ELEXIS-WSD Parallel Sense-Annotated Corpus

Presentation and Pipeline Demo (INCEpTION)
UniDive WG2 T2.2: Design of a lexicon-corpus interface

Jaka Čibej[1], Carole Tiberius[2], Simon Krek[1]
[1]Jožef Stefan Institute, Slovenia
[2]Dutch Language Institute, The Netherlands

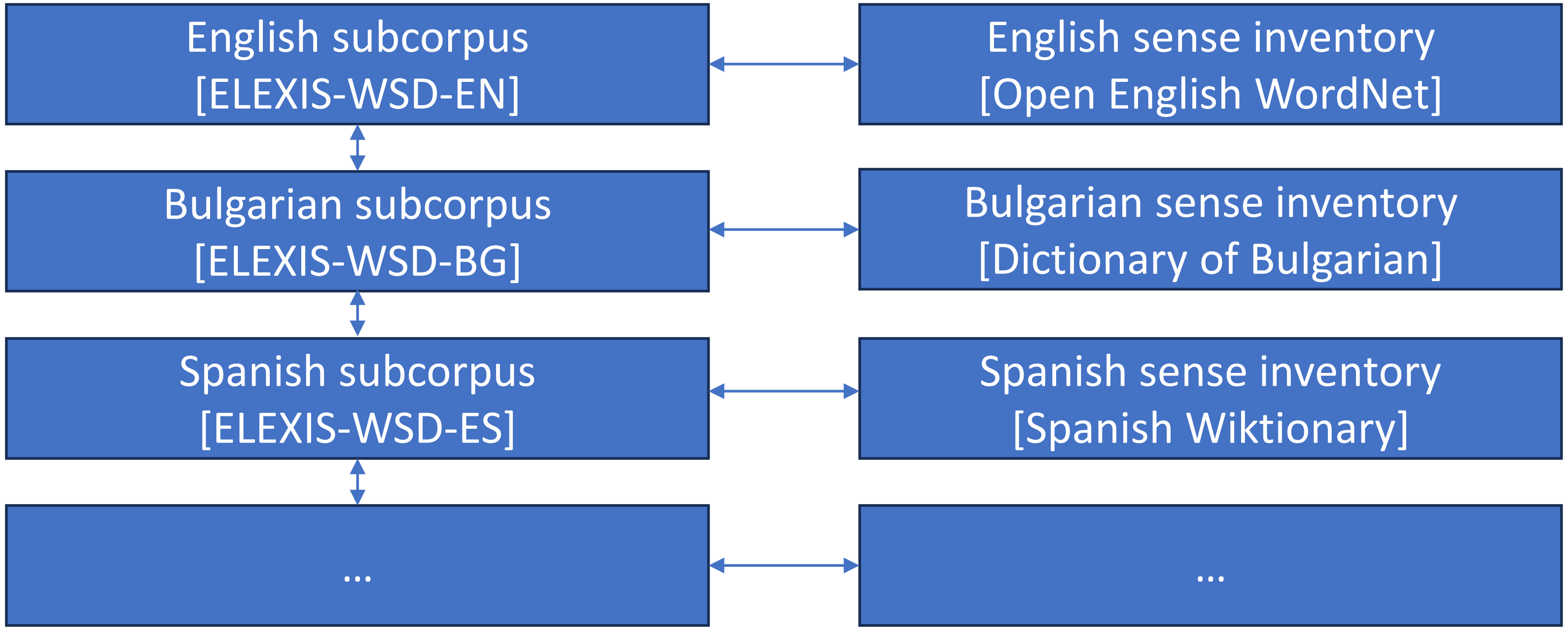
Naples, Italy, 9 February 2024

ELEXIS-WSD Parallel Sense-Annotated Corpus

- Version 1.0 (and 1.1) compiled within the ELEXIS project
  - European Lexicographic Infrastructure, https://elex.is
  - available in the CLARIN.SI repository: http://hdl.handle.net/11356/1842
  - motivation:
    - lack of high-quality manually-curated data/lexical-semantic resources
    - foster collaboration between lexicography and natural language processing
- For further details, see:
  - Martelli et al. (2021)
  - Krek et al. (2023)

What does the corpus consist of?

