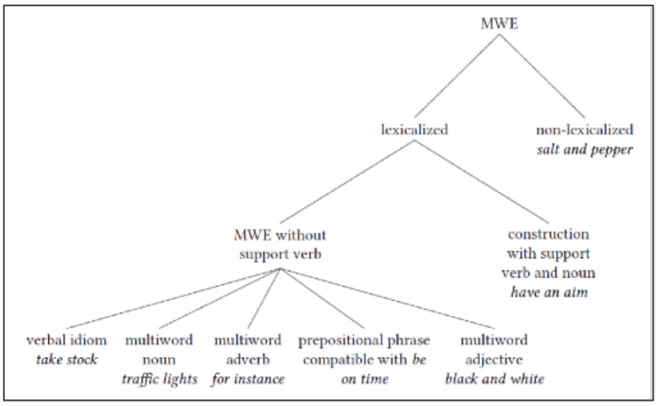


Figure 2 Proposed classification of MWEs.

Laporte (2018) observes that MWEs are "a heterogeneous set with a glaring need for classifications". However, his classification leaves much to be desired: it identifies only seven types of MWE; the top-level division is based on lexicalization – a gradient feature; the second privileges support verb constructions; and the third is a ragbag. There are similar classifications in Sag & al (2002) and Baldwin & Kim (2010). Questions raised:

1. How many different kinds of MWE are there? Can we enumerate them?
2. What types of MWE are widely (and less widely) observed cross-linguistically?

3. What classificatory principles should be applied in order to structure the domain of MWEs (grammatical functions, parts of speech, propositional acts)?
   a. Croft's take: "A primary division between **lexicalization of content words** and grammaticalization of function words. Within content words, a division between **complex predicates** (which are etymologically diverse, and often syntactically flexible) and **complex arguments and modifiers** (which usually are syntactically fixed, though often morphologically flexible). Within function words, a division between **relational function words** (case markers/flags and conjunctions), and **"other"**. Some cases that are problematic/don't fit in are idiosyncratic verb + argument structure combinations (e.g. verb + preposition) – but these are problematic in general – and pragmatically idiosyncratic expressions such as rhetorical questions."
4. Is it possible to infer the types from bottom-up annotation of corpora, for example using embedding-based approaches?

Annotating the morphemic level

1. Is the two-level word segmentation facility in UD, that is, into text (e.g. Fr. *aux*) and words (Fr. *à les*) sufficient to handle issues like the productivity of Bul. *-ica* [fem/dim] and the different orthographic systems (disjunctive and conjunctive) of closely related Bantu languages Northern Soto (Sotho) and Zulu (Nguni)?

	Orthography	Morphological analysis				
Northern Soto	ke a ba rata	*ke*	*a*	*ba*	*rat-*	-a
Zulu	ngiyabathanda	*ngi-*	*-ya-*	*-ba-*	*-thand-*	-a
	"I like them"	SC.1SG	pres	OC.CL2	verb root	inflectional ending

2. Is there a difference regarding the role of productivity (under the word level) and idiomaticity (above the word level)?

MWEs and constructions

How to define "construction"

Participants appear to have rather different notions of "construction", as evidenced by the question *is X an MWE or a construction?* In the absence of a common understanding of the term, communication is much more difficult. From the perspective of Construction Grammar, every MWE is by definition a construction:

- A construction is "the basic unit of morphosyntactic analysis ... a conventional pairing of form and function; its form is morphosyntactic structure, and its function is a combination of meaning (semantic content) and information packaging" (Croft 2022).
- What definitions of 'construction' do other people operate with?

MWEs as a subtype of construction

- If we accept that every MWE is a construction, do they constitute a separate subtype of construction?
- Or is it more reasonable to regard each MWE type a subtype of one or more other more general constructions, some of whose properties they share?
- Does an MWE "correspond to" one construction or (potentially) multiple constructions?

Comparing MWEs across languages

Which principles should be applied when comparing MWEs across languages? For instance, what allows us to perform the following two comparisons:

- The "Excess Construction", as in "so drunk he fell over": English the *so X [that] Y*, Chinese *X dao ('until') Y*, Japanese *Y hodo ('as.much.as') X*
- Eng. *in the black* vs. Port. *in the blue*

Annotation schemes for MWEs

Additional annotation schemes for MWEs

For some applications, some kinds of MWEs require more fine-grained annotation schemes than those offered by UD and PARSEME:

- Example 1: The semantic relation in noun-noun compounds and their functional equivalents is catered for only by **nmod** (and perhaps compound, which is something of a ragbag anyway). Hatcher-Bourque as a proposed annotation scheme.
- Example 2: The distinction between adverbial intensifiers and mitigators in Greek.
- Example 3: Degrees of productivity in English Verb-Particle Constructions.

Capturing types (or degrees) of idiosyncrasy

It is theoretically possible to specify how MWEs deviate from the prototypical constructions whose morphosyntactic strategies they recruit. Whether this is useful or not depends on what we are annotating (say, a corpus vs. a dictionary), and the needs of the application (e.g., parsing vs. text generation or translation). Some questions:

- Should we develop a "taxonomy of idiosyncrasies" that allows to annotate the ways in which MWEs deviate from the constructions they are instances of? For example, *how black and tan*, *salt and pepper*, *by and large* and *to and fro* deviate from the prototypical (conjunctive) coordinate nominal [N and N]N, coordinate adjectival [ADJ and ADJ]ADJ, and coordinate adverbial [ADV and ADV]ADV constructions:
- Does UD annotate enough to capture the full morphosyntactic idiosyncrasies of MWEs in context?
- What about other idiosyncrasies (lexical, semantic, pragmatic, statistical)?
- Should register be annotated?
- Can the STREUSLE approach of distinguishing between "strong" and "weak" semantic opacity be integrated into UD?
- More fundamentally, should semantic annotation (of MWEs or constructions in general) be part of UD, or should this be left to PARSEME?

Issues with UD

Some issues have been identified with UD:

- Please can we have better guidelines for "mischievous nominal constructions": names, dates, numbers; compounds; adverbial NPs?
- And also multiword connectives (*"out of"*, *"along with"*, *"based on"*, etc.)?

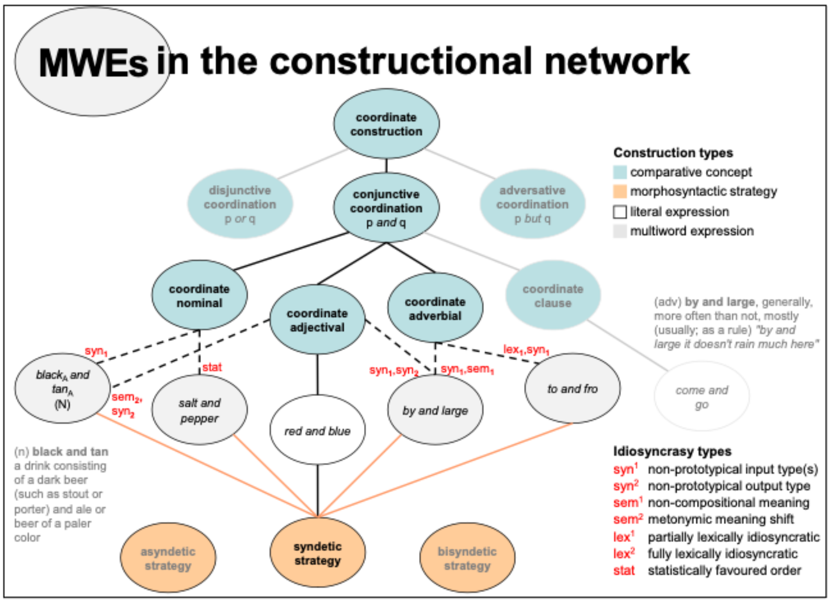

MWEs in the constructional network
coordinate construction
disjunctive coordination p or q
conjunctive coordination p and q
adversative coordination p but q
Construction types
comparative concept
morphosyntactic strategy
□ literal expression
multiword expression
coordinate nominal
coordinate adjectival
coordinate adverbial
coordinate clause
(adv) by and large, generally, more often than not, mostly (usually; as a rule) "by and large it doesn't rain much here"
syn_1
stat
sem_2, syn_2
syn_1,syn_2
syn_1,sem_1
lex_1,syn_1
$black_A$ and tan_A (N)
salt and pepper
red and blue
by and large
to and fro
come and go
(n) black and tan a drink consisting of a dark beer (such as stout or porter) and ale or beer of a paler color
Idiosyncrasy types
syn^1 non-prototypical input type(s)
syn^2 non-prototypical output type
sem^1 non-compositional meaning
sem^2 metonymic meaning shift
lex^1 partially lexically idiosyncratic
lex^2 fully lexically idiosyncratic
stat statistically favoured order
asyndetic strategy
syndetic strategy
bisyndetic strategy

Figure 3 MWEs in the constructional network, with annotation of idiosyncrasies.

Issues with PARSEME

Some issues have been identified with the PARSEME framework when annotating Basque corpora:

- PARSEME expects LVCs to consist of LV+Noun. In Basque they can consist of LV+Adj (when the adjective has a morphologically identical eventive noun, like in Hindi), but also LV+Adv. We can expect the same phenomenon to occur in other languages. Is there any reason why this cannot be easily fixed?
- Basque annotators had trouble treating make a call as an LVC while receive a call should be treated as a collocation (and thus should not be annotated). Since causal verbs like give a headache are accepted in LVCs, could/should more kinds of verbs be included as well?

Some questions for discussion:

- If we want the guidelines to be universal, how acceptable/necessary are language-specific notes? Should we preferably use more general definitions?
- How detailed should definitions be in order to make sure that annotators refer to the same phenomenon/concept in multiple languages, while accepting that these phenomena might vary across languages?
- Should collocations be treated as a purely statistical phenomenon? Why should causal verbs be accepted inside LVCs but other similar cases be discarded?

References

1 Baldwin, Timothy, and Su Nam Kim. "Multiword expressions." Handbook of natural language processing 2 (2010): 267-292.

2 Croft, William. "Morphosyntax: constructions of the world's languages." (2022).

3 Sag, Ivan A., et al. "Multiword expressions: A pain in the neck for NLP." International conference on intelligent text processing and computational linguistics. Springer, Berlin, Heidelberg, 2002.

