

## **Document Overview**

- This is a documentation overview of how to use Batchalign2 to do syntactic and morphological parsing

## **Batchalign2 Overview**

- Batchalign2 is a command-line pipeline developed under the TalkBank project for automatic morphological tagging and Universal Dependencies (UD) syntactic parsing across multiple languages, including Italian.
- It takes linguistic transcripts in the CHAT (.cha) format and produces two key annotation layers:
  - %mor: morphological analysis (lemmas, parts of speech, and inflectional features)
  - %gra: syntactic analysis following the Universal Dependencies (UD) standard.
- Batchalign2 integrates the **Stanza** NLP models and the **UD framework** into the TalkBank/CLAN ecosystem, enabling consistent morphosyntactic annotation of multilingual corpora.
- It automates tokenization, morphological tagging, lemmatization, and dependency parsing, storing all results back in CHAT format—making the output directly usable with CLAN tools for further linguistic or developmental analysis.
- Functions and Commands:
  - asr: ASR!
  - morphosyntax: PoS and dependency analysis
  - fa: Forced Alignment (requires utterance-level timings already)

## **Universal Dependencies Overview**

- **Universal Dependencies (UD)** is an international framework for the consistent annotation of grammar—covering parts of speech, morphological features, and syntactic dependencies—across human languages.
- UD is an open, collaborative project involving over 600 contributors, maintaining more than 200 treebanks in 150+ languages.
- Its goal is to create a cross-linguistically consistent representation of grammatical relations, allowing comparable syntactic analysis across languages like Italian, Mandarin, English, and many others.

## **What we need to do in terms of using CLAN and Batchalign2 to do morphological and syntactic analysis and parsing:**

1. **Download CLAN (MacBook or Windows)**  
<https://dali.talkbank.org/clan/>
2. **Download Batchalign/Environment Setup (Open MacBook Terminal and paste the following)**

```

# Install UV (a lightweight Python environment manager)
curl -LsSf https://astral.sh/uv/install.sh | sh

# Install Batchalign2 for Python 3.11
UV_PYTHON=3.11 uv tool install batchalign

# Ensure ~/.local/bin is in your PATH
echo 'export PATH="$HOME/.local/bin:$PATH"' >> ~/.zshrc
source ~/.zshrc

# Check installation
batchalign --help

```

### 3. Data Preprocessing (R Studio) → also see [txt-to-cha.R](#)

**Aim data cleaning:** Txt file → cha file

- A valid .cha file must follow CHAT conventions: every utterance on its own line and starting with a speaker code (e.g., \*ADU:, \*CHI:, \*MOT:).

**Tips:**

- A Cha file can be automatically converted to a txt file and vice versa
- @Languages: must include the language code (e.g., ita, zho, eng, spa).
  - The Italian model will be loaded automatically from Stanza/UD based on @Languages: ita.
- Sentences must end with a space and punctuation (., !, or ?).
- Delete empty or malformed tokens like <>, which cause lexer errors.

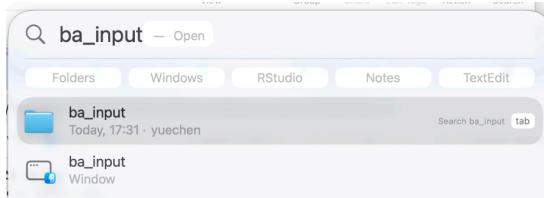
**Things we need to do for the VoxCommunis dataset:**

1. Add speaker tier at the beginning of each sentence (e.g., \*ADU: )
2. Remove all other symbols (e.g., @, \*, % or <>:)
3. Ensure a space before punctuations (e.g., . ! or ?)
4. Add a final period if missing
5. Make sure to add the following at the beginning of the cha file  
@UTF8  
@Begin  
@Languages: ita  
@Participants: ADU Adult  
@ID: ita|troncamento|ADU||||Adult|||
6. Make sure to add the following at the end of the cha file  
@End