- Current content (v1.1): 10 parallel subcorpora + 10 sense inventories



English subcorpus
[ELEXIS-WSD-EN]

English sense inventory
[Open English WordNet]

Bulgarian subcorpus
[ELEXIS-WSD-BG]

Bulgarian sense inventory
[Dictionary of Bulgarian]

Spanish subcorpus
[ELEXIS-WSD-ES]

Spanish sense inventory
[Spanish Wiktionary]

…

…

- Bulgarian, Danish, English, Spanish, Estonian, Hungarian, Italian, Dutch, Portuguese, Slovene

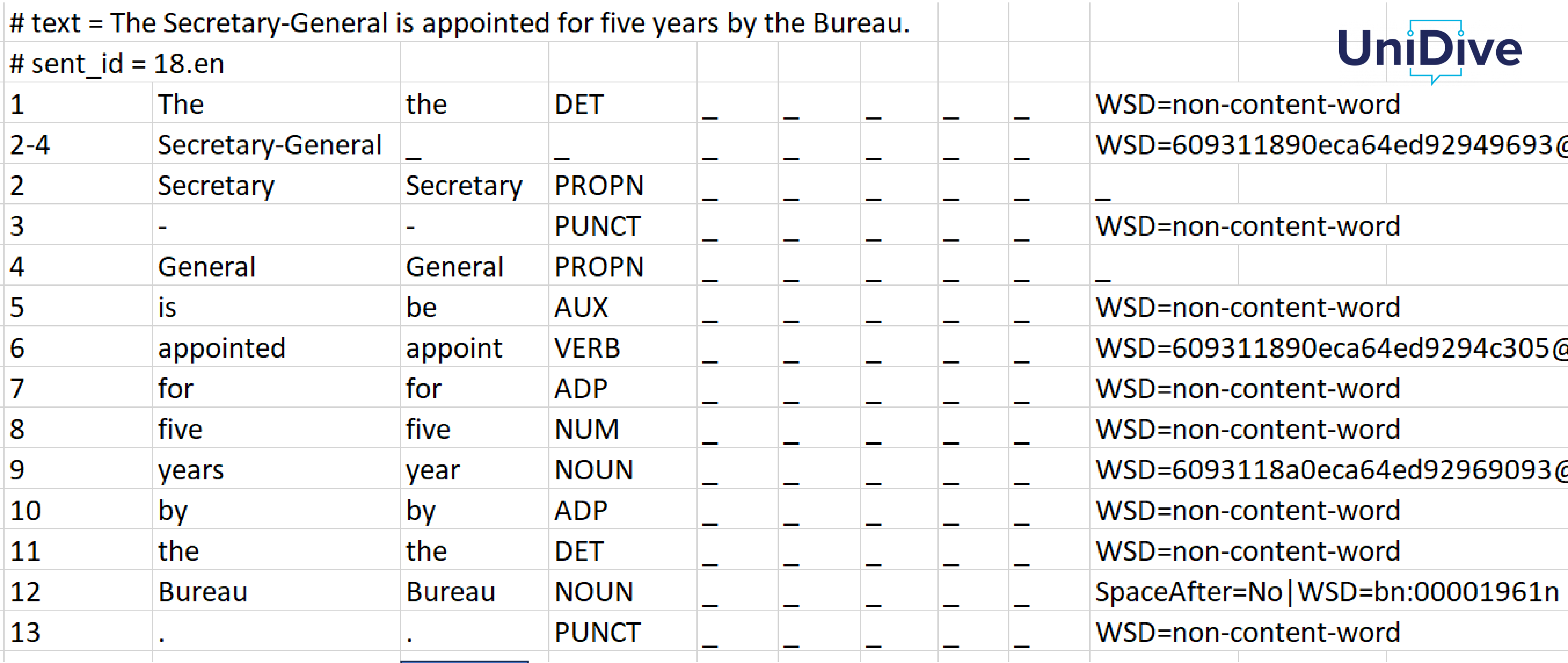
(Sub)corpora

- 2,024 sentences from WikiMatrix (Wikipedia)
- Originally in English, then translated into other languages.

- # sent_id = 4.en → More than 7,000 people visited the film's premiere in Damascus.
- # sent_id = 4.it → Oltre 7.000 spettatori hanno assistito alla prima proiezione del film a Damasco.
- # sent_id = 4.nl → Meer dan 7.000 mensen bezochten de première van de film in Damascus.
- # sent_id = 4.sl → Več kot 7000 ljudi je obiskalo premiero filma v Damasku.