4.3 Working Group 3 (Syntax vs. Semantics)

Emily M. Bender, Jan Hajič, Marie-Catherine de Marneffe, Maria Koptjevskaja Tamm

Syntax vs. Semantics: Issues and Objectives of the Discussions

In line with the topic and goal of the whole seminar, the presentations and followup discussions have concentrated on the design and properties of the syntax/semantics interface and its components in a cross/multilingual setting. The issues are not new, but with the existence and experience of massively multilingual resources at the morphological and syntactic levels (SIGMORPHON Shared Tasks (Coterell et al. 2020) and resources, in particular UniMorph (Syllak-Glassman, 2016) and the Universal Dependencies (Nivre et al. 2016, de Marneffe et al. 2021) syntactic annotation), it is natural that the focus now turns to semantics.

There are many semantic representations, projects and environments (and some were presented here as well), and some are venturing into multilingual annotation schemes and resources. However, there is no common framework such as the one for Universal Dependencies. It was thus natural to ask fundamental questions such as whether it is ever feasible (to have a common scheme), how to design it (if ever yet) to be useful and understandable for both linguists and technologists, but also for the barely initiated (such as Ph.D. students).

Semantic Representations, Grounding and Lexicons

Semantic representations have been tackled, looking at them "top down", from two points of view: annotation-based ones (e.g., the Prague Dependency Treebank (Hajic et al. 2020), UCCA (Abend et al., 2013) and others have been presented or mentioned) and the grammar-based ones (the Delph-in family, HPSG and MRS, ERG, etc.; see e.g (Bender et al., 2015, Copestake et al., 2005, Sag et al., 2003)). While the semantic representations based on an annotation scheme are available in more languages (albeit not many – see e.g. the MRP 2020 Shared Task (Oepen et al., 2020), where each formalism has been represented by just one additional language, besides English), the grammar-based approaches concentrate mostly on English – with some attempts to look at multilingual issues, such as at the Ling567 course at Univ. of Washington. This seems natural, as (hypothetically) unification of annotation guidelines seems to be simpler than building a grammar (necessarily?) different for each language, even if the resulting representation were uniform in their fundamental features.

However, the discussions brought up several interesting points where common schemes across languages might be possible. One example is grounding using language-neutral (or multilingual) ontologies – for example, DBpedia or Wikidata for organizations, people, locations of all sorts, domain ontologies etc. It has been mentioned that proper grounding might help to view things in context and possibly design a more uniform representation – for example, in the area of representing multiword entities (see also the WG2 report), representing synonyms, and dealing with ambiguity, both for single words (whatever a "word" is, see the WG1 report). Attempts have been mentioned to construct or convert existing verb-oriented lexical resources to an event type ontology, which is not well covered by current resources, such as WordNet or FrameNet. Examples have been shown, however, that every categorical scheme – whether grounding-based or more "lexically"-based – will necessarily have to solve several problems, such as granularity, style distinctions (how can they be grounded?), strength/positivity/negativity (and "scales" of all sorts).

Additionally, two other general and omnipresent issues have been mentioned: underspecification (and how to deal with it in the representation) and inferencing (how far to go in actually "interpreting" the utterance (and its "language") while constructing the representation. Relatedly, albeit not discussed very broadly, there was the issue of whether the semantic representation should in fact equal a "knowledge representation" (in the broader sense), or whether such a representation (either in the narrow interpretation of KR as a logic of some sort) will be still added on top of such semantic representation – but this is far away from the original "syntax vs semantics" topic of the WG, and probably deserving a seminar of its own.

Syntax-semantics Interface

While for grammar-engineering-based representations the relation to syntax is an inherent part of the grammar, it is much less clear how this interface is tackled in the annotation-based approaches. In some, there are layers which are linked together at a word (and sometimes subword) level; in some cases, at a MWE level (e.g., the Prague treebanks). In some others, there are no such links – AMR is a prime example, with others somewhere in between. There is a conceptual issue (how is it possible to connect syntax and semantics, represent this connection, at which level, etc.) and a representation (and format) issue (one schema for all – e.g., as in Enhanced UD, or two (or more) independent representations, with or without links). Several constraints are present here and have been discussed (see also the next paragraph about Cost of Access): fit to a certain background theory (of both syntax and semantics in terms of representation), boundary between the syntactic and semantic layers, harmonization across languages, simplicity of the representation(s), maintenance, error-proness, and many more. It is also important to define the purpose – ideally the same annotation scheme should serve theoretical and computational linguists (syntactitians, semanticists, typologists, grammarians, but also researchers in the fields of pragmatics, language acquisition, cognitive science, speech and language deficiencies, etc. etc.) as well as NLP/LT developers. Some specific issues have been presented in this area, e.g., harmonization between the UD and PARSEME annotation schemes, and a consistency of the UD scheme in relation to semantics.

Cost of Access and Maintenance

The design of any representation is inevitably an exercise in compromising between a cleanliness and the effort to be able to represent the real-world language in its full breadth (cf. the Manning Laws in Universal Dependencies). One of the requirements, namely the understandability of such a representation (without an extensive training "course") for, for example, students (both linguists and computer scientists, or technologists at software development companies, etc.) – i.e., the "cost of access". There is a related issue: given that building a large multilingual collection of semantically annotated resource is costly, and therefore only possible with a community effort, how difficult is it to understand the representation and guidelines to actually build an annotated corpus (or convert it from an existing one)? And what is the cost of its maintenance? It has been also mentioned – and agreed – that for such large community efforts to be successful, people must feel a sense of community, be able to get help and guidance, and to contribute not only data, but ideas and provide input and feedback for decisions at the top level.

Conclusions and Open Questions

It is hard to come to a definite conclusions at any seminar, the less so during a two-day online seminar which did not allow, for one, for a real plenary sessions due to time zone differences; it is hard even for a full-length typical face-to-face Dagstuhl meeting. However, in WG3, we believe we have at least identified some of the most pressing questions in the area of the syntax-semantics interface and semantic representation itself from the point of view of the long-term goal, namely a multilingual universal representation of the idiosyncrasies arising in human languages. While necessarily simplifying the contents of the discussions, and perhaps even missing some important topics which might have been mentioned only briefly (and despite having 18 pages of detailed notes taken in turns by the co-leads of this WG...), here is an attempt to summarize some of the points which are worth further investigation in this area:

- What are or should be the units of semantic representation (the "concepts"), where they come from, what is their granularity, how to ensure consistency especially in the universal (multilingual) setting
- Which approach is perhaps more suitable – the grammar-engineering approach vs. the "annotation" approach – what are the advantages of one against the other, or could they be merged or can the strength be combined somehow
- Where is the sweet spot between language-specificity (and adequate representation of the semantics of that particular language) vs. cross-lingual comparability using a simplified, but "universal" or "uniform" representation
- Should the syntax-semantic interface be captured, and how: in separate layers or one annotation scheme, how tight the "linking" between layers should be (whether implicit or explicit)
- What about redundancy in the representation(s) – some is probably inevitable, but how strong should the effort be to avoid it (across layers, within the semantic layer, in the linking or grounding)
- Issues of scaling or discretization/categorization: how fine/grained it should be, how to capture differences among language
- How to represent underspecification and (immediate) inferencing – should it be captured in the representation, and if yes, how far or deep
- Cost of access – for users, contributors, technologists: how to minimize this cost already in the design of such a representation, which compromises it entails

Finally, we would like to thank the organizers for the suggestion to hold this meeting, for their perseverance to organize it under the pandemic restrictions and all the difficulties it brought – it has been very dense two days, but with extremely inspiring presentations and discussions – both in the main talks and in the WGs themselves.

References

1 Abend, O., Rappoport, A. 2013. Universal Conceptual Cognitive Annotation (UCCA). In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). ACL 2013. Sofia, Bulgaria. https://aclanthology.org/P13-1023. Pp. 228-238.

2 Bender, Emily M., Dan Flickinger, Stephan Oepen, Woodley Packard and Ann Copestake. 2015. Layers of Interpretation: On Grammar and Compositionality. In Proceedings of the 11th International Conference on Computational Semantics (IWCS 2015), London. pp. 239-249.

3 Copestake, A., Flickinger, D., Pollard, C., Sag, I. 2005. Minimal recursion semantics: An introduction. Research on Language and Computation. 3(4). 281-332.

4 Flickinger, Dan, Stephan Oepen and Emily M. Bender. 2017. Sustainable Development and Refinement of Complex Linguistic Annotations at Scale. In Ide, Nancy and James Pustejovsky (eds), Handbook of Linguistic Annotation Science. Springer. pp. 353-377.

5 Hajič J. et al. 2020. Prague Dependency Treebank – Consolidated 1.0 (PDT-C 1.0). 2020. LINDAT/CLARIAH-CZ digital library. http://hdl.handle.net/11234/1-3185

6 Hajič J., Bejček E., Hlaváčová J., Mikulová M., Straka M., Štěpánek J., Štěpánková B. 2020. Prague Dependency Treebank – Consolidated 1.0. In: Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020), European Language Resources Association, Marseille, France, ISBN 979-10-95546-34-4, pp. 5208-5218.

7 Nicolai, G., Gorman, K., Cotterell, R. (Editors). 2020. Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology. ACL 2020.

8 de Marneffe, M.-C., Manning, C., Nivre, J., Zeman D. 2021. Universal Dependencies. In: Computational Linguistics, ISSN 1530-9312, vol. 47, no. 2, pp. 255-308.

9 Nivre, J., de Marneffe, M.-C., Ginter, F., Hajič, J., Manning, C., Pyysalo, S., Schuster, S., Tyers, F., Zeman, D. 2020. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020), pp. 4034-4043, European Language Resources Association, Marseille, France, ISBN 979-10-95546-34-4.

10 Oepen S., Abend O., Abzianidze L., Bos J., Hajič J., Hershcovich D., Li B., O'Gorman T., Xue N., Zeman D. 2020. MRP 2020: The Second Shared Task on Cross-Framework and Cross-Lingual Meaning Representation Parsing. In: Proceedings of the CoNLL 2020 Shared Task: Cross-Framework Meaning Representation Parsing, Copyright © Association for Computational Linguistics, Stroudsburg, PA, USA, ISBN 978-1-952148-64-4, pp. 1-22.

11 Sag, I. A.. Wasow, T., Bender, E. M. 2003. Syntactic Theory: a formal introduction, Second Edition. Chicago: University of Chicago Press.

12 Syllak-Glassman, J. 2016. The Composition and Use of the Universal Morphological Feature Schema (UniMorph Schema). Report, Working Draft v.2. CLSP. Johns Hopkins University.

