DepClinBR

A dependency-annotated corpus of clinical text in Portuguese for NLP tasks

Projeto: Exploring deep language models to leverage Portuguese and French biomedical semantic resources

Profa. Dra. Cláudia Moro

Health Artificial Intelligence Lab (HAILab)
Pprograma de Pós-Graduação em Tecnologia em Saúde da Pontifícia / PUCPR
Haute Ecole Spécialisée de Suisse Occidentale / HES-SO

Research team

Health Artificial Intelligence Lab (HAILab) PUC/PR

Coordenador: Claudia Moro

Douglas Teodoro
Elisa Terumi Rubel Schneider
João Vitor Andrioli de Souza
Lucas Emanuel Silva e Oliveira
Lucas Ferro Antunes de Oliveira
Yohan Gumiel

Laboratory for Experimentation in Translation (LETRA) UFMG

Coordenador: Adriana S Pagano
Thiago Castro Ferreira
Ana Clara Souza Pagano
Barbara Cristina Barbalho
Douglas Francisco de Cillo
Iasmin Rabelo
Luiz Felipe Batista Monteiro
Sarah Teixeira
Tarcio do Vale Gomes

Clinical text

- ➔ non-structured format (hand-written notes; patient reports)
- ➔ particular style due to time-pressure and minute detail
- ➔ relevant to enhance clinical activities (to improve patient-healthcare provider interaction; to aid treatment personalization; to automate risk stratification to adverse drug events; to predict associated events)

DALIANIS, 2018

Corpora and models at HAILab

SemClinBr --> semantically-annotated corpus

TempClinBr --> corpus annotated with temporal relations between clinical concepts (University of Rennes - France)

SummClinBr --> corpus of clinical concepts for patient record summarization

BioBertPt --> word embeddings model for clinical text in Portuguese

DepClinBr --> dependency annotated corpus

https://github.com/HAILab-PUCPR

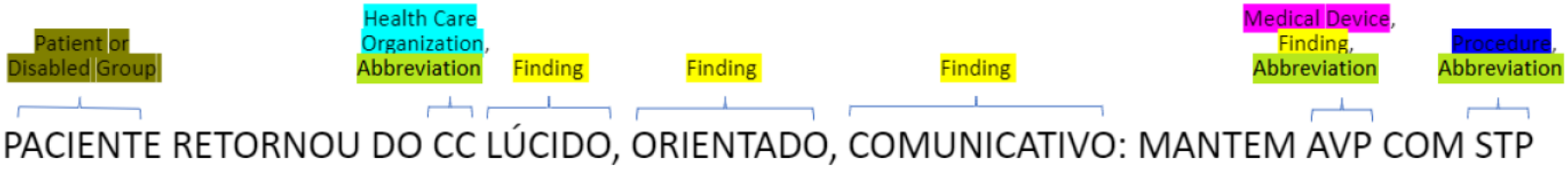
Corpus

- Clinical narratives provided by three Brazilian hospitals
- Including discharge summaries, medical nursing notes, ambulatory records, clinical evolution
- Related to various subdomains cardiology, nephrology, or endocrinology.
- Some following a topic structure (e.g., SOAP - Subjective, Objective, Assessment and Plan)
- Preprocessed, Anonymized and Pseudonymized

Clinical concept annotation

Semantic types from Unified Medical Language System® (UMLS®) Semantic Network https://bioportal.bioontology.org/ontologies/STY/?p=summary

Figure 8 – Named Entity Recognition algorithm applied to a clinical text to identify the categories of each concept found within the text.



