# CoBaLD Annotation: the Enrichment of the Enhanced Universal Dependencies with the Semantical Pattern

Lingotto Conference Centre - Torino (Italia)

Maria Petrova, Alexandra Ivoylova, Anastasia Tishchenkova

LREC-COLING 2024

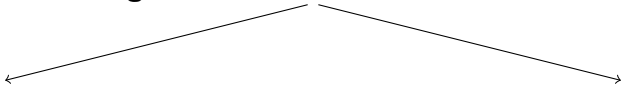
May 2024

## Introduction

**Purposes of the work:**

- to enrich the Enhanced Universal Dependencies (E-UD) annotation with a semantic pattern
- to develop a markup format supporting morphological, syntactic and semantic levels

**Compreno-Based Linguistic Data Annotation, or CoBaLD Annotation**



UD annotation schema
[De Marneffe et al., 2021]

Compreno semantics
[Anisimovich et al., 2012]

# Introduction

Linguistic markup is used for:

- theoretical purposes
- various downstream NLP tasks (e.g., for enriching language model embeddings in such tasks as sentiment analysis [Baly et al., 2017], metaphor detection [Li et al., 2023], or cross-lingual transfer [Ponti et al., 2018])

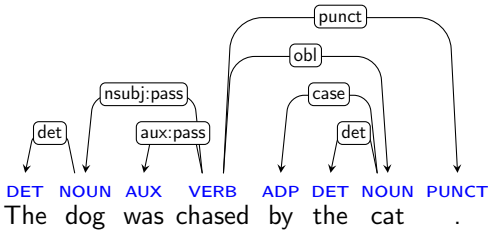
# Introduction

Practical applicability of linguistic annotation depends on its

- simplicity
- fullness
- accuracy
- suitability for machine processing

# Introduction

## Universal Dependencies




- Convenient for automatic processing ⇒ is widely used for practical tasks
- Surpasses other standards in popularity
- Has treebanks in many languages
- Annotates morphology and dependencies
- Aims to be linguistically universal
- Lacks semantics

# Introduction

Semantic annotation standards:

- Universal Networking Language (UNL) [Uchida and Zhu, 2001]
- Abstract Meaning Representations (AMR) [Banarescu et al., 2013]
- Prague Tectogrammatical Graphs (PTG) [Hajic et al., 2001]
- Universal Conceptual Cognitive Annotation (UCCA) [Abend and Rappoport, 2013]
- Discourse Representation Structures (DRS) [Bos et al., 2017]
- Universal Decompositional Semantics (UDS) [White et al., 2016]
- Compreno [Anisimovich et al., 2012]

## Introduction

**Requirements for the semantic annotation model**

- compatibility with UD
- universality
- simplicity
- full semantic description
  - all word meanings
  - all relations between words

## Compreno

Advantages:

1. Compreno and UD have similar token labeling principles, so integrating the Compreno semantics into the UD annotation is a feasible task
2. Universality - aimed at the description of any language
3. Simplicity - very simple categorial system, containing only two category sets: deep slots (DSs) for semantic relations and semantic classes (SCs) for word meanings
4. Fullness
   - SCs are organized in the form of a thesaurus-like tree, where all meanings of each word are stored
   - every level gets all possible DSs for each SC, which provides full description of all possible semantic dependencies including actants, adjuncts, modifiers, and so on

## Compreno

Disadvantages:

1. Compreno presents morphosyntactic information in the form of parsing trees – not in the markup itself, which makes its usage more inconvenient in comparison with the UD presentation, and, most important, depends on the parser's work
2. The number of the SCs and the DSs is too big due to the detailness of the description which seems excessive for most applicational tasks

## Integration

To overcome these problems, we have:

- converted the Compreno morphosyntax to UD
- used the hyperonym SCs to reduce the number of the SCs
- used the generalized version of the DSs to reduce the number of the DSs

# Compreno semantics

**Compreno word meanings**

Lexical meanings are presented in the form of so called Semantic Classes (SCs).

The whole hierarchy includes more than 200,000 universal SCs.