5 Open problems

5.1 Harmonizing semantic representations in multilingual grammar engineering & otherwise

Emily M. Bender (University of Washington – Seattle, US)

Semantic harmonization refers to the process of designing semantic representations such that they are similar across languages, to the extent possible, while still staying true to the ".*" specific to each language. In that context, the following are points for discussion:

- What are the constraints that lead us to want to harmonize/what are the use cases for harmonized representations? Possible answers here include resources like the Grammar Matrix, Grammar-informed machine translation and downstream tasks that can handle different languages.

- What are the constraints that lead us to keep things language specific? Possible answers here include considerations of grammar design, considerations of annotation schema design and the annotation process itself.
- Finally, there are questions that can help situate the broader goal of harmonizing semantic representations and thus potentially help achieve it or at least approach it productively: At what point are we making these design decisions? Is it sensible to try to have one set of conventions? Across what domain (a multilingual project, multiple multilingual projects)? Who might benefit from such a set of conventions?

5.2 Idiosyncrasy and Derivational Morphology: From Distribution to Annotation

Cem Bozşahin (Middle East Technical University – Ankara, TR)

Joint work of Cem Bozşahin, Ozkan Aslan, Burcu Can, Arzu B Guven, Umut Ozge, Tzu-Ching Kao, Utku C Kunter, Gizem N Ozdemir, Barbaros Yet

Derivational morphology is commonly considered to be one morphological formant of the lexicon. Even in languages in which it is very productive, it is considered not as free as phrasal combination, and less compositional than inflection. It can become lexicalized, and idiosyncratic. For example, the Turkish word in (a) is formed by a very productive affix, *-lık*, with allomorphs *-lik, -lük, -luk*, but nowadays considered to be idiosyncratic; cf. its productive use in (b), and phrasal scope in (c), both of which are compositional. (DV: denominal verb; IRR: Irrealis; NN: noun to noun derivation.)

(1) a. *bakanlık*
bakan-lık
attend-NESS

‘ministry’ lit. bak-an-lık: see-REL-NESS Turkish

b. *gözlemlenebilirlik*
göz-lem-le-n-ebil-ir-lik
eye-NN-DV-PASS/REFL-ABIL-IRR-NESS
‘observability’

c.

[*her*	*konuda*	*resmi*	*görüş-lü*] *-lük*	*sana*	*yakışmadı.*
every	topic	official	view-with-NESS	you-DAT	fit-NEG-PAST

‘Officialese in every matter is not you.’

The question of unpredictability of combined meaning is pervasive crosslinguistically. Finnish is known for its agglutinating word structure. [4] report a case study in which Finnish children were asked about meaning of complex words, which is scored by a panel of linguists. Third graders scored better in low-frequency complex words if they happen to have highly productive affixes. Sixth graders were better in low-productive affixes if root frequencies were higher. It seems that idiosyncrasy in word structure needs time to settle in.

We can put this finding in a more general perspective in the microcosm of agglutination with results from Turkish. It has been known since [2] that Turkish children rarely go against adult’s ordering of affixes in a word. [3] has found that in a database of 18–36 month old children [10], frequency of affixes in child-speech are proportional to that in child-directed speech. Whether the child uses the derivational ones as conventional/idiosyncratic as the adult's is a question of interest.

In an effort to add adult performance to these findings, we have made two pilot studies on annotated Turkish corpora, reported in [5, 7]. Turkish has more than one hundred derivational affixes (see [6]/2014 for a full list). We chose the ten most frequent from the annotations, judged from annotation's selection of the analyses of word forms. We morphologically disambiguated a set of complex words with the ten derivational affixes. This step gives us correct stem forms of the chosen affixes.

In order to make maximal use of meaning annotations, we analyzed the verbal stems of the affixes to explore semantic properties of the verbal stem in annotation banks, to understand the affixs' selectivity in a stem (agent, patient, beneficiary etc. for thematic annotations such as Proposition Bank, kinds of scenes and participants in UCCA of [1], dependencies in UD frameworks). One such verb-to-verb derivation that we analyzed is (it also has deverbal noun interpretation):

(2)	gel	come	gel-iş	grow/develop	Turkish
	kaç	escape	kaç-ış	run away	
	sığ	fit	sığ-ış	accomodate	
	böl	divide	böl-üş	share	
	koş	run	koş-uş	scurry	
	kok	smell	kok-uş	rot	

As a first approximation we performed unsupervised clustering in the Proposition Bank of [8] and UD Bank of [9], using three methods: k-means, agglomerative clustering, and Gaussian mix. We then translated the verbal stems to English, in an effort to find more results in larger proposition banks of English, with the assumption that semantically the verb stem itself may have similar cross-linguistic conceptual ontology, theme, or dependency properties, although affixing to verb forms is Turkish-specific. We have also performed similar analyses on UCCA-annotated Turkish database.

We have observed no clustering of features. Manual dimension reduction techniques, for example conflating theme and experiencer, did not help.

In retrospect, it seems that trying to infer the meaning of complex words from annotations by unsupervised methods is too much to expect from labels, given current trends in annotation, in which clause and phrase meanings are the targets, and in approximation. Derivational meanings are more conventional than thematic or compositional, and annotators can hardly be expected to pay attention to subleties of conventional use, or have access to guidelines for their annotation.

One case in point is the convention of using "run' above in its derived forms. Consider a scenario where the schoolbell rings and kids run to their classes. As native speaker, I would not use "scurry' alone and say *çocuk-lar koşuş-tu* (kids scurried), because it would mean disorganized and unpurposeful action. I would use *çocuk-lar koş-uş-tur-du* (kids run-COLL-CAUS-PAST), to mean kids ran around for a reason, with the collective and causative imposing some kind of purposefulness. Notice that now the derived verbal stem is compositional, considered to be collective run.

One can hardly expect annotators attending to clausal meaning to distinguish the two uses of *koşuş*. It is also questionable whether guidelines can be set up for the task. It seems that we need an entirely different mechanism or sources to capture these differences manifesting themselves as conventions, that is, higher-order uses of grammatical elements, which are idiosyncratic references to events.

References

1 Abend, O. and A. Rappoport (2013). Universal conceptual cognitive annotation (UCCA). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 228–238.

2 Aksu-Koc, A. A. and D. I. Slobin (1985). The acquisition of Turkish. In D. I. Slobin (Ed.), *The Crosslinguistic Study of Language Acquisition, vol.I: The Data*. New Jersey: Lawrence Erlbaum.

3 Avcu, E. (2014). Nouns-first, verbs-first and computationally easier first: A preliminary design to test the order of acquisition. Master's thesis, Cognitive Science department, Middle East Technical University (ODTÜ), Ankara.

4 Bertram, R., M. Laine, and M. M. Virkkala (2000). The role of derivational morphology in vocabulary acquisition: Get by with a little help from my morpheme friends. *Scandinavian Journal of Psychology 41*(4), 287–296.

5 Kunter, U. C., G. N. Özdemir, and C. Bozşahin (2020). Distributional and lexical exploration of semantics of derivational morphology. In *Proc. of Int. Symp. on Brain and Cognitive Science, ISBCS 2020*, Ankara.

6 Oflazer, K., E. Göçmen, and C. Bozşahin (1994). An outline of Turkish morphology. Technical report, METU and Bilkent Univ. re-issued in 2014.

7 Özdemir, G. N. (2021). Distributional investigation of some frequent Turkish derivational affixes for exploring their semantics. Master's thesis, Middle East Technical University. Cognitive Science Dept., Ankara.

8 Şahin, G. G. and E. Adalı (2018). Annotation of semantic roles for the Turkish proposition bank. *Language Resources and Evaluation 52*(3), 673–706.

9 Türk, U., F. Atmaca, Ş. Özateş, G. Berk, A. Köksal, and A. Özgür (2020). Resources for Turkish dependency parsing: Introducing the BOUN treebank and the BoAT annotation tool. *Arxiv preprint 2002.10416*.

10 Ural, A. E., D. Yuret, F. N. Ketrez, D. Koçbaş, and A. C. Küntay (2009). Morphological cues vs. number of nominals in learning verb types in Turkish: The syntactic bootstrapping mechanism revisited. *Language and Cognitive Processes 24*(10), 1393–1405.

5.3 How to best account for the semantics of MWEs?

Voula Giouli (Athena Research Center, GR)

Joint work of Voula Giouli, Aggeliki Fotopoulou
Main reference Aggeliki Fotopoulou, Voula Giouli: "MWEs and the Emotion Lexicon: Typological and cross-lingual considerations", Language Science Press, 2018.
URL http://dx.doi.org/10.5281/zenodo.1182591

The identification of MWEs involves lexical, morphosyntactic and semantic criteria (Gross 1982; 1998b; Lamiroy 2003), to be taken into account, namely: *non-compositionality*, i.e., the meaning of the expression cannot be computed from the meanings of its constituents; *non-substitutability*, i.e., at least one of the expression constituents does not enter in alternations at the paradigmatic axis; and *non-modifiability*, in that they enter in syntactically rigid structures, posing further constraints over modification, transformations, etc. However, the criteria mentioned do not apply in all cases in a uniform way, and the variability attested brings about the notion of *degree of fixedness* (Gross 1996). In this regard, idiomatic expressions bear a meaning that cannot be computed based on the meaning of their constituents and the rules used to combine them. Light verb constructions (LVCs), on the other hand, have a rather transparent meaning due to the presence of the predicative noun (Npred) which retains its original sense.

The problem posed can be defined as follows:

(a) What is the best way to represent the semantics of MWEs – more precisely VIDS (verbal idiomatic expressions) and LVCs – in a uniform way?

(b) One step further, the limits between LVCs and VIDs are in some cases fuzzy: despite the semantic transparency in LVCs (entailed by the Npred) the overall structure is often susceptible to a number of constraints as shown in the example, and the overall semantics of the final expression is less transparent:

(1) πετάω από χαρά

petao	*apo*	*chara*
fly.1SG	from	happiness.ACC.SG

“to be very happy”

What are the criteria to account for these fuzzy cases?

(c) Is mapping of a MWE to a concept sufficient for representing its sense efficiently? esp. when a (near-)synonymous single-word expression exists? For example, how to account for the VID in (2) and its near-synonymous single-word verb?