Annotation layers in the latest version (v1.1)

- tokenization (including multiword-tokens)
- lemmatization
- pos-tagging (UPOS)
- word-sense disambiguation
- only for some languages: MWE-annotation (**only spans**, no categorization!)



text = The Secretary-General is appointed for five years by the Bureau.
sent_id = 18.en

1	The	the	DET	_	_	_	_	_	WSD=non-content-word
2-4	Secretary-General	_	_	_	_	_	_	_	WSD=609311890eca64ed92949693@
2	Secretary	Secretary	PROPN	_	_	_	_	_	_
3	-	-	PUNCT	_	_	_	_	_	WSD=non-content-word
4	General	General	PROPN	_	_	_	_	_	_
5	is	be	AUX	_	_	_	_	_	WSD=non-content-word
6	appointed	appoint	VERB	_	_	_	_	_	WSD=609311890eca64ed9294c305@
7	for	for	ADP	_	_	_	_	_	WSD=non-content-word
8	five	five	NUM	_	_	_	_	_	WSD=non-content-word
9	years	year	NOUN	_	_	_	_	_	WSD=6093118a0eca64ed92969093@
10	by	by	ADP	_	_	_	_	_	WSD=non-content-word
11	the	the	DET	_	_	_	_	_	WSD=non-content-word
12	Bureau	Bureau	NOUN	_	_	_	_	_	SpaceAfter=No|WSD=bn:00001961n
13	.	.	PUNCT	_	_	_	_	_	WSD=non-content-word

(MW-)tokens

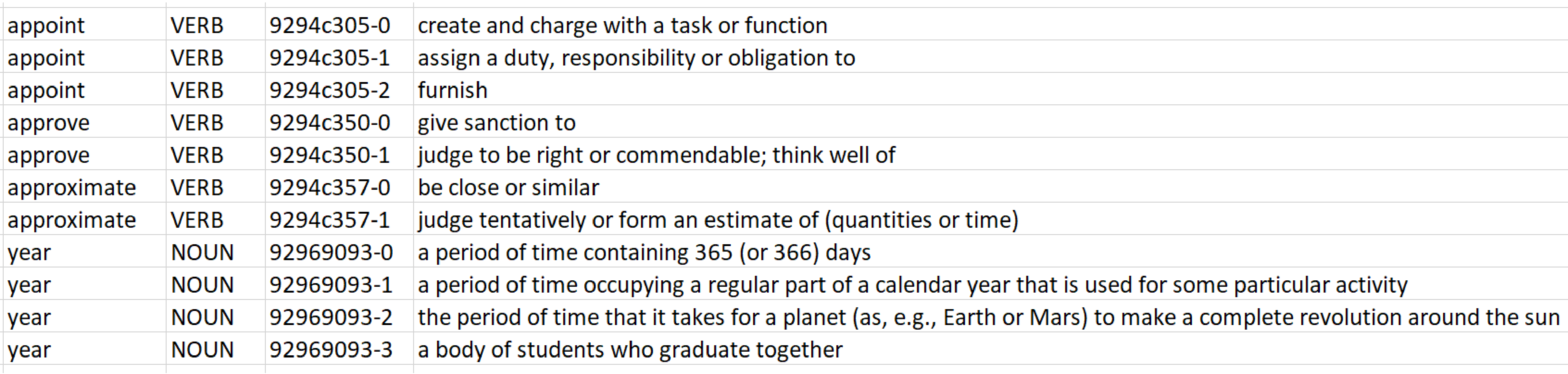
lemmas

UPOS

sense annotation (sense ID from sense inventory)

Sense Inventories

- consist of lemmas, their UPOS, and their senses (definitions)



appoint	VERB	9294c305-0	create and charge with a task or function
appoint	VERB	9294c305-1	assign a duty, responsibility or obligation to
appoint	VERB	9294c305-2	furnish
approve	VERB	9294c350-0	give sanction to
approve	VERB	9294c350-1	judge to be right or commendable; think well of
approximate	VERB	9294c357-0	be close or similar
approximate	VERB	9294c357-1	judge tentatively or form an estimate of (quantities or time)
year	NOUN	92969093-0	a period of time containing 365 (or 366) days
year	NOUN	92969093-1	a period of time occupying a regular part of a calendar year that is used for some particular activity
year	NOUN	92969093-2	the period of time that it takes for a planet (as, e.g., Earth or Mars) to make a complete revolution around the sun
year	NOUN	92969093-3	a body of students who graduate together

lemma

UPOS

Sense ID

Sense Definition

Want to participate with your language?