### **Successful Example:**

@UTF8 → Must  
@Begin → Must  
@Languages: ita → Must  
@Participants: ADU Adult → Must  
@ID: ita|troncamento|ADU||||Adult|| → Must  
@Media: 10192025, audio → optional  
@Date: 19-OCT-2025 → optional  
\*ADU: Il libro ha suscitato molte polemiche a causa dei suoi contenuti . → Must  
@End → Must

### **Sample Input file (where you should put it? Search ba\_input!):**



```
1 @UTF8
2 @Begin
3 @Languages: ita
4 @Participants: ADU Adult
5 @ID: ita|troncamento|ADU||||Adult||
6 @Media: 10192025, audio
7 @Date: 19-OCT-2025
8 *ADU: Il libro ha suscitato molte polemiche a causa dei suoi contenuti .
9 @End
10
```

### **4. Parsing: Commands used for Batchalign2 (MacBook Terminal)**

```
# check input file
sed -n '1,25p' ~/ba_input/troncamento.cha

# run batch aligner
batchalign morphotag ~/ba_input ~/ba_output

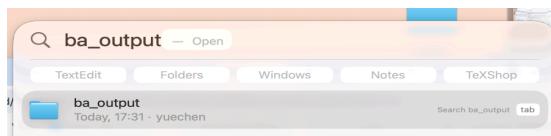
# check output file %mor / %gra
```

```
grep -n "%^mor:\|^%gra:" ~/ba_output/troncamento.cha | head
```

What you will see if it's successfully running:

```
● ● ●  yuechen — python -m batchalign morphotag ~/ba_input ~/ba_output — 8...  
  
/Users/yuechen/.local/share/uv/tools/batchalign/lib/python3.11/site-packages/pratio/utilities/utils.py:9: UserWarning: pkg_resources is deprecated as an API. See https://setuptools.pypa.io/en/latest/pkg_resources.html. The pkg_resources package is slated for removal as early as 2025-11-30. Refrain from using this package or pin to Setuptools<81.  
    from pkg_resources import resource_filename  
/Users/yuechen/.local/share/uv/tools/batchalign/lib/python3.11/site-packages/pyanote/audio/core/io.py:212: UserWarning: torchaudio._backend.list_audio_backends has been deprecated. This deprecation is part of a large refactoring effort to transition TorchAudio into a maintenance phase. The decoding and encoding capabilities of PyTorch for both audio and video are being consolidated into TorchCode c. Please see https://github.com/pytorch/audio/issues/3902 for more information.  
    It will be removed from the 2.9 release.  
    torchaudio.list_audio_backends()  
  
Mode: morphotag; got 1 transcript to process from /Users/yuechen/ba_input:  
  
Downloading https://raw.githubusercontent.com/stanfordnlp/stanza-resources/main/  
Downloading  
https://raw.githubusercontent.com/stanfordnlp/stanza-resources/main/resources_1.  
11.0.json: 436kB [00:00, 242MB/s]  
  
: troncamento.cha -————— 3% 0:32:56 Running: Morpho-Syntax
```

Sample Output file (where you can find it? Search ba\_output!):



```
● ● ●  /Users/yuechen/Downloads/troncamento.cha  
  
1 @Begin  
2 @Languages: ita  
3 @Participants: ADU Adult  
4 @ID: ita|troncamento|ADU||||Adult|||  
5 @Media: 10192025, audio  
6 @Date: 19-OCT-2025  
7 *ADU: Il libro ha suscitato molte polemiche a causa dei suoi contenuti .  
8 %mor: det|il-Masc-Def-Art-Sing noun|libro-Masc aux|avere-Fin-Ind-Pres-S3 verb|suscitare-Part-  
9     Past-S det|molti-Fem-Def-Ind-Plur noun|polemica-Fem-Plur-Acc adp|a noun|causa-Fem det|di-  
10    Masc-Def-Ind-Plur det|suo-Masc-Def-Prs-Plur noun|contenuto-Masc-Plur-Acc .  
11 %gra: 1|2|DET 2|4|NSUBJ 3|4|AUX 4|11|ROOT 5|6|DET 6|4|OBJ 7|8|CASE 8|4|OBL 9|11|DET  
12     10|11|DET-POSS 11|4|OBJ 12|4|PUNCT  
@End
```