(2) κάνω σκόνηά

kano	*skoni*
make.1SG	dust.ACC.SG

“to defeat thoroughly”

(d) What other types of semantic representation are feasible, i.e, semantic role labelling? And how to account for a sound annotation of MWEs? What are the best practices for assigning semantic roles to syntactic constituents (non-fixed elements) of MWEs (syntax-semantics interface), esp. with regard to cross-linguistic issues?

(3) Τον τρώει η ζήλιαά

Ton	*troi*	*i*	*zilia*
Him.3.SG.ACC.EXPERIENCER	eats.3.SG.NM	the.SG.NOM	jealousy.SG.NOM

“He is very jealous.”

References

1 Gross, Maurice. 1982. Une classification des phrases “figées”du français. *Revue Québécoise de Linguistique (RQL)* 11(2). 151–185.

2 Gross, Maurice. 1998a. La fonction sémantique des verbes supports. *Travaux de linguistique* 37. 25–46.

3 Gross, Maurice. 1998b. Les limites de la phrase figée. *Language* 90. 7–23.

4 Lamiroy, Béatrice. 2003. Les notions linguistiques de figement et de contrainte. *Lingvisticae Investigationes* 26(1). 1–14.

5.4 MWE Identification using Embedding-based Approaches

Tunga Güngör (Bogaziçi University – Istanbul, TR)

Joint work of Zeynep Yirmibeşoğlu, Tunga Güngör
Main reference Zeynep Yirmibesoglu, Tunga Gungor: "ERMI at PARSEME Shared Task 2020: Embedding-Rich Multiword Expression Identification", in Proc. of the Joint Workshop on Multiword Expressions and Electronic Lexicons, MWE-LEX@COLING 2020, online, December 13, 2020, pp. 130–135, Association for Computational Linguistics, 2020.
URL https://aclanthology.org/2020.mwe-1.17/

Rather than being directly related to the annotation of multiword expressions (MWE), this talk is on the computational side of processing of MWEs. We present a general schema for automatically identifying MWEs using embedding-based approaches. Normally, in all computational models based on the deep learning paradigm, the input is represented in terms of embeddings. An embedding is a short vector (a vector of dimensions typically between 100 and 500) that is used as a representation of a particular entity (e.g. a word). These embedding-based approaches can be used in a multilingual context.

We have used three embedding-based models in previous research for MWE identification in the scope of the PARSEME shared task (Ramisch, et al. (2020)). The basic model is named as ERMI (embedding-rich MWE identification). Two of the models are supervised, while one of them is a semi-supervised model. In that research, as the deep learning model, LSTM-CRF network was used. The details of the model are not important for this talk; other neural network models can also be used. We focus on the embeddings in these models. In the first model, the input for each word is formed of the concatenation of three types of information: the word itself, its part-of-speech, and its dependency relation to the head word in the UD treebank (Nivre, et al. (2016)). Each of these three parts is formed of embeddings. In this way, for a word, the input consists of both morphological and syntactic information. In the second model, a fourth component is added to the input, which is the head word of this word (i.e. its embedding). All these embeddings are learned from a corpus for a language. These two models use annotated corpus. The third model is different in the sense that there is an additional raw (not annotated) corpus, which is much larger than the annotated corpus. A MWE identification model is trained as in the other models, then this model is used to annotate the raw corpus. Then this much larger annotated corpus is used for learning a MWE identification model.

To conclude: These models can be used as multilingual NLP tools in computationally tractable ways. We have tested such tools in about 15 languages having different typologies. And they showed promising results in detecting MWEs.

The open question in this research is: What should be good input representations in terms of embeddings for different types of languages?

References

1 Ramisch, C., Savary, A., Guillaume, B., Waszczuk, J., Candito, M., Vaidya, A., Mititelu, V.B., Bhatia, A., Inurrieta, U., Giouli, V., Gungor, T., Jiang, M., Lichte, T., Liebeskind, C., Monti, J., Ramisch, R., Stymne, S., Walsh, A. and Xu, H., Edition 1.2 of the PARSEME Shared Task on Semi-supervised Identification of Verbal Multiword Expressions, Joint Workshop on Multiword Expressions and Electronic Lexicons (MWE-LEX 2020) at COLING 2020, Barcelona, p.107-118, December 2020.

2 Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D., Universal Dependencies v1: A Multilingual Treebank Collection, Proceedings of International Conference on Language Resources and Evaluation (LREC 2016), European Language Resources Association (ELRA), Portoroz, Slovenia, p.1659–1666, 2016.

5.5 Harmonizing Semantic Representations

Jan Hajič (Charles University – Prague, CZ) and Daniel Zeman (Charles University – Prague, CZ)

Joint work of Stephan Oepen, Omri Abend, Lasha Abzianidze, Johan Bos, Jan Hajič, Daniel Hershcovich, Bin Li, Tim O'Gorman, Nianwen Xue, Daniel Zeman
Main reference Stephan Oepen, Omri Abend, Lasha Abzianidze, Johan Bos, Jan Hajic, Daniel Hershcovich, Bin Li, Tim O'Gorman, Nianwen Xue, Daniel Zeman: "MRP 2020: The Second Shared Task on Cross-Framework and Cross-Lingual Meaning Representation Parsing", in Proc. of the CoNLL 2020 Shared Task: Cross-Framework Meaning Representation Parsing, CoNLL Shared Task 2020, Online, November 19-20, 2020, pp. 1–22, Association for Computational Linguistics, 2020.
URL http://dx.doi.org/10.18653/v1/2020.conll-shared.1

While there are very good examples of harmonized representations on lexical, morphological and syntactic levels (UniMorph, Universal Dependencies), work is still ongoing on various semantic representations (AMR/UMR, Prague Dependency Treebanks, UCCA, DMR, DRT, RMS, PMB, DM, ...), as exemplified, for example in the MRP Shared Tasks in 2019 and 2020. If ever possible, such a harmonization should work across formalisms and across languages. However, even the term "semantics" is understood wildly differently by the authors of the various existing representations – from just slightly above syntax (PDT, Enhanced UD) all the way to logic (e.g., PMB) or knowledge representations and ontologies. The related presentation(s) (see the WG3 slides and talks) look at it from different points of view. Perhaps a similar effort could be launched, like UD, for semantic representation and annotation...?

5.6 VMWE degree modification: somewhere between "compositional" and "idiomatic"

Stella Markantonatou (Athena Research Center, GR)

This discussion is about issues regarding the implications of the phenomenon of degree modification (achieved with modifiers) for the lexicographic codification and UD annotation of VMWEs. Degree modifiers may apply on the verb head (1a) or on a lexicalized phrase (fixed subject (1b), object (1c), complement of the copula).

(1) a. *δάγκωσα* ***για τα καλά/γϵρά/άσχημα*** *τη λαμαρίνα*
I.bit for good/strongly/ugly.**ADVERB** the tin

"I fell for somebody"

b. *θα σου πιουν* ***όλο*** *το αίμα*
will you.GEN they.drink all.**DET** the blood

"they will exhaust you;;

c. *βγήκϵ* ***μϵγάλη*** *βρώμα πριν λίγο στο Τρωκτικό για Σάββα*
came.out big.**ADJECTIVE** dirt before little in.the Rodent for Savvas

"very nasty news about Savvas has just been published in the "Rodent""

The distribution of adverbs such as κυριολεκτικά (“literally”), πραγματικά, πράγματι (“really”), ειλικρινά (“sincerely”) and PPs such as στην κυριολεξία (“literally”), στ′ αλήθεια (“truly”) seems to be determined by discourse only. These adverbs are used to confirm the true of the denotation of the utterance that contains the VMWE; it is precisely the speaker’s commitment about his utterance that creates the intensification effect (Mexa and Markantonatou 2020; Israel 2002; Paradis 2003; Bordet 2017).

The adverbs εντελώς, τελείως (“completely”) apply to VMWEs that are closed scale predicates (Kennedy and McNally 2005; Gavriilidou and Giannakidou 2016; Mexa and Markantonatou 2020). Manner adverbs are used for degree modification of VMWEs (and verbs) (Κλαίρης και Μπαμπινιώτης 2004: 857): άγρια (“wildly”), άσχημα (“ugly”), για τα καλά (“for good”), γερά, δυνατά (“strongly”), etc. Mexa and Markantonatou (2020) found them with VMWEs from the semantic domains of ANGER (2) and LOVE but not of SURPRISE. They possibly form collocations with certain VMWEs.

(2) *αρχίζω και φορτώνω πολύ/άσχημα/ϵπικίνδυνα/;;γϵρά*
I.start and I.load ugly/dangerously/??strongly

“I am getting dangerously angry”

Modification of a lexicalized NP is described below. Certain adjectives, such as ανήμερος (“untamed”) (3), τσουχτερός (“biting”) seem to form **collocations**.

(3) *έγινϵ θηρίο ανήμϵρο*
he.became beast untamed

“he got furious”

Μεγάλος (“large”, “big”), τεράστιος (“huge”) (4a), (4b) have less constrained distribution; other adjectives have a messy distribution probably determined by the ““literal” meaning of the noun. The definite and the indefinite article, the determiners όλος (“all”) (1b), πολύς (“much”, “a lot”), and the conjunction και (“and”) can be used as intensifiers:

(4) a. *έφαγα μεγάλη/τεράστια/τρϵλλή/άγρια/*βαριά φρίκη*
I.ate large/huge/mad/wild/*heavy horror

“I went through a horrendous experience”

b. *μου ήρθϵ μϵγάλη/τϵράστια/βαριά/χοντρή/;;τρϵλλή κϵραμίδα*
to.me came large/huge/heavy/fat/?mad rooftile.NOM

“I experienced a big unpleasant surprise”

In sum: (i) Modification by adverbs of the type of “literally” is not determined by the semantics of the VMWE (ii). Certain, but not all, adjective+noun combinations strongly collocate (iii). In between are several adverb+verb, adjective+noun combinations.

A. What should be encoded in a lexicon?

- A collocation/combination as an independent VMWE?
- A collocation/combination as a variation of the “original” VMWE that could be derived from it via modification with (very often) prespecified adjectives/adverbs? Consider the θηρίο (“beast”) set of VMWEs:

(5) a. *ϵ́γινα ϑηρίο*
became.1.SG beast

“I got very angry”

b. *ϵ́γινα άγριο/σωστό/πραγματικό ϑηρίο*
became.1.SG wild/right/real beast

“I got very, very angry”

c. *ϵ́γινα ϑηρίο ανήμϵρο*
became.1.SG beast untamed

“I got furious”

B. Degree modification in UDs

- “Degree modification using modifiers (not using morphology)” ... but it is reminiscent of degree modification via morphology, e.g.:

(6) a. *χϵ́ρι μικρό χϵ́ρι/χϵρ-άκι*
hand

b. *μϵγάλο χϵ́ρι/χϵρ-ούκλα*
large hand/hand-magnifier

UD at the moment have only one morphology feature Dim(inutive) and only for Afrikaans[5] that allows to connect the lemma of a noun e.g. χέρι with a diminutive form e.g. χεράκι. Probably something similar could be defined for the “beast”-set (5) and its kin.