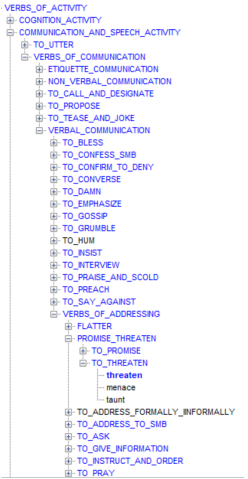
In CoBaLD, we use the shortened variant which consists of the hyperonym SCs:

## SC reduction

*CAT, TORTOISE, ELEPHANT* → *ANIMAL*

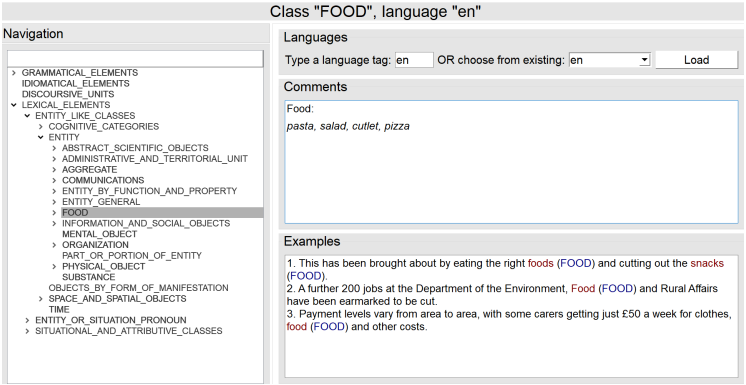
| | |
|---|---|
| FULL: | The players "PLAYER_OF_GAMES" ran "TO_RUN" hard "INTENSITY_OF_ACTIVITY" all the time "TIME". |
| SHORT: | The players "HUMAN" ran "MOTION" hard "CH_OF_INTENSITY" all the time "TIME". |

For CoBaLD, the hyperonym hierarchy includes about 650 classes and is available on Github.



# Compreno semantics

## Hierarchy of hyperonym SCs



## Compreno semantics

**Word meanings in the hyperonym hierarchy**

Hyperonym hierarchy is usually enough to differentiate between most homonyms, but there are still cases where generalisation leads to the loss of distinction between closely related homonyms:

### Homonym distinction

*pour* as 'flow in the stream' ⇒⇐ *pour* as 'to rain hard'
'The river poured into the sea' ⇒⇐'The rain is just pouring down'
*pour* as 'flow in the stream' ⇒⇐ *pour* as 'to move in large amounts or numbers'
'The river poured into the sea' ⇒⇐'The people poured along the street'

The feedback in this respect is very important and will help us to define the optimal detailness level of the hierarchy.

## Compreno semantics

### Relations between words

In Compreno there are:

- syntactic roles = surface slots: language-specific, surface relations
- semantic roles = deep (semantic) slots, DSs: universal, deep relations

| Deep Slots | Surface Slot |
|---|---|
| (a) I talked [to Peter] – Addressee;<br>(b) The rule refers only [to children] – Object;<br>(c) The tune [to the song] – Purpose;<br>(d) His reply [to a question] – Stimulus. | Object_Indirect_To |

## Compreno semantics

### Deep Slots

DSs in Compreno are not only actant dependencies, but all dependencies a word can attach.

Each DS can be filled with a strict set of SCs.

Most dependencies get both a surface and a semantic role. The exceptions are grammatical dependencies (articles or prepositions, for instance), and idiomatical ones (like *beans* in 'spill the beans').

## Relations between words

Surface slots do not strictly depend on semantic roles ⇒ can correspond to several types of UD relations, e.g. *iobj* or *obl*.

Besides, verbal and nominal cores attach the same surface slot:

(a) The house [on the hill];

(b) He stood [on the hill].

# E-UD vs Compreno

**Differences between E-UD and Compreno**

| Model | E-UD | Compreno |
|---|---|---|
| Have semantic dependencies | no | yes |
| Have syntactic dependencies | yes | yes |
| Actant and circumstantial dependencies with similar surface realizations get different syntactic roles | yes | no |
| Syntactic dependencies with similar surface realizations which depend on nominal vs verbal cores get different syntactic roles | yes | no |

## Generalized deep slots

The inconvenience is that the full list of the Compreno DSs is more than 300.

While it may be beneficial for some (mostly theoretical) tasks, such as semantic sketch creation [Ponomareva et al., 2021], for most practical purposes it is too detailed.

$\Rightarrow$ we joined all characteristic slots together, parentheticals, specifications, and united a number of slots with the same semantics but different filling.

The shortened number of the DSs is 143.