Fonte: OLIVEIRA, 2020

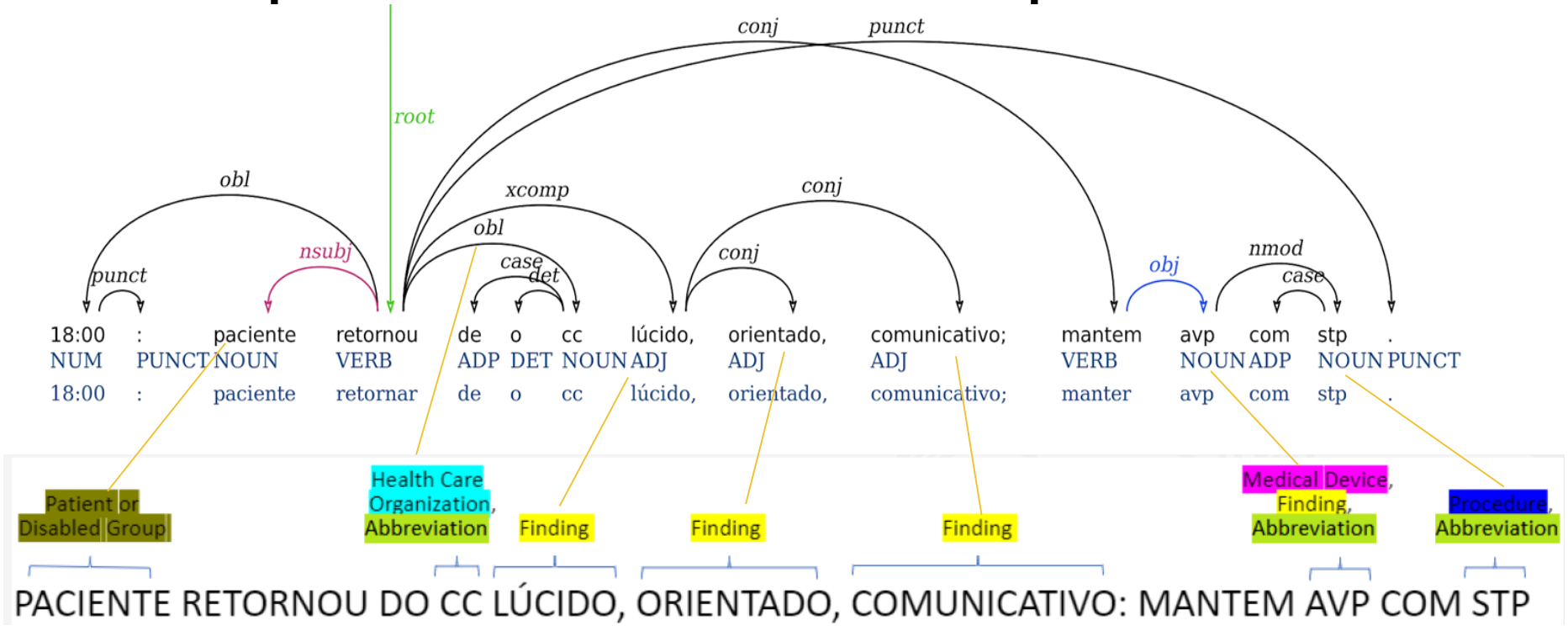
Clinical concept annotation

Table 11 – Text samples containing the most used semantic types (STY) and their corresponding semantic groups (SGR). The third column shows the original examples, and the fourth shows the translated versions. The underlined passages indicate annotated concepts.

SGR	STY	Original examples	Translated examples
Anatomy	Body Location or Region	MEIA TALA GESSADA EM MIE apresenta edema em região craniana ABDÔMEN PLANO E FLÁCIDO	Half-length plaster cast in LLL presents edema in the cranial region FLAT AND FLACID ABDOMEN
Anatomy	Body Part, Organ, or Organ Component	acesso venoso central em jugular D ACESSO VENOSO PERIFERICO EM BRAÇO DIREITO	right jugular central venous access RIGHT ARM PERIPHERAL VENOSOUS ACCESS
Chemicals & Drugs	Organic Chemical	Fez uso de atenolol por 3 anos cefaléia em regiao parietal bilateral que melhora com dipirona	used atenolol for 3 years headache in bilateral parietal region improved with dipyrone
Chemicals & Drugs	Pharmacologic Substance	asmatica em uso de salbutamol e budesonida	asthmatic person using salbutamol and budesonide
Concepts & Ideas	Temporal Concept	POI DE LAVAGEM + CURETA DE TECIDO NECRÓTICO Paciente em Pré-operatório de FX fêmur	WASHING IP + NECROTIC TISSUE CURETAGE Preoperative patient of femur fracture
Devices	Drug Delivery Device	cloreto de potassio a 42 ml/h em bomba de infusão	potassium chloride at 42 ml/h in infusion pump
Devices	Medical Device	AVP em MSE com soroterapia em curso SVD com diurese efetiva	PVA in LUL with ongoing serotherapy DBP with effective diuresis
Disorders	Disease or Syndrome	REFERE HIPERTENSÃO E DIABETES EM USO DE INSULINA. SINDROME DE GUILLAIN BARRE.	REFERS HYPERTENSION AND DIABETES IN INSULIN USE GUILLAIN BARRE SYNDROME
Disorders	Finding	RETORNOU DO CC LÚCIDO, ORIENTADO, COMUNICATIVO consciente, comunicativo, pupilas isocóricas fotoreagentes	RETURNED LUCID FROM SC CONSCIOUS, COMMUNICATIVE conscious, communicative, photoreagent isochoric pupils
Disorders	Injury or Poisoning	TRAUMA CRÂNIOCERVICAL APÓS QUEDA FRATURAS MULTIPLAS DA COLUNA TORACICA.	SKULL-CERVICAL TRAUMA AFTER FALL MULTIPLE FRACTURES IN THORACIC COLUMN
Disorders	Sign or Symptom	relata cefaléia SINAIS VITAIS ESTAVÉIS, REFERE ALGIA	reports headache STABLE VITAL SIGNS, REFERS PAIN

Fonte: OLIVEIRA, 2020

Concept annotations and dependencies



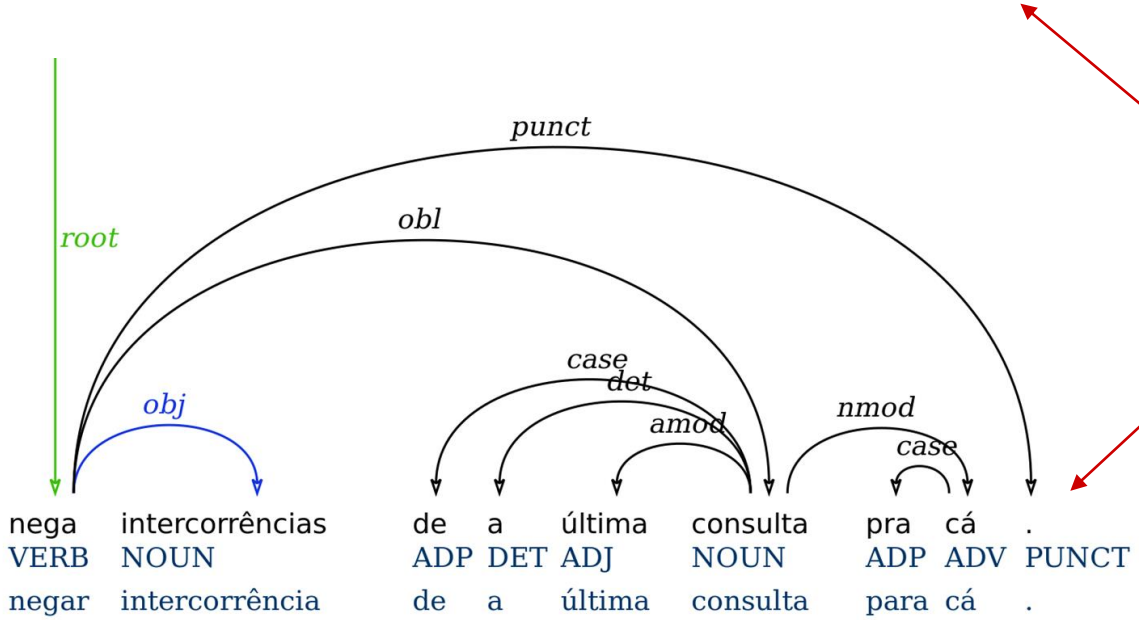
AVP >>> acesso venoso periférico STP >>> Soroterapia