## 5. Understanding the Output:

**Sample sentence:** Il libro ha suscitato molte polemiche a causa dei suoi contenuti .

### Morphological layer (%mor):

```
%mor: det|il-Masc-Def-Art-Sing noun|libro-Masc aux|avere-Fin-Ind-Pres-S3  
verb|suscitare-Part-Past-S det|molto-Fem-Def-Ind-Plur noun|polemica-Fem-Plur-Acc  
adp|a noun|causa-Fem det|di-Masc-Def-Ind-Plur det|suo-Masc-Def-Prs-Plur  
noun|contenuto-Masc-Plur-Acc .
```

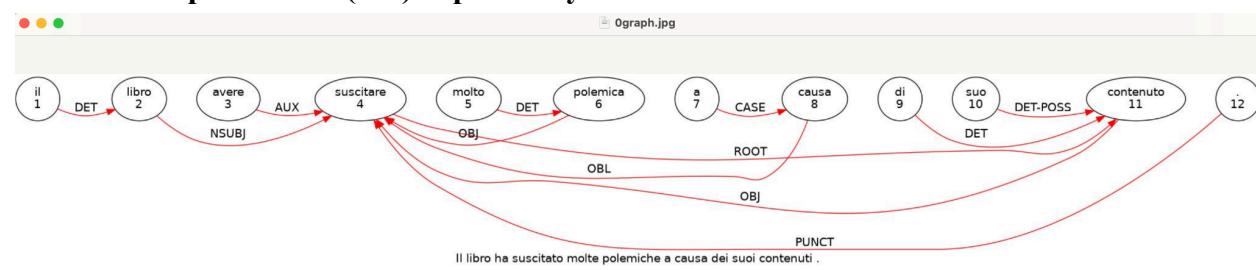
- Breaks the sentence down word-by-word and encodes part of speech + grammatical features (gender, number, tense, etc.).
  - Lemma (dictionary form): e.g. libro, suscitare
  - Part of speech: noun, verb, det, adp, etc.
  - Morphosyntactic features: Gender: Masc/Fem, Number: Sing/Plur, Tense: Pres, Past, Fut, Mood: Ind (indicative), Subj (subjunctive), Person: S1, S2, S3 (subject 1st, 2nd, 3rd person), Case: Nom (nominative), Acc (accusative), etc.

### Syntactic layer (%gra):

```
%gra: 1|2|DET 2|4|NSUBJ 3|4|AUX 4|11|ROOT 5|6|DET 6|4|OBJ 7|8|CASE 8|4|OBL  
9|11|DET 10|11|DET-POSS 11|4|OBJ 12|4|PUNCT
```

- These lines show token IDs (word order, such as word 1, word 2, word 3), dependency heads, and UD relations (e.g., NSUBJ = subject, OBL = oblique).

### Universal Dependencies (UD) dependency tree:



### What this shows:

- Each oval = a token (word), labeled with
  - its lemma (dictionary form), and
  - its index (1–12 here).
- Each red arrow = a dependency relation between two words:

- The arrow points from the head (governing word) to its dependent (modified word).
- The label on the arrow (e.g., NSUBJ, AUX, OBJ, OBL, etc.) indicates the syntactic relation type.

## **6. Some CLAN Post-processing (optional, if you want to look at a different analysis)**

- Once %mor and %gra tiers exist, you can analyze the Italian corpus with CLAN tools:
  - kwal +t%mor ~/ba\_output/troncamento.cha # view morphology
  - kwal +t%gra +s"OBL" ~/ba\_output/troncamento.cha # search dependencies
  - freq +t%mor -t\* ~/ba\_output/troncamento.cha # POS frequency

## **7. Citation and Documentation**

- Batchalign2: <https://talkbank.github.io/batchalign2/>
- [CLAN Manual](#), section on Batchalign and role conversion
- **Reference article:** [Liu, H. & MacWhinney, B. \(2024\). Morphosyntactic Analysis for CHILDES. Language Development Research, 4\(1\), 233–258.](#)