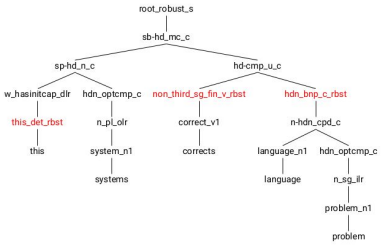
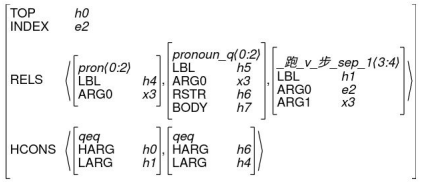
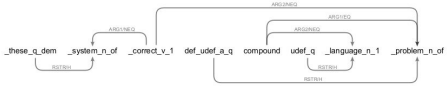
DELPH-IN Parsers/Demo @VU Amsterdam







Luis Morgado da Costa
VU Amsterdam

2nd July, 2024, Olomouc

VU VRIJE UNIVERSITEIT AMSTERDAM

DELPH-IN Grammarium

First introduced at Bellingham/Virtual Summit (2020)

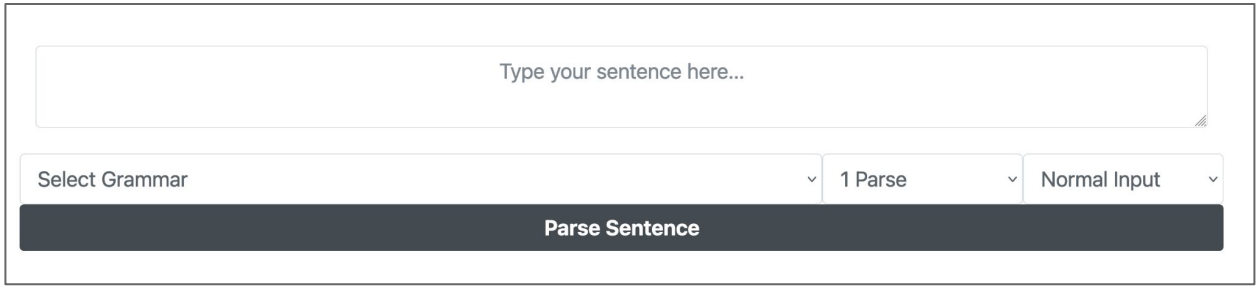
Grammarium:

- Part of larger web-app (iTELL: GED, CALLIG, etc.)
- Web-based toolkit for grammar development
- DELPH-IN Viz libraries
- Grammar Update/Compile (from ~~SVN~~/GitHub)
- Inspect Profiles (/ Basic Regression Testing)

Today's Focus:

- Grammar/Parser demo
- Learn what do people want moving forward

Grammar/Parser demo



Here is a list of grammars currently included:

- English Resource Grammar (aka ERG)
- ZHONG (for Mandarin Chinese)
- JACY (for Japanese)
- INDRA (for Indonesian)
- PorGram (for Portuguese)
- YueGram (for Cantonese)
- Spanish Resource Grammar (partial support, through YY input)

Shout out to Olga for the friendly pressure!

I'm happy to add any other grammar that has ACE support.

https://compling.cltl.labs.vu.nl/itell

Text Input: ZHONG, JACY, YueGram, etc.

- Some grammars expect special input – e.g. tokenization (ZHONG and Co.)
- While grammarians are most certainly always right (😉), the righteous path is often not clear for end users

JACY: 犬が寝ています。[The dog is sleeping]

Expects: 犬　が　寝　て　い　ます　。

ZHONG: 你会不会开车？ [Can you drive?]

Expects: 你　会不会　开　车　？

Accepts: 你　会不会　开　车　？

Accepts: 你　会不会　开车　？

YY Input: JACY vs SRL

JACY:

太郎 が 寝る

グッドマン が 寝る

組織 [?] – Organization[?]