References

1 Bordet, Lucile. “From vogue words to lexicalized intensifying words: the renewal and recycling of intensifiers in English. A case-study of very, really, so and totally.” Lexis. Journal in English Lexicology 10 (2017).

2 Ernst, Thomas. “Grist for the linguistic mill: Idioms and “extra’adjectives.” Journal of Linguistic Research 1.3 (1981): 51-68.

3 Gavriilidou, Zoe, and Anastasia Giannakidou. “Degree modification and manner adverbs: Greek: poli “very” vs. kala “well’.” Selected papers on theoretical and applied linguistics 21 (2016): 93-104.

4 Israel, Michael. “Literally speaking.” Journal of Pragmatics 34.4 (2002): 423-432.

5 Kennedy, Christopher, and Louise McNally. “Scale structure, degree modification, and the semantics of gradable predicates.” Language (2005): 345-381.

6 Mexa, M., and S. Markantonatou. “Intensifiers/moderators of verbal multiword expressions in Modern Greek.” EURALEX XIX (2021).

7 Paradis, Carita. “Between epistemic modality and degree: the case of really.” Modality in contemporary English. De Gruyter Mouton, 2012. 191-222.

8 Κλάιρης, Χρήστος, ανδ Γεώργιος Δ. Μπαμπινιώτης. Γραμματική της νέας ελληνικής: δομολειτουργική-επικοινωνιακή. Το όνομα της νέας Ελληνικής. Ι. ὅλ. 1. Ελληνικά Γράμματα, 1996.

[5] https://universaldependencies.org/af/feat/Degree.html

5.7 Projects are humans – on the trade-off between complexity and diversity

Carlos Ramisch (Aix-Marseille University, FR)

Initiatives to create highly multilingual resources for morphological, syntactic and semantic processing abound nowadays in the computational linguistics community. Among these initiatives, three are represented in this seminar: UniMorph for morphology, Universal Dependencies (henceforth UD) for dependency syntax, and PARSEME for multiword expressions. These projects deal with the representation, especially in annotated corpora, of complex linguistic phenomena. To model these phenomena, one reasonable assumption is that one should use different layers to group phenomena according to their similarity into a single layer whereas phenomena that seem too distant are pushed to an upper/lower layer.

One of the most concrete examples of this approach is UD's CoNLL-U format. An annotated corpus file in CoNLL-U format contains 10 columns which can be seen as more or less independent layers representing the segmentation, form, lemma, POS, morphological features, dependencies etc. of naturally occurring text. This mechanism can be generalised further, as in the "CoNLL-U plus" format[6], which was adapted by PARSEME to add an 11th column representing MWEs in the CUPT format.[7]

Layers are very useful because they allow focusing on a single subproblem of language representation at each time inside an ambitious annotation project such as UD. In addition, different projects can then deal with different groups of phenomena, such as UniMorph, UD and PARSEME, thus manipulating autonomous layers in parallel without conflicts. Different layers can use different underlying structures to represent language: subword morphological features, word-level POS tags, dependency graphs, MWE subsequences (or sub-graphs), etc. Layers of different levels can be connected using unique IDs and references. This helps accounting for the complexity of language in a modular way that is also very convenient for computational processing.

On the other hand, a layered approach to language annotation and representation poses two major challenges. First, it is hard to define completely hermetic layers, especially in the light of cross-layer phenomena such as multiword expressions, which challenge the traditional borders between lexicon, syntax and semantics (or even morphology, e.g. idiomatic compounds). Therefore, different initiatives may decide to represent a single phenomenon in a language in different ways. For instance, while UD sees verb-particle constructions in English (e.g. *make up*) as a type of compound, PARSEME rather models the non-compositional nature of the particle modifier, providing different tests and scope to annotate the same phenomenon. This introduces redundancy when we bring layers together: many verb-particle constructions will be annotated twice, once in UD and once in PARSEME, with sometimes inconsistent decisions.

While consistency and redundancy are well known technical issues in large annotation projects, there is a second problem that arises from the complexity of multi-layered approaches. Suppose a new language wants to join UD and PARSEME, and that both projects are now completely integrated, with MWE annotation representing an extra layer over UD's morpho-

[6] https://universaldependencies.org/ext-format.html
[7] http://multiword.sourceforge.net/cupt-format/

syntactic layers. Even if all consistency and redundancy issues were solved, annotating all these 11 layers at once would still be extremely complex, especially for new annotators. Now suppose that not only MWEs but also finer morphological annotation is included as an extra layer (e.g. to introduce morpheme-based tags). Besides, more abstract semantic and pragmatic layers can be added such as abstract meaning representation semantics trees, named entities, terms, and so on, each on a separate layer. Reading, understanding and becoming familiar with the whole annotation guides of all these layers might sound like a scary task, so the **access cost** to this new integrated annotation project would increase.

The open issue at hand here could be summarised in the following question: *how can we account for the complexity of language in our multi-layered projects without increasing the access cost for new languages beyond what would be acceptable?* In other words, given the "universal" nature of these projects, we must keep in mind that the diversity of languages covered is a crucial aspect. Resource creation and enhancement cannot be limited to a small group of initiated project members who have been there for enough time to master the maze of annotation layers and guides.

To decompose the question into more focused ones, we can think about several aspects that can make a resource creation project attractive. First, the benefits for contributors must be made visible from the beginning. Most of these projects cannot fund their members, so the benefits are rather indirect. They include co-authoring papers on the topic, networking, improving the resources for their own languages, and being supported by a large international initiative in their local grant applications.[8]

Second, the management of the community must be well designed. It may require some structure, some specific roles, so that new members know who to contact when they get lost. New members should be able to get quick assistance, so that they do not feel discouraged by the weight of all the layers.

Third, collaborative projects usually require a deep sense of community. Members have to feel engaged, belonging to something pleasant, interesting, dynamic. For instance, how can we ensure that the discussions carried out on git issues are taken into account into the guidelines? Should the guidelines point to the discussions (e.g. git issues) that led to given a decision? How can we make guidelines evolve so that experts feel listened to, without requiring prohibitive updates to existing data? Creating and keeping this feeling of belonging, especially in the context of the current pandemic crisis, is a real challenge for these communities.

Fourth, a certain number of tools and resources can make integration of a new language less traumatic for newcomers. These include online forums, chat platforms, shared software infrastructure, tutorials, videos, zoom calls with more experienced members, and probably many more tools that we still have to imagine and develop. While free tools can be used in many cases, developing specific tools tailored to our needs can also make a difference, although it is not always easy to request funding for software development and engineering in research grant applications.

In short, the question of how to integrate (many) annotation layers in resources created by different communities to account for the complexity of linguistic phenomena poses real challenges for (computational) linguistics. In addition to the well known problems of redundancy and consistency, it is important to put some effort into keeping these projects welcoming and inclusive so that the **access cost** for new languages is kept to a reasonable minimum, helping keep diversity at the core of these initiatives. This provides us with a

[8] See an example from PARSEME: [9]

great opportunity to think out of the box and employ our creative energy to come up with innovative solutions that make people feel committed and engaged into connecting their resources and putting their linguistic expertise at the service of cross-linguistic connections and deeply multilingual computational applications.

5.8 Multiword expressions as multiword constructions

Manfred Sailer (Goethe-Universität Frankfurt am Main, DE)

Joint work of Manfred Sailer, Sascha Bargmann

The formal modelling of idioms has been alternating between phrasal/holistic and lexical/combinatorial approaches. At least in HPSG and SBCG, a combinatorial analysis seems to have been widely recognized recently. I think that, notwithstanding the important arguments in favor of a combinatorial modelling, the unit-like character of an MWE should be accounted for as well. I see two main flaws of combinatorial analyses: First, they fail to capture the fact that the MWE-specific reading of the components of the complex expression should not be represented outside the MWE. Second, the literal meaning of MWE components is accessible for metaphoric and other processes even in non-decomposable MWEs.

I will sketch an attempt of a solution to this dilemma, which may at least work for HPSG, though it is an open question what this would mean for other frameworks, for corpus annotation, or parsing.

Summary of the presentation

The formal modelling of idioms has been alternating between phrasal/holistic and lexical/combinatorial approaches. At least in HPSG and SBCG, a combinatorial analysis seems to have been widely recognized for all syntactically regular MWEs recently. I think that, notwithstanding the important arguments in favor of a combinatorial modelling, the unit-like character of an MWE should be accounted for as well.

Combinatorial modellings are very well equipped to capture the fact that MWEs of the same degree of decomposability differ with respect to their syntactic flexibility across languages. This observation has been made already in Nunberg et al. 1994, and more systematically in Schenk 1995. Bargmann & Sailer 2018 show that it can follow directly from a parallel specification of the lexical entries of the parts of the MWEs (*such as kick the bucket* 'die' and its German analogue *den Löffel abgeben* (lit: the spoon away.give)), combined with language-specific characterizations of the fronting constructions. In this approach, the constraints on the syntactic flexibility of MWEs follows from the interaction of the lexical entries of their component parts and the analysis of the critical syntactic constellations (passive, fronting, etc).

In so-called phrasal approaches, MWEs are encoded as phrasal units, which means that their "lexical" description contains both information on their component words/morphemes and on their syntactic combination (such as VP for kick the bucket). Such phrasal approaches face problems:

The discourse conditions can be so special that even usually syntactically non-flexible MWEs may appear in a particular constellation, such as non-decomposable MWEs in English passive:

(1) When you are dead, you don't have to worry about death anymore. ... The bucket will be kicked.
(internet example, reported in Bargmann & Sailer 2018:5)

There is interaction with other special constructions, such as the N-after-N construction:

(2) All those people behind them pulling string after string for them
(internet example, reported in Bargmann 2019: chapter 6)

One part of an MWE may be associated with several occurrences of the MWE:

(3) The beans have not been spilled yet, but will be spilled very soon.
(constructed, reported in Sailer & Bargmann 2021)

Parts of an MWE can be pronominalized:

(4) Eventually she spilled all the beans. But it took her a few days to spill them all.
(Riehemann 2001:207)

Webelhuth, et al. 2018 and Bargmann 2019 show how these data can be captured in an approach that reduces the description of an MWE to the description of its component words/morphemes, ignoring their concrete phrasal combination.