(OLIVEIRA, 2020)

Negation detection and dependencies

NEGA [INTERCORRÊNCIAS DA ÚLTIMA CONSULTA] PRA CÁ

Fonte: DALLOUX et al, 2020



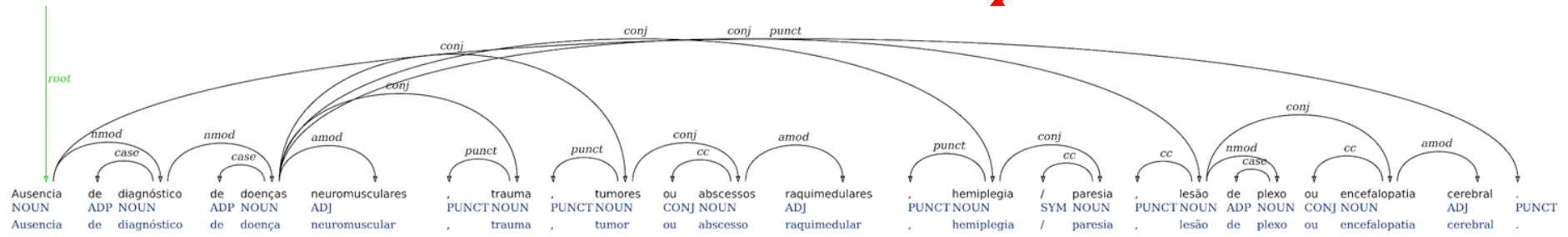
dependencies annotation solves noun phrase delimitation problem to aid negation scope detection

Negation detection and dependencies

Ausencia de diagnóstico [de doenças neuromusculares], [trauma], tumores ou [abscessos raquimedulares], hemiplegia/paresia , lesão de plexo ou [encefalopatia] cerebral .

Fonte: DALLOUX et al, 2020

dependencies annotation solves noun phrase delimitation problem to aid negation scope detection



Raw Corpus

121,465 tokens

11,065 types

Vocabulary Density: 0.091

Most frequent words in the corpus:

paciente (645)
uso (502)
nega (429)
mg (395)
d (374)
retorno (374)
pa (324)
acesso (322)
dieta (309)
dor (305)
queixas (293)
renal (291)
venoso (287)
s (285)
refere (281)

"paciente" realized in text

BUT

implicit in the majority of segments

"paciente nega" - 6 occurrences

"nega" - 423 occurrences

"paciente refere" - 34 occurrences

"refere"- 247 occurrences

Discourse actors, actions

- Patient - realized as Subject ("paciente") or implicit

pcte com rnm de crânio agendada para hoje a as 23:00h.

apresenta curativo + tala gessada em mse , referindo algia moderada , edema distal , mobilidade diminuida , apresentou 1 episodio de emese , sendo medicada , diurese presente , segue cuidados .

hidantalizado por a r1 vital brasil de a neurocirurgia , procedimento realizado sem intercorrências .

- Healthcare provider (nurse, physician) - 1st person implicit

oriento manter tala metalica a noite e retirá-la durante o dia.

converso com doutor vital brazil brandão e explico a impossibilidade de a realização de o exame hoje devido a não liberação de a guia e o fato de o paciente não trabalhar com cheque.

recebo guia com solicitação de rnm de coluna lombar.

Most frequent word: "paciente"

Term	Collocate	Count
paciente	refere	34
paciente	relata	27
paciente	interna	23
paciente	apresentou	16
paciente	consciente	15
paciente	apresenta	14
paciente	encaminhado	10
paciente	encaminhada	7
paciente	apresentando	7
paciente	retorna	6
paciente	recebe	6
paciente	queixa	6
paciente	nega	6
paciente	realizou	5
paciente	proveniente	5
paciente	diabética	5
paciente	segue	4
paciente	retornou	4
paciente	portadora	4

Corpus characteristics

words not recognized by POS taggers / lemmatizers	**Hidantalizado** **Facietomia** **flutter atrial**
extensive use of acronyms and abbreviations	**mantendo monitorização p 81, pa 103/74, sat o2 97%.** **po le + tadf em 01/12/15.** **colar cervical c/ queixas de dor**
misspelling	**em repoiso em o leito.** **outras inetrcorrências**
numerical expressions	**ssvv a as 05:45h pa = 133/74mmhg , fc = 114bpm, spo2 = 93% .** **glasgow: 9.** **mantendo monitorização p 81, pa 103/74, sat o2 97%.**
no punctuation	**dois ave um há 1 ano e outro 2 anos.**