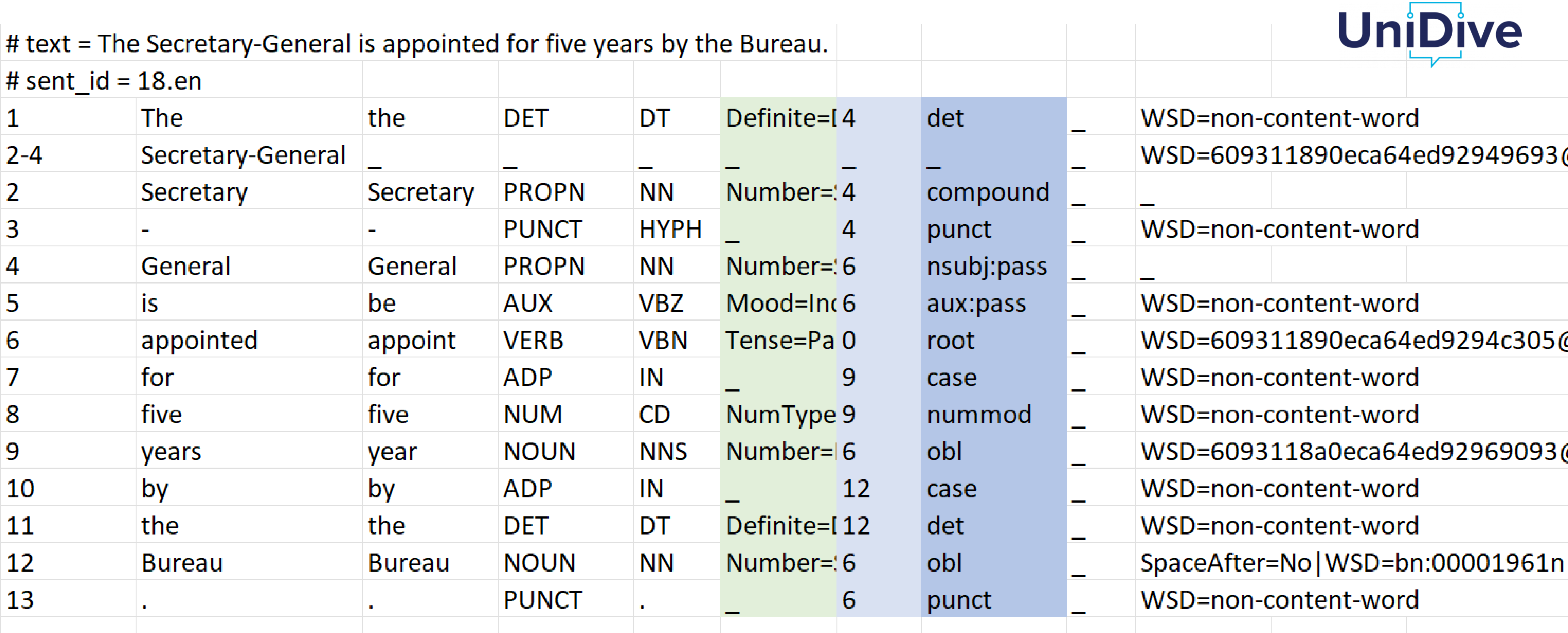
- **Main precondition: sense inventory available under CC BY-SA 4.0**
- 5.000 lemmas (not necessarily a full dictionary!)
- We prefer lexicographic data (not possible for all languages because of copyright issues/restrictions on data accessibility)

Additional annotation layers in UniDive

- UD-parsing (added automatically through UDPipe)
- Annotation of MWEs
- Annotation of named entities

- + extension of the corpus for other languages:
  - So far: Serbian, Croatian, Polish, Romanian, Greek

UniDive



text = The Secretary-General is appointed for five years by the Bureau.