# Generalized deep slots

## An example of a generalized deep slot

| Time DSs in Compreno | Time examples | Time DS in CoBaLD |
|---|---|---|
| Time | He came [yesterday/at 5 o'clock]. | Time |
| Time_Being | [post-Bush] economy | |
| Time_Entity | [After a sandwich and a pint], we headed to Trinity College. | |
| Time_Situation | [When the war started], nobody believed it. | |
| Time_Source | someone [from 1860] | |

## Corpora

### CoBaLD Rus

- In 2023 we proposed a 400,000 token news corpus for Russian in our standard, but using 'base' UD
- The applicability of the format was tested during the SEMarkup-2023 Shared Task [Petrova et al., 2023]
- The baseline parser using ruBERT-tiny embeddings achieved around 90% F1-score for both SCs and DSs

## Corpora

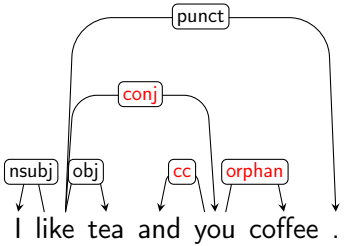
### CoBaLD Eng

Now we propose a new 150,000 token corpus for English available on Github.

- The resource for texts - the BBC dataset ([Greene and Cunningham, 2006]), divided into five topics (business, entertainment, politics, sport and tech)
- We switched to CONLL-Plus file standard
- We provide E-UD annotation as well as base UD

## UD modifications

In the standard version of UD, functional words may never be heads, and only surface realisations are marked up, thus sometimes strange annotations may appear:



## UD modifications

### SUD vs Enhanced UD

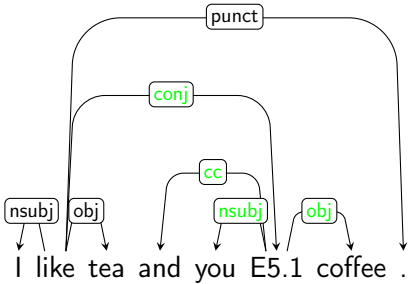
As the 'base' version of UD is not purely syntactical, other annotation standards were proposed:

- Surface Universal Dependencies [Gerdes et al., 2018]: it is based on the theory of I. Melčuk [Mel'cuk et al., 1988], treats functional words as heads, including copula
- Enhanced Universal Dependencies [Schuster and Manning, 2016]: an enhancement of 'base' UD version, restores ellipsis and adds several more elaborate dependency types

## UD modifications

### Enhanced Universal Dependencies

E-UD standard [Schuster and Manning, 2016] is a popular development: there have been two shared tasks [Bouma et al., 2020, Bouma et al., 2021] devoted to it. While retaining all of the UD features, E-UD adds a new, more detailed and syntax-oriented layer of annotation.



## Corpora

### Corpus creation

The process of creation included several steps:

1. The chosen texts were automatically annotated with the Compreno rule-based parser;
2. The annotation was checked by professional linguists;
3. Syntax level was converted automatically to E-UD standard, then to UD;
4. Morphology level was also converted automatically;
5. Semantic labels were processed according to our modifications and applied to the CONLL-Plus format in columns 11-12.

# Corpora

## CoBaLD Annotation

The resulting annotation looks like follows:

```
# text = The full economic costs of the disaster remain unclear.
1   The the DET Article Definite=Def|PronType=Art   4   det 4:det    _   _    ARTICLES
2   full    full    ADJ Adjective   Degree=Pos  4   amod    4:amod  _   Characteristic  CH_SPHERE_OF_COVERAGE
3   economic    economic    ADJ Adjective   Degree=Pos  4   amod    4:amod  _   Sphere  ECONOMY
4   costs   cost    NOUN    Noun    Number=Plur 8   nsubj   8:nsubj _   Object  MONEY
5   of  of  ADP Preposition _   7   case    7:case  _   _   PREPOSITION
6   the the DET Article Definite=Def|PronType=Art   7   det 7:det    _   _    ARTICLES
7   disaster    disaster    NOUN    Noun    Number=Sing 4   nmod    4:nmod:of   _   Object_Situation    BAD_DANGEROUS_EVENT
8   remain  remain  VERB    Verb    Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin   0   root    0:root  _   Predicate   BE
9   unclear unclear ADJ Adjective   Degree=Pos  8   xcomp   8:xcomp SpaceAfter=No   State   CH_PERCEPTIBILITY
10  .   .   PUNCT   PUNCT   _   8   punct   _   _   _   _
```

column 11 = DSs, column 12 = SCs

## Inter-annotator agreement

Inter-annotator agreement for semantics level is high and the inconsistencies are usually caused by homonymy.