Corpus characteristics

special use of symbols	**hma: inchaço, principalmente em o peridodo de a manha e em mmii (++++/++++), piora de o quadro ha 15 dias, estagio ii de irc.** **solicitada svd + coleta de gasometria arterial.** **# 61 # professora.** **a as 11:30hs: realizado endoscopia digestiva + broncoscopia, sem intercorrência durante o exame e transporte.** **antes 2 carteiras/dia.**
coordination	**apresenta curativo + tala gessada em mse , referindo algia moderada , edema distal , mobilidade diminuida , apresentou 1 episodio de emese , sendo medicada , diurese presente , segue cuidados .**
parenthetical comments	**refere cx em a bexiga devido a tumor (sic) em 2013.** **hmp : pais falecidos po ca (em a o soube especificar)** **controle glicêmico com glicemia em jejum abaixo de 100 (não costuma anotar).**

Corpus characteristics

reduction	**# retorno 7 dias.**
ellipsis	**apresentou problemas.** **aceitando pouco a dieta vo.**
variable segmentation	**ausculta pulmonar com mv livres.** **18:30: desacordado, faz abertura ocular espontânea, agitado.** **realizado banho de leito e higiene oral, após contido em o leito novamente devido a agitação.** **02:19: lúcida, chorosa, disfásica, em repouso em o leito, apresenta edema em região craniana, hemiparesia a a direita, afebril até o momento, sem queixas álgicas, aceita dieta via oral, diurese presente, seguem cuidados.** **no momento, consciente, orientada, calma, eupnéica, respirando em ar ambiente, comunicativa, em repouso em o leito, monitorizada, fc 99, pa 121/75, sato2 95, acesso venoso central a a direita com dopamina a 25ml/h e soroterapia em curso, uso de fralda com evacuação ausente, svd com débito amarelo claro, sem queixas álgicas, liberado dieta por a residente florence, segue sob cuidados.**

Corpus characteristics -very long segments

a as 07:30hrs: ramsay 6, pupilas iso / foto + , traqueostomizado, aspirado media quantidade de secreção de aspecto espesso, em vm modo vc (530), peep 5, fio2 40%, pa 137x75 mmhg, fc 97 bpm, fr 18 mpm, sat 100%, sne fechada, acesso venoso central em subclavia d com curativo oclusivo limpo e seco, recebendo dormonid a 20 ml/h + fentanil a 10 ml/h e noradrenalina a 12 ml/h, pam radial d, incisao cirurgica em regiao dorsal apresentando odor fetido e exsudato piossanguinolento, dreno de torax a d sem debito em o momento e com fuga aerea, curativo em local de inserção de dreno com aspecto seroso, ap diminuida em base d e roncos difusos, abdomen globoso e flacido, rha hipoativos, up sacra, svd diurese em pequena quantidade concentrada e com grumos, extremidades aquecidas, perfusao periferica <3s, mantem monitorização cardiaca, pam, oximetria de pulso e cabeceira elevada a 45°, segue sob cuidados e observação.

Pipeline


corpus preparation
automatic annotation using Stanza
human annotation in Arborator Grew
evaluation of inter-annotator agreement
case lowering
segmentation
contraction splitting
Stanza outperformed Spacy in preliminary test
web interface or Conllu file in case of very long segments
Python script

Methodology for dependencies annotation

1. Review of dependency syntax theory
2. Review of UD project & acquaintance with UD guidelines
3. Review of discussions of UD for Portuguese and treebanks
4. Review of literature on clinical text annotation
5. Search for lists of MWEs (multi-word expressions) in Portuguese
6. Corpus analysis - discourse, grammar, lexis
7. Compilation of supplementary material (clinical text abbreviations, terminology)
8. First draft of guidelines
9. Annotator training on UD dependencies and annotation software
10. Individual annotation per batch and weekly discussions
11. Guidelines update with troubleshooting
12. Assessment of inter-annotator agreement and model training for new batch
13. Evaluation of each annotated batch and fine-tuning of guidelines

Guiding principles

1. Adherence to UD version 2 guidelines and decisions in Portuguese forums
2. Alignment with decisions made in international projects on clinical text annotation
3. Alignment with approaches made in international projects on other types of text with similar features to clinical text (e.g. disfluencies and syntax in tweets)
4. Troubleshooting: solution geared towards efficient PLN tasks (NER, negation detection and time expression identification)

Guidelines for dependencies annotations

Symbol resolution

We follow published sources on clinical text

Table 2. Definition, examples and numbers of symbol senses in clinical documents.