However, existing lexical approaches fail to capture the fact that the MWE-specific reading of the components of the complex expression should not be represented outside the MWE (see the criticism in Riehemann 2001). For example, some phenomena seem to link the literal and the MWE-specific (or idiomatic) reading of an MWE. Egan 2008 and Findlay, et al. 2019 discuss so-called extended uses of MWEs as in (5).

(5) If you let this cat out of the bag, a lot of people are going to get scratched.
(Egan 2008: 392)

Here, the MWE *let the cat out of the bag* is interpreted idiomatically with respect to the current world in the if-clause. To make sense of the main clause, we need to interpret the MWE literally, but with respect to some figurative or metaphoric world. Finally, the overall interpretation needs to be mapped back to current world by some analogy between the figurative world and the current world (i.e. revealed secret can hurt many people, just as a released cat can scratch many people). For such a reasoning to be available, we need (i) access to both the literal and the idiomatic interpretation of the MWE, (ii) access to the MWE as a unit.

Ernst 1981 and Bargmann, et al. 2021 show that so-called conjunct modification as in (6) is another instance which requires simultaneous availability of the literal and the idiomatic meaning of an MWE.

(6) With the recession, oil companies are having to tighten their Gucci belts.
(Ernst 1981:60)

In the talk, I sketched Sailer & Bargmann 2021, which treates MWEs as multiword constructions, i.e. as constructions that specify words they contain, but no concrete phrasal syntactic pattern. I suggested a way to expand this to integrate the analysis of data as (5) from Findlay, et al. 2019.

Summary of the discussion

The discussion brought up a number of interesting issues.

First, the notion of a construction was debated. If the current approach is on the right track, we should not only have constructions as complex phrasal patterns that may specify some lexical components, but we should also assume cases where we have fixed lexical components but no pre-determined phrasal pattern.

Second, the treatment in Sailer & Bargmann 2021 is in part similar to the IOB-encoding proposed in Schneider, et al. 2014, but may differ in cases like (3).

Third, the extent to which we find purely idiomatic uses of an MWE, purely literal uses, or combined uses such as those in (5) and (6) is an interesting topic that needs further investigation.

References

1 Bargmann, S. 2019. Chopping up idioms: Towards a combinatorial analysis. Goethe-University Frankfurt a.M. dissertation.

2 Bargmann, Sascha, Berit Gehrke, and Frank Richter. "Modification of literal meanings in semantically non-decomposable idioms." One-to-many relations in morphology, syntax, and semantics (2021): 245.

3 Bargmann, Sascha, and Manfred Sailer. "The syntactic flexibility of semantically non-decomposable idioms." Multiword expressions: Insights from a multi-lingual perspective 1 (2018): 1-29.

4 Egan, Andy. "Pretense for the complete idiom." Noûs 42.3 (2008): 381-409.

5 Ernst, Thomas. "Grist for the linguistic mill: Idioms and "extra'adjectives." Journal of Linguistic Research 1.3 (1981): 51-68.

6 Findlay, Jamie Y., et al. "Why the butterflies in your stomach can have big wings: combining formal and cognitive theories to explain productive extensions of idioms." Talk given at the EUROPHRAS 2019 Productive Patterns in Phraseology Conference. Vol. 24. 2019.

7 Nunberg, Geoffrey, Ivan A. Sag, and Thomas Wasow. "Idioms." Language 70.3 (1994): 491-538.

8 Riehemann, Susanne Zalta. A constructional approach to idioms and word formation. stanford university, 2001.

9 Sailer, M. & S. Bargmann. 2021. A phraseo-combinatorialanalysis of idioms. Talk presented at HPSG 21.

10 Schenk, André. "The syntactic behavior of idioms." Idioms: Structural and psychological perspectives (1995): 253-272.

11 Webelhuth, Gert, Sascha Bargmann, and Christopher Götze. "Idioms as evidence for the proper analysis of relative clauses." Reconstruction effects in relative clauses. De Gruyter (A), 2018. 225-262.

5.9 Word and resources for little-resourced languages

Emmanuel Schang (University of Orleans, FR)

The definition (its boundaries and the method to discover it) of word has been identified as a difficult topic for a very long time. These difficulties can be found in the Cours de Linguistique Générale (Saussure 1915):

> "En résumé la langue ne se présente pas comme un ensemble de signes délimités d'avance, dont il suffirait d'étudier les significatiosnet l'agencement ; c'est une masse indistincte où l'attention et l'habitude peuvent seules nous faire trouver des éléments particuliers. L'unité n'a aucun caractère phonique spécial, et la seule définition qu'on puisse en donner est la suivante : une tranche de sonorité qui est, à l'exclusion de ce qui précède et de ce qui suit dans la chaîne parlée, le signifiant d'un concept. [...] Cependant nous sommes mis immédiatement en défiance en constatant qu'on s'est beaucoup disputé sur la nature du mot, et en y réfléchissant un peu, on voit que ce qu'on entend par là est incompatible avec notre notion d'unité concrète.

De Saussure concludes in an optimistic way and gets around the problem of the definition of word by saying that it is not a necessary unit:

> "Lorsqu'une science ne présente pas d'unités concrètes immédiatement reconnaissables, c'est qu'elles n'y sont pas essentielles. En histoire, par exemple, est-ce l'individu, l'époque, la nation? On ne sait, mais qu'importe? On peut faire oeuvre historique sans être clair sur ce point."
> [...]
> "La langue présente donc ce caractère étrange et frappant de ne pas offrir d'entités perceptibles de prime abord, sans qu'on puisse douter cependant qu'elles existent et que c'est leur jeu qui la constitue. C'est là sans doute un trait qui la distingue de toutes les autres institutions sémiologiques."

More than a century later, Haspelmath (2017) concludes "that we do not currently have a good basis for dividing the domain of morphosyntax into morphology and syntax, and that linguists should be very careful with general claims that make crucial reference to a cross-linguistic "word' notion."

Having this in mind, any annotation of corpus based on the notion of word has to be taken with caution.

This is true for well-known languages, but crucial for little-known languages. In absence of a standardized writing system, the linguist makes theoretical choices which frequently conflict with the speakers "intuitions' and practice. This has been well described in Hazaël-Massieux (1993) for the Antillean Creoles for instance.

This becomes a real problem when building a treebank for little-known languages since the resources are rare and expensive, and the opportunities to recode it (new segmentation etc.) is difficult and costful. The necessities of a particular project often leads to a particular coding which too often blocks the reuse of the resource.

One can wish that a coding of a resource is not destructive and that an alternative coding could be considered.

References

1 Haspelmath, M. (2017). The indeterminacy of word segmentation and the nature of morphology and syntax. Folia linguistica, 51(s1000), 31-80.

2 Hazaël-Massieux, M. C. (1993). écrire en créole: oralité et écriture aux Antilles. éditions l'Harmattan.

3 Saussure, F. De (1915). Cours de linguistique generale (Payot, Paris).

5.10 Listen to the data: comprehensive MWE annotation and emergent challenges in English UD

Nathan Schneider

Overview

STREUSLE Corpus:

- Lexical semantic annotation of MWEs in English reviews
- Bottom-up, comprehensive: no preconceived notion of which categories/types we were looking for; original annotators did not see syntax:
  - Lots of variety! Not just verbal and nominal MWEs – also PPs, functional expressions, etc.
  - Also discovered some partially productive constructions in this process
- Strong (semantically opaque) vs. weak expressions (̃formulaic expressions, statistically idiomatic)
- Lexcat (lexical category): syntactic subcategorization of strong MWEs (and single-word expressions); adapted from UD UPOS; draws on PARSEME for VMWEs

UD issues investigated in English: better guidelines needed for

- "Mischievous nominal constructions": names, dates, numbers; compounds; adverbial NPs
- Multiword connectives ("out of", "along with", "based on", etc.)

Details and Links

- **STREUSLE:** MWEs can be annotated comprehensively ina corpus, without prefiltering for syntactic status, which unearths all sorts of interesting expressions and constructions. (Schneider et al. 2014)
  - PARSEME efforts thus far have done a great job for **verbal** MWEs across languages, but MWEs in general are syntactically open-ended.
  - STREUSLE annotators received general criteria for what counts as an MWE: **strong** (semantically opaque) or **weak** (formulaic expression). Developed guidelines for specific constructions as we went[10].

[10] see: original guidelines, prepositional verbs, PARSEME 1.1 VMWEs

        - Investigation was limited to English. Can this be done on a larger scale and multilingually?
        - Suggestion: for a corpus sample, annotate MWEs bottom-up, without predefined syntactic constraints. Then revise, taxonomize, and harmonize.
        - Description of data format (CONLLU-Lex), **lexcats** to sub-categorize strong expressions
    - Are there annotation tools that make it easy to alternate between consecutive and type-based annotation flows?
    - Tagger trained on STREUSLE, evaluated on MWE corpora
- UD guidelines need clarification/improvement with respect to various productive nominal constructions in English: **“mischievous nominal constructions”**:
    - names, dates, numbers, measurements; adverbial NPs; compounds
    - Better accommodate these with current top-level relations, clarifying boundaries of flat, compound, appos, nmod, nummod, etc. See proposals with Amir Zeldes.
- UD guidelines around **multiword functional connectives** need improvement
    - UD annotators need (at least a small) construction!
        - See: current version.
    - Double case analysis for “out of” etc. Why? See: Issue #795.
    - Deverbal connectives: “according to”, “based on”, etc. See: EWT issue #179.
        - Validator considers VERB/mark an error.
    - Conjunction-like connectives: “rather than”, “instead of”, “along with”, etc. See: Issue #679.
    - “Next to”. See: Issue #496.
- UD needs a way to annotate idioms where the internal head POS does not correspond to the phrase’s syntactic distribution, e.g. see `ExtPos=Yes` and other ideas in Issue #807.
- UD unclear on complex determiners: “a few”, “a little” (see EWT Issue #170), and in general whether “few” and “many” should be ADJ or DET (see Issue #786)
- Syntax vs. semantics: UD is a sufficiently established standard that we should clarify morphosyntactically tricky aspects of constructions, but keep semantics in a separate layer

5.11 Unifying UD and PARSEME frameworks

Sara Stymne (Uppsala University, SE)

Universal dependencies (UD) is a framework for consistent annotation of morphological features, aprt-of-speech tags and dependency syntax, across different human languages. Version 2.8 covers 114 languages. For each annotation layer, UD contains a pool of categories that languages can use, and it is also possible to add language specific extensions. For more information, see the UD web page[11].