(0, 0, 1, <0:5>, 1, "グッドマン", 0, "null", "名詞-固有名詞-**組織**:n-n" 1.0)(1, 1, 2, <5:6>, 1, "が", 0, "null", "助詞-格助詞-一般:n-n" 1.0)(2, 2, 3, <6:8>, 1, "寝る", 0, "null", "動詞-自立:一段-基本形" 1.0)

SRL:

(1, 0, 1, <0:3>, 1, **"el" "los"**, 0, **"da0mp0"**, "da0mp0" 0.99272812) (2, 1, 2, <4:9>, 1, "gato" "gatos", 0, "ncmp000", "ncmp000" 1.00000000) (3, 2, 3, <10:17>, 1, "dormir" "duermen", 0, "vmip3p0", "vmip3p0" 1.00000000) (4, 3, 4, <17:18>, 1, "." ".", 0, "fp", "fp" 1.00000000)

YY Input: JACY vs SRL

JACY:

太郎 が 寝る

グッドマン が 寝る

組織 [?] – Organization[?]

(0, 0, 1, <0:5>, 1, "グッドマン", 0, "null", "名詞-固有名詞-**組織**:n-n" 1.0)(1, 1, 2, <5:6>, 1, "が", 0, "null", "助詞-格助詞-一般:n-n" 1.0)(2, 2, 3, <6:8>, 1, "寝る", 0, "null", "動詞-自立:一段-基本形" 1.0)

SRL:

(1, 0, 1, <0:3>, 1, **"el" "los"**, 0, **"da0mp0"**, "da0mp0" 0.99272812) (2, 1, 2, <4:9>, 1, "gato" "gatos", 0, "ncmp000", "ncmp000" 1.00000000) (3, 2, 3, <10:17>, 1, "dormir" "duermen", 0, "vmip3p0", "vmip3p0" 1.00000000) (4, 3, 4, <17:18>, 1, "." ".", 0, "fp", "fp" 1.00000000)

Turns out SRL is also using yy-rules (I have a rough idea but I'm not 100% sure what that is) — **This must to be documented for (even expert) users!**

What may be nice to show users:

Summary: ZHONG is a mid-sized grammar, with targeted coverage. Much of its development has been centered around educational data, which means it is able to provide adequate analyses for most of the sentence patterns found in early to intermediate educational materials. From a linguistic point of view, ZHONG offers in-depth analyses of interesting and well-studied linguistic phenomena including: serial-verbs, 的 (DE) constructions, 把 (BA) constructions, interactions between aspect and negation, sentence final particles, and interrogatives.

Input: ZHONG expects its input to be segmented at the word level. This means that words must be separated by spaces (see some of the examples below). Some of our analyses expect input to be tokenized differently than most Mandarin Chinese word segmentation tools (e.g., 这个 should be tokenized as two tokens 这 and 个). To the best of its ability, the grammar uses REPP (internal preprocessing) to help further segment the input (e.g., if 这个 is provided as a single token, it will be internally split as two tokens).

Examples:

我 正在 看 电视 呢 。
这 个 商场 很 大 。
我们 还 要 三 碗 汤 。
你 会 不 会 开车？
语言 学院 有 多少 个 系？

Visiting Profiles

Simple Profile View

Show 5 entries

Showing 1 to 5 of 12 entries (filtered from 107 total entries)

Previous 1 2 3 Next

i-id	i-wf	readings	i-input	i-comment	
831	1	50	It took Abrams ten minutes to arrive.	Det tok Abrams ti minutter å ankomme.	PARSE
971	1	45	The cats found a way to go.	Kattene fant en måte å bjeffe på.	PARSE
881	1	34	Abrams barked from ten to three.	Abrams bjeffet fra ti til tre.	PARSE
841	1	26	Abrams left it to Browne to bark.	Abrams overlot til Browne å bjeffe.	PARSE
791	1	11	Abrams promised Browne to go.	Abrams lovte Browne å bjeffe.	PARSE
i-id	i-wf	reading	" to "	i-comme	

erg2018.dat

Parse Skeleton

Thoughts…

- I’m interested to develop/host these tools further, if there is interest
- It may not be advisable to hardcode some of this info (which changes as grammars develop) – an easy solution would be a JSON config [???]

User External Info:

- Intro to the grammar
- Expected input (and which mode to use?)
- Examples (with or without notes?)

Grammarium:

- Which ACE config to use?
- PATH to profiles to expose?
- API to process the input / Easy to use Python script to process it?

Thank you!