Symbol	Clinical Corpus Sense	Definition	Example	N†
+	pulse	used in pulse degree format	pulses are 2 + bilaterally	287
	edema(swelling)	used in edema degree format	4 + brawny edema	187
	reflexes	used in reflexes degree format	2 + patellar reflexes	148
	pregnancy dating	using in pregnancy dating format	38 + 3 weeks' gestation	115
	excess	more than the given number	20 + years, 37 + weeks	68
	strength	used in strength degree format	strength of the upper extremities is 5+	52
	plus	addition between two numbers	49 + 5 cm	35
	heart murmur	used in heart murmur degree format	there was + 1 mitral regurgitation	23
	blood type	indicates antigen to blood type	a blood type A +	21
	positive (laboratory test)	react to laboratory test	blood pressures with 1-2 + protein	18
	uncommon rating*	uncommon rating	left knee has a 2+ effusion	15
	and	functions like the conjunction *and*	caltrate 600 + vitamin D1	11
	present	exist or react	+/+ trigger points	11
	fetal position during labor	position format during labor	the cervix at + 1 to +2 station	6
	tonsil size	indicates of size of tonsil	3 + tonsils	3
–	quotative, or expressing a quantity or rate	appears in quotatives, or constructions expressing a quantity or rate	5-years-old, once-in-a-lifetime	252
	compound pre-modifier	appears in compound pre-modifiers	seizure-like symptoms	226
	compound	links components of a non-modifier compound	K-Dur, x-ray, E-coli, break-through	157
	lexical hyphen	links small word formatives and content words	non-medically, ex-smoker	126
	to	indicates a range	3-4 times	111
	typographic convention	typographic-conventional hyphen or dash	allergies – none	54
	junction	notes the junction of two elements, usually vertebrae	status post C3-C4 laminectomy	24
	phone number	used in phone-number formatting	612-555-5555	13
	and (fraction)	links an integer and fraction to form a non-integer number	37-1/2 weeks gestation	10
	obstetrical data	appears in what is usually four-pronged data about a patient's pregnancy history	para 0-0-1-0	7
	hyphenated name	links two components of a hyphenated name, usually a surname	Avera-McKennan Hospital	7
	and	functions like the conjunction *and*	type II-III odontoid fracture	3
	date	used in date formatting	05-17-2003	3
	negative	indicates a negative number	-2.132	2
	line-breaking hyphen	follows the first portion of a word that is split by a line break	postopera- tively	2
	ZIP+4 code	separates a zip code and ZIP+4 code	55433-5841	1
	protocol number	serves specification function in an institution's protocol-numbering system	per our protocol #2005-02	1
	minus	indicates subtraction operation	normal 24 + or - 3 ml/kg	1
/	date	used in date formatting	05/17/2003	499
	over(e.g., blood pressure)	couples systolic and diastolic blood pressure measurements, or inhalation and exhalation with BiPAP settings	blood pressure 140/90, we will continue BiPAP at 10/5	196
	either meaning	used in constructions indicating either/both words	and/or, DNR/DNI, Heme/Onc	119
	of	separates a specific rating and the maximum value possible given the scale	regular rate and rhythm with a 2/6 systolic murmur	60
	separates two doses	indicates two separate dosages, usually in drugs with multiple drug constituents	advair 250/50	43
	divided by	separates the numerator and denominator in a fraction	1/2 day, 3-5/7 weeks	39
	per	shorthand for *per*	mg/dL	30
	abbreviation	used to abbreviate, or to link components of an acronym	OB/GYN	6
	respectively	couples values that are each respective to a distinct measure	DP and PT are 1+/4+	6
	phone number	used in phone-number formatting	612/123-4567	2
#	number	shorthand for *number*	hospital day #2	856
	quantity	indicates a quantity, usually of pills dispensed	#10 tablets, #20 dispensed	130
	gauge	indicates gauge specification	aortic valve replacement with #23 medtronic Mosaic valve	13
	level	indicates what level a measurement is at	hemoglobin at #10	1

† N = the number of samples per sense of given symbol in 1000 random samples

* Uncommon rating = subspecialty or other uncommon standard rating

Fonte: MOON, S., PAKHOMOV, S., RYAN, J., & MELTON, G. B. (2011). Automated non-alphanumeric symbol resolution in clinical texts. AMIA ... Annual Symposium proceedings. AMIA Symposium, 2011, 979–986.

Examples

# indicating text section **annotated as discourse relation**	**# s : paciente nega queixas, nega dor, dispnéia.** **# a : sepse pulmonar em d8 tazocin** **# o : paciente em beg** **# p : plano de manter atbterapia até d10**

+ indicating conjunction **annotated as cc relation**	**apresentou 1 episódio de vômito + diarréia** **apresentando tremores + pele fria** **cd: reforço de gesso + orientações + retorno em 1 semana.** **com dormonid 15 ml + fentanil 10 ml/h + noradrenalina 30 ml/h + dopamina 20 ml/h**

Treatment of ellipsis

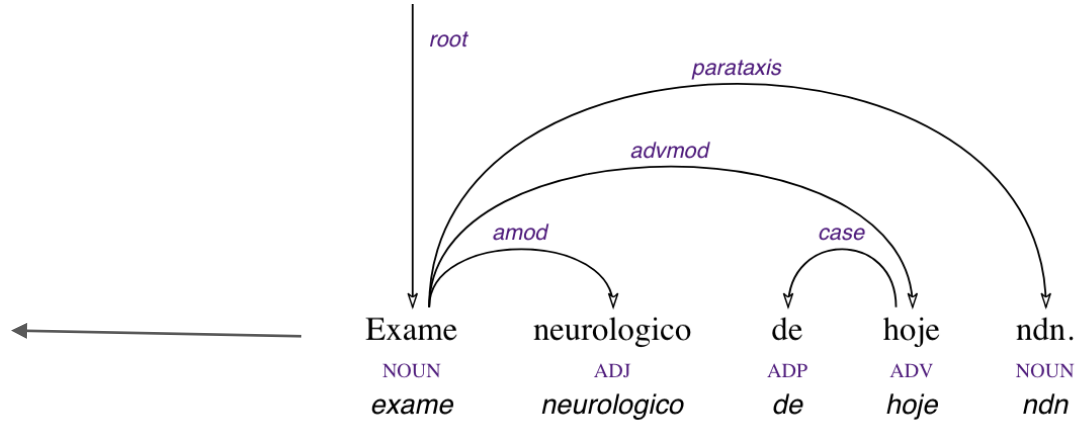
Two approaches available

1. Treating cases as typical ellipsis in UD, by promoting one of the overt dependents as head word
2. Detecting and normalizing ellipses as a preprocessing procedure