PARSEME (PARSing and Multi-word Expressions), started out as a EU COST action (2013–2017), but the initiative remains. One of the PARSEME activities is to organize shared tasks on the identification of verbal multiword expressions (VMWEs). There has been three

[11] https://universaldependencies.org/

editions in 2017, 2018 and 2020. As part of the shared task, a large annotation effort has taken place for 26 languages (in edition 1.2). PARSEME has general guidlines targeted at all languages, with language specific extensions in a few cases.

Both these initiatives target harmoized annotation of language phenomena across languages, but with differnt phenomena in focus. PARSEME encouraged participants to annotate texts from UD, in order to have basic annotations as well, but there are also other texts annotated in the PARSEME corpora.

The main points for discussion proposed were:

- Is it desirable to unify UD and PARSEME?
- How should it be done technically?
  - The PARSEME CUPT format is an extension of the UD ConLLU format
  - Potentially PARSEME anntoations can be added to UD in a similar way as extended universal dependencies.
- What are the potential relations between the two projects?
- What can the projects learn from each other?

The maybe most important outcome of the discussion in WG3 was that it was seen as desirable to synchronize UD and PARSEME annotations. There seem to be advantages to having resources in the same location, especially when they are annotated on the same texts. While the PARSEME annotations are more semantic than UD, this was not seen as a major issue. There are also “extended universal dependencies” for some UD treebanks, which cover semantic aspects.

We noted that there are some overlap in annotations, especially as language-specific features in UD. Particles are attached to their main verbs with the edge label “compund:prt”, and similarly the label “compound:lvc” is used for light verbs. However, these labels are only used for a small number of languages. While a language like English do have LVCs, such as “make a decision”, it is not annotated as such in UD, but seen as a regular syntactic construction. We discussed that language-specific constructions will likley mainly be used in languages where such constructions are pervasive. This can be problametic for instance for corpus-studies, when the phenomenon is not represtend equally across languages. It is also the case that some syntactic constructions, like particle verbs, can be used idiomatically, but there are also cases where the semantics are regular, and they would not be annotated as VMWEs in PARSEME.

5.12 Are Chinese idioms Multi-Word Expressions (MWEs)?

Nianwen Xue (Brandeis University – Waltham, US)

Chinese has many (typically four-character) idioms that are based on some ancient stories. Overtime the moral of a story has become the meaning of the idiom. The question of how to identify MWEs are intricately linked to the question of wordhood. One key test for MWEs is non-compositionality, which is also a key test for wordhood in Chinese. An idiom is by definition non-compositional. Does that mean that all idioms are words (thus no longer MWEs)? Or are cases where idioms can still be MWEs? Answers to these questions are crucial to arriving at a definition of MWEs that are cross-linguistic applicable.

PARSEME definitions of Multiword Expressions (MWEs):

- Some degree of orthographic, morphological, syntactic or semantic idiosyncrasy with respect to what is considered general grammar rules of a language.
- Their component words include a head word and at least one other syntactically related word. Most often the relation they maintain is a syntactic (direct or indirect) dependence but it can also be e.g. a coordination.
- At least two components of such a word sequence have to be "lexicalized (fixed)" (others are "open slots").
- How to recognize MWEs: Probably the most salient property of MWEs is semantic non-compositionality, but since non-compositionality is subjective, use inflexibility as proxy.

Chinese has many (typically four-character) idioms that are based on some ancient stories. Over time the moral of a story has become the meaning of the idiom:

(1) 请司空摘星拿主意,无异于缘木求鱼、刻舟求剑毫无 可操作性。
ask

"Asking Sikongzhaixing for ideas is no different from climbing trees to catch fish or carve a mark on boat to find the missing sword, and is not at all practical."

Other idioms are metaphors that can be very long:

(2) 我哑巴吃黄连有苦说不出。
I.aphastic

"I feel like an aphasic who has taken coptis Chinesis, and cannot speak out even though I am wronged."

In other cases apparent MWEs may be just discontinuous words:

(3) 我摔了一个大 跤。
I.fall

"I had a big fall / I fell hard."

Questions for discussion:

- Are Chinese idioms multi-word expressions? To answer that, we need to know how many words these idioms have.
- How what counts as a word in Chinese, where text is not written with orthographic word boundaries?
- For languages like Chinese where there are no natural word boundaries, tests for MWEs need to be built on tests for wordhood.
- Wordhood in Chinese is a complicated issue, and there are many factors involved: morphophonological, syntactic, semantic, lexical, rhythmic, etc.
- Any definitions or classifications of MWEs that are cross-linguistically applicable would have to take into account languages that do not have orthographical word boundaries, and the determination of MWEs cannot be easily separated from the determination of wordhood.

- Coming up with consistent criteria to identify MWEs helps with consistent annotation of syntactic (e.g., UD) and semantic representations (e.g., AMR/UMR).
- Thinking about the syntactic / semantic structure can sometimes crytalize our judgments of MWEs.

5.13 Issues in UD Consistency, Phrases, and Semantics – With a Focus on Double Subjects and Nested Copula Predication

Amir Zeldes (Georgetown University – Washington, DC, US)

In this talk I give an overview of Universal Dependencies-related activities at Georgetown University, with a focus on annotation problems encountered while annotating the UD English Georgetown University Multilayer corpus (GUM) and some new UD data in Hebrew. I focus on the problematic guideline advocating the annotation of nesting copulas with a top-level clause headed by the copula and governing the nested clause as a complement clause (ccomp). I argue that this guideline is linguistically wrong, as it implies that verbs like “be’ are transitive; that it is inconsistent, since other nesting functional structures (nesting PPs, adverbial clauses) do not behave this way; that it is incoherent, since it gives radically different sub-graph analyses to different cases of the same construction; and that it is cross-linguistically untenable due to languages with zero copula constructions, which may nest in the same way.

5.14 Wordhood issues: Toward a typology

Tim Zingler (University of New Mexico – Albuquerque, US)

Main reference Zingler, Tim: “Wordhood issues: Typology and grammaticalization”. PhD dissertation, University of New Mexico, 2020
URL https://digitalrepository.unm.edu/ling_etds/71/

Researchers in both phonology and morphosyntax agree that the structural unit of the “word” is difficult to define on the basis of concrete formal criteria (e.g., Bickel et al. 2009; Haspelmath 2011). What is of particular interest is that this issue manifests itself in both language-specific and cross-linguistic work (e.g., Schiering et al. 2010; Tallman 2021). One response to this impasse has been to posit two different word units, a phonological (or prosodic) word and a grammatical (or morphological, or morphosyntactic) word (cf. Dixon 2010: ch. 10). In addition, it has become a convention to classify as “clitics” all those elements that show some kind of mismatch between phonological and grammatical wordhood criteria (cf. Haspelmath & Sims 2010: 198, 202). Yet, since phonological and grammatical words, as well as the mismatches within and between them, are language-specific and variegated, such a coarse classification does not shed much light on the problem or on its causes (cf. also Tallman 2020). The net result of this situation is thus a paradox. The majority of linguists would arguably agree that words are an important “building block” of language, and words have psychological reality even for users of traditionally unwritten languages (e.g., Evans 1995: 62; Mithun 2014: 73). Yet, every effort to define words seems to suggest that they do not in fact exist.

One way to accommodate this conundrum is to approach it from the opposite angle. That is, once it is acknowledged that mismatches between wordhood criteria (henceforth "wordhood issues") are inevitably found in the world's languages, it becomes a major desideratum to typologize these wordhood issues. To the extent that these mismatches tend to involve some combinations of wordhood criteria more often than others, this would suggest that these common mismatches would have to be treated in more depth by linguistic theories than the less common, language-specific outliers. Arriving at such a typology of wordhood issues was the central aim of Zingler (2020), which looked at exponents of definiteness, case, indexation ("agreement"), and tense across 60 unrelated languages from five geographical macro-areas. This focus on the grammatical domain was itself motivated by two major cross-linguistic insights. One the one hand, words tend to lose both their phonological and morphological word properties during the process of grammaticalization (e.g., Bybee et al. 1994; Hopper & Traugott 2003). On the other hand, the four functions selected bear a relatively low degree of relevance to the meanings of their respective nominal or verbal stems, which typically coincides with a lower degree of formal fusion between the stem and the grammatical marker (cf. Bybee 1985). In conjunction, these findings suggest that markers of those four functions are particularly likely to be formally ambiguous, that is, to constitute wordhood issues.

The methodological basis of Zingler (2020) was a set of four criteria of phonological wordhood and four criteria of grammatical wordhood. These eight criteria are displayed in Table 1 and were drawn from the relevant literature because they are sufficiently general to be applicable to a large number of different languages. Furthermore, these criteria were subsumed under the more general concept of "formal dependence," which was in turn defined as the degree to which the shape or distribution of a morpheme is determined by other morphemes. Hence, an element that falls short of any given criterion of wordhood is "dependent" in terms of that criterion. It also follows that an element that is dependent in terms of more criteria of phonological wordhood than of grammatical wordhood is more "prosodically" dependent than "syntagmatically" dependent, abbreviated here as "P > S." The converse, in which a morpheme is dependent on more criteria of grammatical wordhood than of phonological wordhood, is rendered as "S > P" in this work. One major benefit of this approach is that it does not require the vague and unhelpful "clitic" label. Rather, it permits the classification of any given element as a P > S or S > P issue, which then leaves open the possibility to further specify which criteria underlie the relevant mismatch.

The main finding of Zingler (2020) was that P > S issues are considerably more common than S > P issues. The 72 wordhood issues in the database come from 41 of the 60 languages in the sample, which suggests that there might be languages without wordhood issues in the grammatical domain. It is also important to highlight that the predominance of P > S issues holds in each of the four grammatical domains and in each of the five macro-areas investigated. So, the typological generalization seems to be that more prosodic dependence is the norm among grammatical exponents and that more syntagmatic dependence is the exception. Meanwhile, the grammaticalization perspective of this distribution is that the emergence of prosodic dependence typically precedes that of syntagmatic dependence, which can largely be explained by frequency-driven accounts of phonological change such as those by Bybee (2001, 2015). Finally, the S > P issues exclusively occur in contexts of high morphological synthesis, where the relevant exponent will often be syntagmatically fixed before it is fully prosodically reduced. This further suggests that wordhood issues are systematic. Yet, future research will have to establish why synthesis is a necessary rather than a sufficient condition for S > P issues.