sent_id = 18.en

1	The	the	DET	DT	Definite=[	4	det	_	WSD=non-content-word
2-4	Secretary-General	_	_	_	_	_	_	_	WSD=609311890eca64ed92949693(
2	Secretary	Secretary	PROPN	NN	Number=!	4	compound	_	_
3	-	-	PUNCT	HYPH	_	4	punct	_	WSD=non-content-word
4	General	General	PROPN	NN	Number=!	6	nsubj:pass	_	_
5	is	be	AUX	VBZ	Mood=Inc	6	aux:pass	_	WSD=non-content-word
6	appointed	appoint	VERB	VBN	Tense=Pa	0	root	_	WSD=609311890eca64ed9294c305(
7	for	for	ADP	IN	_	9	case	_	WSD=non-content-word
8	five	five	NUM	CD	NumType	9	nummod	_	WSD=non-content-word
9	years	year	NOUN	NNS	Number=I	6	obl	_	WSD=6093118a0eca64ed92969093(
10	by	by	ADP	IN	_	12	case	_	WSD=non-content-word
11	the	the	DET	DT	Definite=[	12	det	_	WSD=non-content-word
12	Bureau	Bureau	NOUN	NN	Number=!	6	obl	_	SpaceAfter=No|WSD=bn:00001961n
13	.	.	PUNCT	.	_	6	punct	_	WSD=non-content-word

(MW-)tokens

lemmas

UPOS

XPOS

FEATS

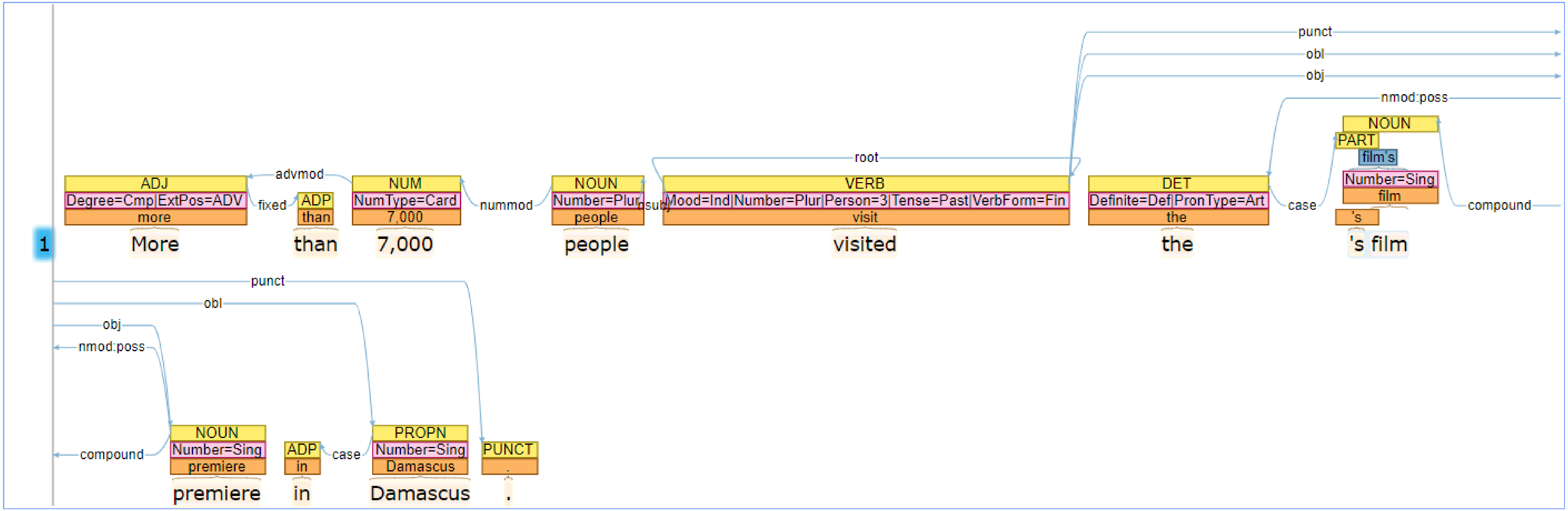
HEAD

DEPREL

sense annotation (sense ID from sense inventory)

Manual Annotation/Corrections in INCEpTION

- https://inception-project.github.io/


punct
obl
obj
nmod:poss
NOUN
PART
film's
Number=Sing
film
's
root
ADJ
Degree=Cmp|ExtPos=ADV
more
advmod
fixed
ADP
than
NUM
NumType=Card
7,000
nummod
NOUN
Number=Plur
people
nsubj
VERB
Mood=Ind|Number=Plur|Person=3|Tense=Past|VerbForm=Fin
visit
DET
Definite=Def|PronType=Art
the
case
compound
1
More
than
7,000
people
visited
the
's film
punct
obl
obj
nmod:poss
compound
NOUN
Number=Sing
premiere
ADP
in
case
PROPN
Number=Sing
Damascus
PUNCT
.
premiere
in
Damascus
.