Adoption of approach No. 1

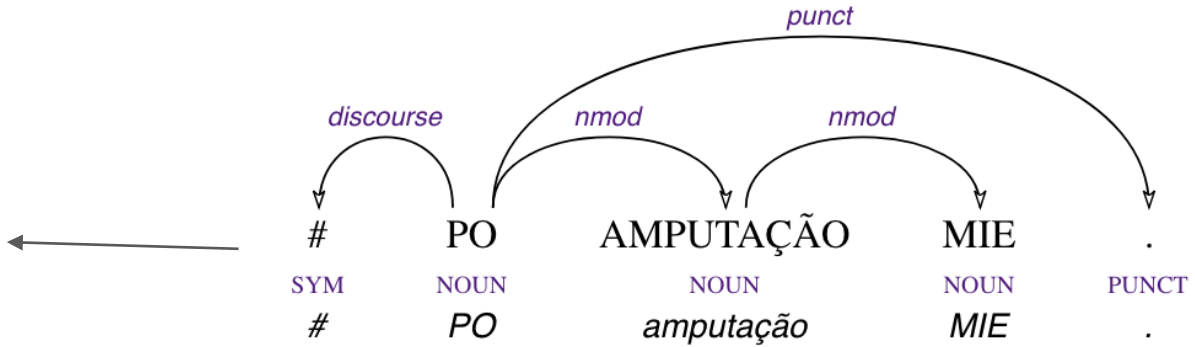
Examples

exame neurológico de o dia
de hoje : nada digno de nota



ndn>>> nada digno de nota

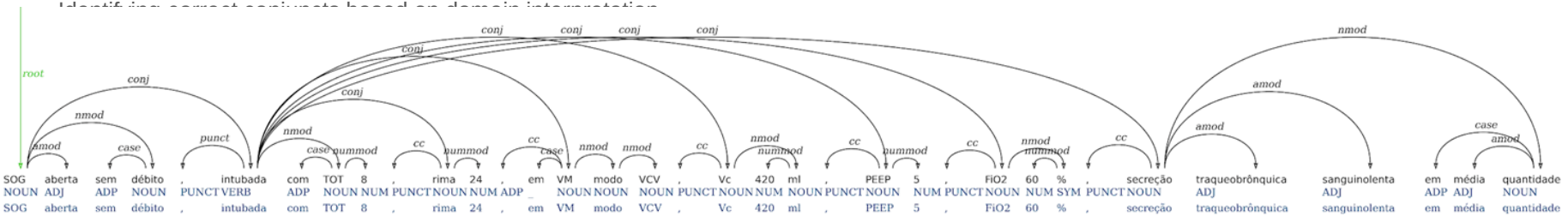
pós-operatório de
amputação de membro
inferior esquerdo