While the quantitative discrepancy between P > S issues and S > P issues already helps to constrain the problem of wordhood, a further empirical fact that emerges from Zingler (2020) is of even greater interest. Specifically, 54 of the 63 P > S issues involve an item that has the syntactic distribution of a word (and thus meets the wordhood criterion of non-selectivity) but that falls short of phonological wordhood because it is integrated into a larger domain in terms of prominence and/or allomorphy (and thus violates the wordhood criteria of prosodic features and/or phonological rules). Overall, then, 54 of the 72 wordhood issues are defined by a specific interaction of only three of the eight wordhood criteria investigated. In order to account for this pattern, one might thus define a word as a single domain of prominence and allomorphy in which all non-root constituents are selective. However, it should be emphasized that this definition is no more than a heuristic to be used in cross-linguistic research on phenomena that require some definition of wordhood. Yet, while this definition obviously fails to account for notions such as vowel harmony, this omission derives precisely from the fact that vowel harmony proved to be a marginal indicator of wordhood in most of the languages sampled. Hence, the question of how to treat such language-specific wordhood issues will still have to be left for specialists working on the languages at issue.

Table 1 Wordhood criteria used in Zingler (2020).

phonological word	*grammatical word*
Free occurrence: Word constitutes a well-formed utterance	Cohesiveness: The constituent elements of a word always occur together
Segmental structure: Word is in the domain of phonotactic constraints, minimum weight	Conventionalized meaning: Word is the smallest psychologically real sign for users
Prosodic features: Word is the domain of stress / tone assignment, vowel harmony	Fixed order: The relative position of each morphological unit within a word is fixed
Phonological rules: Word is in the domain of phonologically conditioned allomorphy	Non-selectivity: Words can co-occur with different word classes (unlike affixes)

Table 2 Number and distribution of wordhood issues in Zingler (2020).

Dependence	*Definiteness*	*Case*	*Indexation*	*Tense*	*Total*
P>S	6	28	15	14	63
S>P	2	1	4	2	9

References

1 Bickel, Balthasar, Kristine Hildebrandt & René Schiering. 2009. The distribution of phonological word domains: A probabilistic typology. In Grijzenhout, Janet & Barış, Kabak (eds.), Phonological domains, 47-75. Berlin: De Gruyter.

2 Bybee, Joan. 1985. Morphology. Amsterdam: Benjamins.

3 Bybee, Joan. 2001. Phonology and language use. Cambridge: Cambridge University Press.

4 Bybee, Joan. 2015. Language change. Cambridge: Cambridge University Press.

5 Bybee, Joan, Revere Perkins & William Pagliuca. 1994. The evolution of grammar. Chicago: The University of Chicago Press.

6 Dixon, R. M. W. 2010. Basic linguistic theory, vol. 2. Oxford: Oxford University Press.

7 Dixon, R. M. W. & Alexandra Aikhenvald. 2003. Word: A typological framework. In Dixon & Aikhenvald (eds.), Word, 1-41. Cambridge: Cambridge University Press.

8 Evans, Nicholas. 1995. A grammar of Kayardild. Berlin: De Gruyter.

9 Haspelmath, Martin. 2011. The indeterminacy of word segmentation and the nature of morphology and syntax. Folia Linguistica 45, 31-80.

10 Haspelmath, Martin & Andrea Sims. 2010. Understanding morphology. 2nd ed. London: Hodder.

11 Hopper, Paul & Elizabeth Traugott. 2003. Grammaticalization. 2nd ed. Cambridge: Cambridge University Press.

12 Mithun, Marianne. 2014. Morphology: What's in a word? In Genetti, Carol (ed.), How languages work, 71-99. Cambridge: Cambridge University Press.

13 Schiering, René, Balthasar Bickel & Kristine Hildebrandt. 2010. The prosodic word is not universal but emergent. Journal of Linguistics 46, 657-709.

14 Tallman, Adam. 2020. Beyond grammatical and phonological words. Language and Linguistics Compass 14(2), e12364, 1-14.

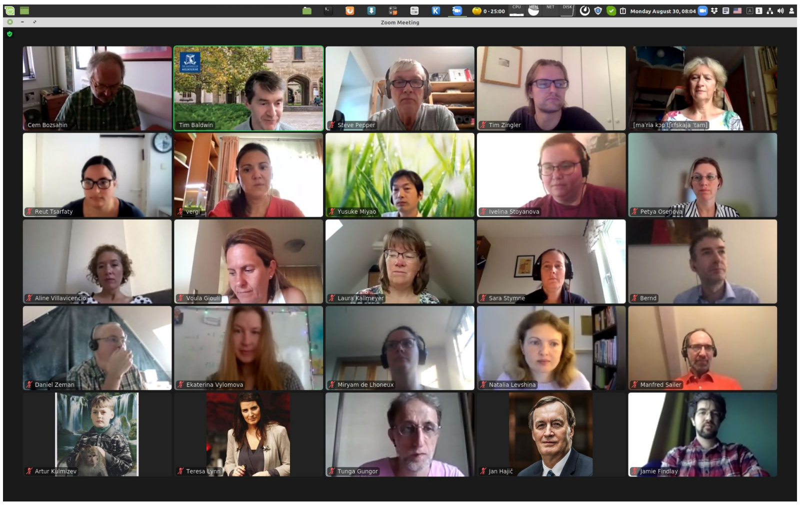
15 Tallman, Adam. 2021. Constituency and coincidence in Chácobo (Pano). Studies in Language 45, 321-383.

16 Zingler, Tim. 2020. Wordhood issues: Typology and grammaticalization. PhD dissertation, University of New Mexico.

Remote Participants

- Timothy Baldwin
The University of Melbourne, AU
- Verginica Barbu Mititelu
Research Institute for A.I. – Bucharest, RO
- Emily M. Bender
University of Washington – Seattle, US
- Archna Bhatia
Florida IHMC – Ocala, US
- Bernd Bohnet
Google – Amsterdam, NL
- Francis Bond
Nanyang TU – Singapore, SG
- Cem Bozsahin
Middle East Technical University – Ankara, TR
- Ryan Cotterell
ETH Zürich, CH
- William Croft
University of New Mexico – Alburquerque, US
- Miryam de Lhoneux
University of Copenhagen, DK
- Marie-Catherine de Marneffe
Ohio State University – Columbus, US
- Jamie Findlay
University of Oslo, NO
- Daniel Flickinger
Stanford University, US
- Kim Gerdes
University Paris-Saclay – Orsay, FR
- Voula Giouli
Athena Research Center, GR
- Tunga Gungor
Bogaziçi University – Istanbul, TR
- Jan Hajic
Charles University – Prague, CZ
- Dag Haug
University of Oslo, NO
- Uxoa Iñurrieta
Donostia, ES
- Laura Kallmeyer
Universität Düsseldorf, DE
- Christo Kirov
Google – New York, US
- Maria Koptjevskaja Tamm
Stockholm University, SE
- Artur Kulmizev
Uppsala University, SE
- Lori Levin
Carnegie Mellon University – Pittsburgh, US
- Natalia Levshina
Max-Planck-Institute for Psycholinguistics – Nijmegen, NL
- Teresa Lynn
Dublin City University, IE
- Stella Markantonatou
Athena Research Center, GR
- Nurit Melnik
The Open University of Israel – Raanana, IL
- Paola Merlo
University of Geneva, CH
- Yusuke Miyao
University of Tokyo, JP
- Kadri Muischnek
University of Tartu, EE
- Joakim Nivre
Uppsala University, SE
- Petya Osenova
Bulgarian Academy of Sciences – Sofia, BG
- Stephen Pepper
University of Oslo, NO
- James Pustejovsky
Brandeis University – Waltham, US
- Alexandre Rademaker
IBM Research – Sao Paulo, BR
- Carlos Ramisch
Aix-Marseille University, FR
- Manfred Sailer
Goethe-Universität Frankfurt am Main, DE
- Agata Savary
Université de Tours – Blois, FR
- Emmanuel Schang
University of Orleans, FR
- Nathan Schneider
Georgetown University – Washington, DC, US
- Ivelina Stoyanova
Bulgarian Academy of Sciences – Sofia, BG
- Sara Stymne
Uppsala University, SE
- Reut Tsarfaty
Bar-Ilan University – Ramat Gan, IL
- Francis M. Tyers
Indiana University – Bloomington, US
- Meagan Vigus
University of New Mexico – Alburquerque, US
- Aline Villavicencio
University of Sheffield, GB
- Veronika Vincze
University of Szeged, HU
- Ekaterina Vylomova
The University of Melbourne, AU
- 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
- Tim Zingler
University of New Mexico – Alburquerque, US

Joakim Nivre


Zoom Meeting
Monday August 30, 08:04
Cem Bozsahin
Tim Baldwin
Steve Pepper
Tim Zingler
[maˈria kɔpˈt͡ʃɛfskaja ˈtam]
Reut Tsarfaty
vergi
Yusuke Miyao
Ivelina Stoyanova
Petya Osenova
Aline Villavicencio
Voula Giouli
Laura Kallmeyer
Sara Stymne
Bernd
Daniel Zeman
Ekaterina Vylomova
Miryam de Lhoneux
Natalia Levshina
Manfred Sailer
Artur Kulmizev
Teresa Lynn
Tunga Gungor
Jan Hajič
Jamie Findlay

Zoom Meeting
ceramisch
Stella Markantonatou
Joakim Nivre
Agata Savary
Cem Bozsahin
Tim Baldwin
Steve Pepper
Tim Zingler
[maˈria kɔpˈt͡ʃɛfskaja ˈtam]
Reut Tsarfaty
vergi
Yusuke Miyao
Ivelina Stoyanova
Petya Osenova
Aline Villavicencio
Laura Kallmeyer
Sara Stymne
Bernd
Daniel Zeman
Ekaterina Vylomova
Miryam de Lhoneux
Natalia Levshina
Manfred Sailer
Artur Kulmizev
Teresa Lynn