| | | Full markup, % | Simplified markup, % | | | Full markup, % | Simplified markup, % |
|---|---|---|---|---|---|---|---|
| Tech | SemSlot | 97.12 | 97.36 | Politics | SemSlot | 98.75 | 98.8 |
| | SemClass | 97.7 | 97.78 | | Semclass | 98.34 | 98.65 |
| | Heads | 98.29 | 98.29 | | Heads | 99.58 | 99.58 |
| | Overall | 93.10 | 93.44 | | Overall | 96.67 | 97.04 |

## Conversion

### Morphology

In morphology, the asymmetry between the models concerns the following areas:

- tokenization,
- lemmatization,
- POS-tagging,
- defining the sets of grammatical features.

The description of the Russian converter is given in [Ivoylova et al., 2023]. In the current version, we optimised the converter and adapted it for English.

## Conversion

### Morphology

There were several differences in Compreno and UD standard we had to resolve (in favor of UD):

- Tokenization: Compreno splits texts into tokens accordingly to semantics, so we had to re-tokenize some cases like complex prepositions and such ('as to'). In other cases we had to merge tokens.
- Lemmatization: in Compreno, proper nouns sometimes get a special lemma instead of the word itself, and in Russian there was a difference for verb aspect, otherwise the conversion didn't need any additional operations.

## Conversion

### Morphology

- POS-tagging: several Compreno POS-tags do not correspond fully to UD; e.g., Compreno does not distinguish 'Auxiliary' and 'Verb' POS-tags. Such POS-tags were converted with the help of morphological features and syntax information. A special list was created to convert a Compreno 'Invariable' tag which can correspond to multiple UD tags.
- Grammatical features: we sometimes had to resort to syntax information, and there was also a problem with gerunds: we decided to follow the UD instruction in case of their definition despite the fact that it seems to contradict both Compreno principles and English grammar rules.

## Conversion

### Syntax

The conversion of relations had several challenges:

- One of most notorious cases is the UD distinction between *nmod* and *obj*, *iobj*, *obl* tags which are determined by the POS-tag of their head.
- There were some problems with the conversion of direct speech as Compreno may mark it up differently.
- Compreno tends to analyze quoted expressions by restoring an empty node, so that tokens inside quotes may get a special 'Internal' syntactic slot which makes it extremely difficult to convert their relations to UD.

## Conversion

### Syntax

Concerning head conversion, there were also several problems we had to solve:

- Compreno differently treats several cases as with ‘according to’, ‘including’ and some others. We had to switch heads in order to comply with the UD standard.
- Another crucial difference lies in the fact that Compreno builds constituency trees instead of dependency ones; so in cases of inverted word order there may be problems for conversion.
- In the case of quoted expressions, we had to remove empty nodes as they don’t correspond to E-UD rules of ellipsis restoration; otherwise, there are other differences in ellipsis restoration which have to be inverstigated yet.

## Conversion

### Quality

In the course of our work, we found out that it would be more practical to have human annotators (linguists) check the results of automatic conversion and fix it according to their findings; this is an iterative process. So far, we assess the quality of our morphosyntactic conversion as below:

| Lemma | POS | Feats | Heads | Deprel | E-UD | Overall |
|---|---|---|---|---|---|---|
| 98.68 | 95.48 | 94.92 | 95.48 | 96.80 | 93.97 | 85.12 |

# Conclusion

- We propose a new annotation standard based (and highly compatible) with E-UD and UD, which includes semantic level of markup;
- We make a 150,000 token English dataset in our standard available and publish it on Github;
- We continue improving our conversion process and we enlarge CoBaLD Eng;
- We are developing a DL-based parser for our updated format;
- We are annotating several small datasets in other languages such as Serbian and Hungarian;
- We may add a Compreno syntactic annotation layer as well.

Thank you for your attention!

# References I

[Abend and Rappoport, 2013] Abend, O. and 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)*, pages 228–238.

[Anisimovich et al., 2012] Anisimovich, K., Druzhkin, K. Y., Zuev, K., Minlos, F., Petrova, M., and Selegei, V. (2012).
Syntactic and semantic parser based on abbyy compreno linguistic technologies.
In *Computational Linguistic and Intellectual Technologies*, pages 91–103.