PO >>> Pós-Operatório

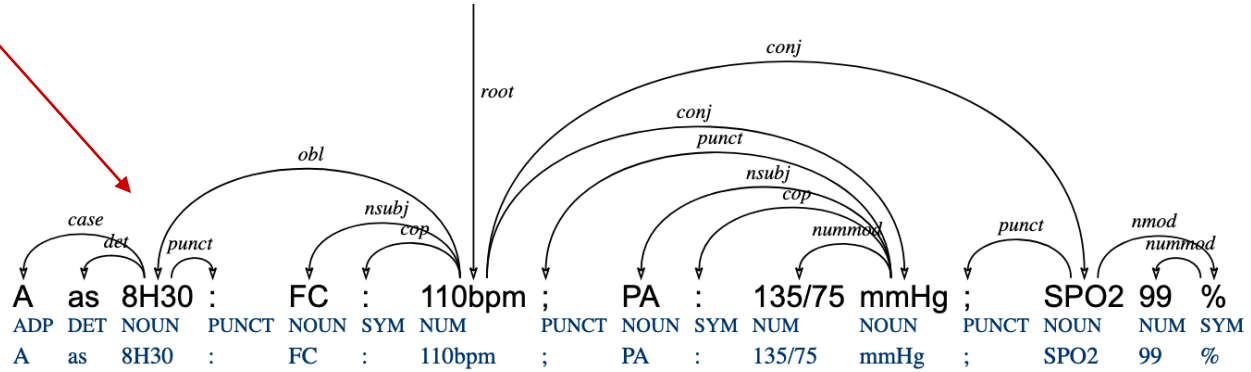
MIE >>> Membro Inferior Esquerdo

Coordination resolution

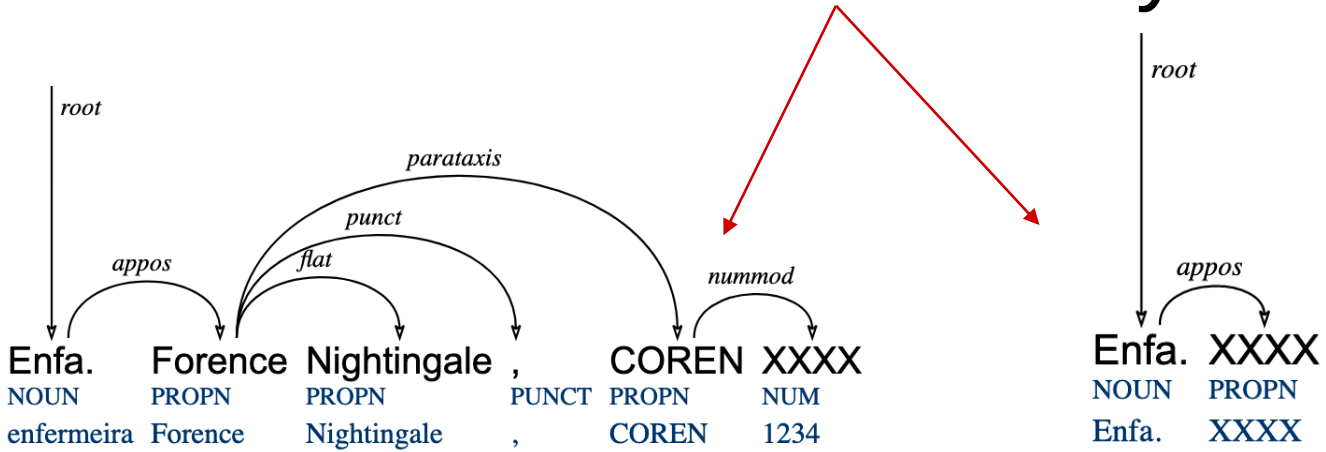


sog aberta sem débito, intubada com tot 8, rima 24, em vm modo vcv, vc 420 ml, peep 5, fio2 60%, secreção traqueobrônquica sanguinolenta em média quantidade, puncionado acesso venoso central d, com dormonid 30 ml/h + propofol 20 ml/h + noradrenalina 3 ml/h, monitorizada, estável hemodinamicamente, mv + com roncos e bolhosos difusos, pam em radial d, abdome plano, flácido, rha +, svd com diurese efetiva, evacuação ausente, extremidades quentes e bem perfundidas, higienizada, pele íntegra.

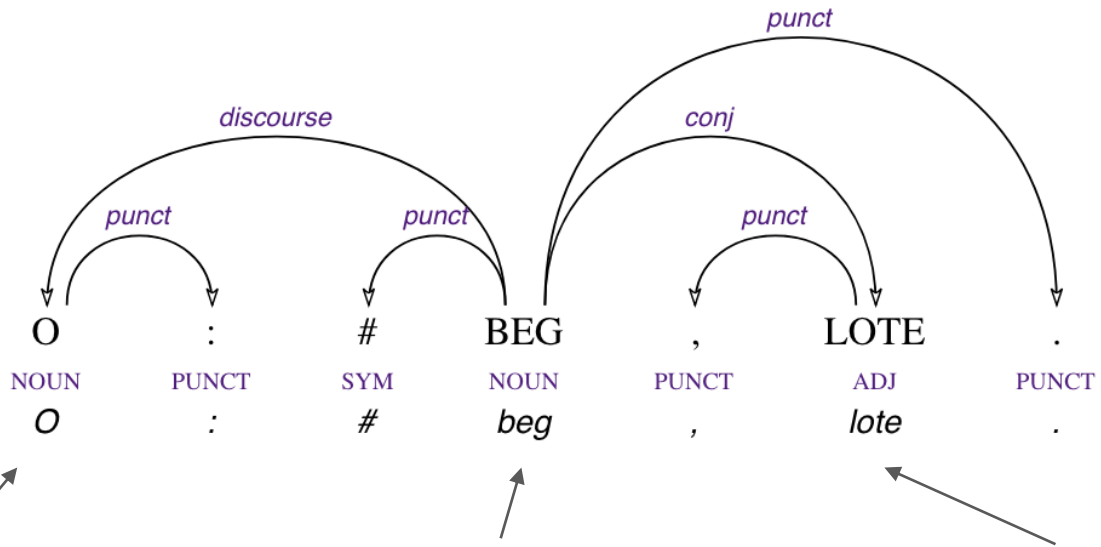
Time entries as obl dependency relations



POS & lemmatization of anonymized tokens



POS & lemmatization of abbreviations



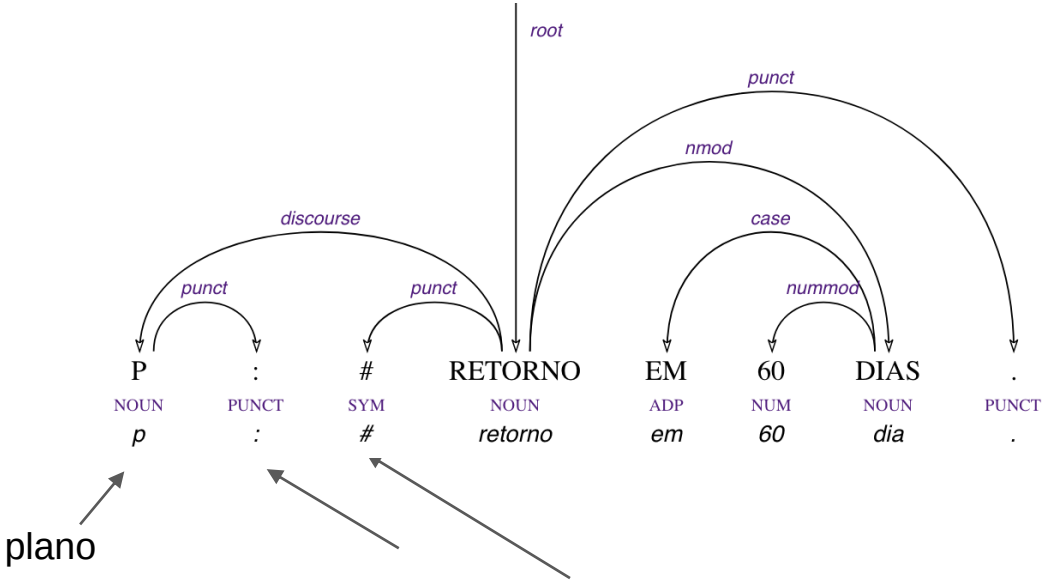
objetivo

bom estado geral

lúcido, orientado em tempo e espaço

#S → Subjetivo
#O → Objetivo
#A → Avaliação
#P → Plano

Punctuation & symbols



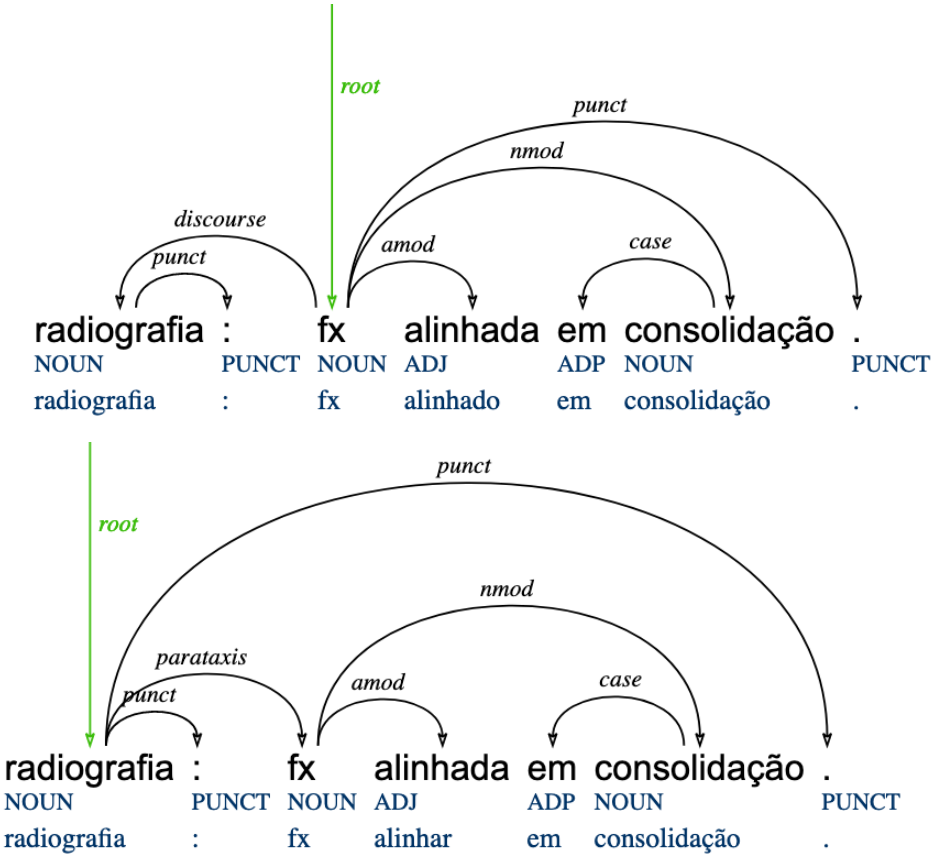
#S → Subjetivo
#O → Objetivo
#A → Avaliação
#P → Plano

punctuation

symbol

Discourse vs Parataxis

radiografia : fx alinhada em consolidação .



References

DAIBER, J., van der GOOT, R. The Denoised Web Treebank: Evaluating Dependency Parsing under Noisy Input Conditions. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), Portoro, Slovenia, European Language Resources Association (ELRA), 2016, p. 649-653

DALIANIS, H. Characteristics of Patient Records and Clinical Corpora. Clinical Text Mining. Cham: Springer International Publishing, 2018. p. 21–34.

DALLOUX, C., CLAVEAU, V., GRABAR, N., OLIVEIRA, L. E. S., MORO, C. M. C., GUMIEL, Y. B., & CARVALHO, D. R. (2020). Supervised learning for the detection of negation and of its scope in French and Brazilian Portuguese biomedical corpora. *Natural Language Engineering*, 1-21.

JIANG, Zh. et al. Developing a Linguistically Annotated Corpus of Chinese Electronic Medical Record. IEEE International Conference on Bioinformatics and Biomedicine. 2014.

KONG, L., SCHNEIDER, N., SWAYAMDIPTA, S., BHATIA, A., DYER, C., SMITH, N. A. (2014). A dependency parser for Tweets. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), p. 1001–1012, Doha, Qatar, October.

LIU, Y et al. Parsing Tweets into Universal Dependencies. Proceedings of NAACL-HLT 2018, p. 965–975.

OLIVEIRA, L. Assembling Natural Language Processing Resources To Perform The Summarization Of Clinical Narratives. PhD. Dissertation. PUC/PR, 2020.

OLIVEIRA, L. Peters, A., Silva, A, Gebeluca, C., Gumiel, Y. Cintho, L., Carvalho, D., Hasan, S., Moro, C. Semclinbr – a multi institutional and multi specialty semantically annotated corpus for portuguese clinical nlp tasks, 2020. Available from: https://arxiv.org/abs/2001.10071

SAVKOV, A., CARROLL, J., CASSELL, J. Chunking Clinical Text Containing Non-Canonical Language. In Proceedings of the 2014 Workshop on Biomedical Natural Language Processing (BioNLP 2014), p. 77–82, Baltimore, Maryland USA, June 26-27 2014.

Links

- https://grew.fr/
- http://match.grew.fr/
- https://universaldependencies.org/
- http://ucrel.lancs.ac.uk/usas/ Portuguese MWE Semantic Lexicon