[Baly et al., 2017] Baly, R., Hajj, H., Habash, N., Shaban, K. B., and El-Hajj, W. (2017).
A sentiment treebank and morphologically enriched recursive deep models for effective sentiment analysis in arabic.
*ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 16(4):1–21.

[Banarescu et al., 2013] Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., and Schneider, N. (2013).
Abstract meaning representation for sembanking.
In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, pages 178–186.

[Bos et al., 2017] Bos, J., Basile, V., Evang, K., Venhuizen, N. J., and Bjerva, J. (2017).
The groningen meaning bank.
*Handbook of linguistic annotation*, pages 463–496.

[Bouma et al., 2020] Bouma, G., Seddah, D., and Zeman, D. (2020).
Overview of the iwpt 2020 shared task on parsing into enhanced universal dependencies.
In *58th Annual Meeting of the Association for Computational Linguistics*.

# References II

[Bouma et al., 2021] Bouma, G., Seddah, D., and Zeman, D. (2021).
From raw text to enhanced universal dependencies: The parsing shared task at iwpt 2021.
In *Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies (IWPT 2021)*, pages 146–157. Association for Computational Linguistics (ACL).

[De Marneffe et al., 2021] De Marneffe, M.-C., Manning, C. D., Nivre, J., and Zeman, D. (2021).
Universal dependencies.
*Computational linguistics*, 47(2):255–308.

[Gerdes et al., 2018] Gerdes, K., Guillaume, B., Kahane, S., and Perrier, G. (2018).
Sud or surface-syntactic universal dependencies: An annotation scheme near-isomorphic to ud.
In *Universal dependencies workshop 2018*.

[Greene and Cunningham, 2006] Greene, D. and Cunningham, P. (2006).
Practical solutions to the problem of diagonal dominance in kernel document clustering.
In *Proc. 23rd International Conference on Machine learning (ICML'06)*, pages 377–384. ACM Press.

[Hajic et al., 2001] Hajic, J., Vidová-Hladká, B., and Pajas, P. (2001).
The prague dependency treebank: Annotation structure and support.
In *Proceedings of the IRCS workshop on linguistic databases*, pages 105–114.

[Ivoylova et al., 2023] Ivoylova, A., Dyachkova, D., Petrova, M., and Michurina, M. (2023).
The problem of linguistic markup conversion: the transformation of the compreno markup into the ud format.
In *International Conference on Computational Linguistics and Intellectual Technologies «Dialog*.

# References III

[Li et al., 2023] Li, Y., Wang, S., Lin, C., Guerin, F., and Barrault, L. (2023).
Framebert: Conceptual metaphor detection with frame embedding learning.
*arXiv preprint arXiv:2302.04834*.

[Mel'cuk et al., 1988] Mel'cuk, I. A. et al. (1988).
*Dependency syntax: theory and practice*.
SUNY press.

[Petrova et al., 2023] Petrova, M., Ivoylova, A., Bayuk, I., Dyachkova, D., and Michurina, M. (2023).
The cobald annotation project: the creation and application of the full morpho-syntactic and semantic markup standard.
In *Proceedings of the International Conference "Dialogue*, volume 2023.

[Ponomareva et al., 2021] Ponomareva, M., Petrova, M., Detkova, J., Serikov, O., and Yarova, M. (2021).
Semsketches2021: experimenting with the machine processing of the pilot semantic sketches corpus.
In *Komp'juternaja Lingvistika i Intellektual'nye Tehnologii*, pages 560–570.

[Ponti et al., 2018] Ponti, E. M., Vulić, I., Glavaš, G., Mrkšić, N., and Korhonen, A. (2018).
Adversarial propagation and zero-shot cross-lingual transfer of word vector specialization.
*arXiv preprint arXiv:1809.04163*.

[Schuster and Manning, 2016] Schuster, S. and Manning, C. D. (2016).
Enhanced english universal dependencies: An improved representation for natural language understanding tasks.
In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2371–2378.

# References IV

[Uchida and Zhu, 2001] Uchida, H. and Zhu, M. (2001).
The universal networking language beyond machine translation.
In *International Symposium on Language in Cyberspace, Seoul*, pages 26–27.

[White et al., 2016] White, A. S., Reisinger, D., Sakaguchi, K., Vieira, T., Zhang, S., Rudinger, R., Rawlins, K., and Van Durme, B. (2016).
Universal decompositional semantics on universal dependencies.
In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1713–1723.