Alternation-neutralized deep syntactic graphs (for French)

Marie Candito, Djamé Seddah,
LLF, Paris Diderot University
ALMANACH, Inria

joint work with
C. Ribeyre, G. Perrier, B. Guillaume

CAS Oslo
March 2018

Outline

- Deep syntactic graphs project
  - Covered phenomena: recovering shared arguments
  - Neutralizing marked syntactic alternations
  - Quantitative analysis
  - Building deep syntactic graphs
- Impact for FrameNet parsing

(For ease of reading:
examples in English in case of strong French/English parallelism)

(Sorry for duplicate talk)

Deep syntactic graphs project

Joint work with Corentin Ribeyre, Bruno Guillaume and Guy Perrier
First in FTBdep annotation scheme (Candito et al. 14; Perrier et al. 14)
More recently in enhanced UD (Candito et al. 17)

Bottom-up approach starting by (easily available) dependency trees:

- Make the most of syntactic dependency trees
- without disambiguation of predicates
- without access to lexical entries with semantic-syntactic linking

Beyond dependency trees

- Many proposals towards bilexical predicate-argument structures
  - Stanford deps (de Marneffe and Manning, 08)
  - cf. in depth analysis of 4 English graph-banks by Kulhman and Oepen (CL, 2016)
  - Semeval 2014 Shared task on "broad coverage semantic dependency parsing" (Oepen et al., 14)
  - Tectogrammatical structures in Prague dependency bank (Czech, English) (Hajič et al., 06)
  - "Deep syntax"
    - Spanish: MTT deep trees AnCora-UPF corpus (Mille et al., 13)
    - French: Deep syntactic graphs (Candito et al. 14; Perrier et al. 14)
  - Enhanced UD graphs
    - for English (Schuster and Manning, 16), French (Candito et al. 17) ...

Deep syntactic graphs: key characteristics

Aim = complete and normalize the syntactic arguments of predicates

Work on verbs and adjectives only so far

Main enhancements, concerning very well known phenomena:

- distributing shared arguments
- neutralizing marked syntactic alternations

- (comparatives)
- (by-passing morpho/syntactic markers)
- (resolving syntactic anaphora (relative pronoun antecedents))

Distributing shared arguments

“Deep” syntax: recovering shared arguments

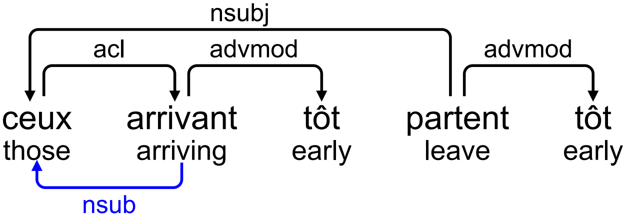
“Subjects” of non finite verbs: cases fully determined by syntax

- raising/control verbs: ***Paul*** *seems/wants to* ***sleep.***
- control nouns: ***Paul****'s desire to* ***sleep.***
- control adjectives: ***Paul*** *is ready to* ***sleep.***

“Deep” syntax: recovering shared arguments

“Subjects” of non finite verbs: cases fully determined by syntax

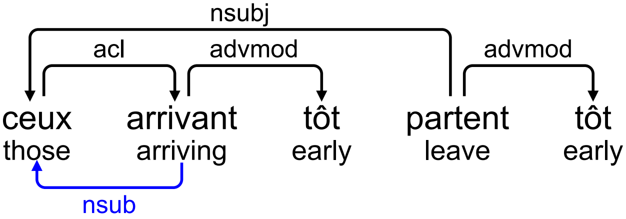
- raising/control verbs: ***Paul*** *seems/wants to* ***sleep.***
- control nouns: ***Paul****’s desire to* ***sleep.***
- control adjectives: ***Paul*** *is ready to* ***sleep.***
- noun-modifying participles: ***those arriving*** *early /* ***arrived*** *at 9am.*



"Deep" syntax: recovering shared arguments

"Subjects" of non finite verbs: cases fully determined by syntax

- raising/control verbs: ***Paul*** *seems/wants to* ***sleep.***
- control nouns: ***Paul****'s desire to* ***sleep.***
- control adjectives: ***Paul*** *is ready to* ***sleep.***
- noun-modifying participles: ***those arriving*** *early /* ***arrived*** *at 9am.*



Objects of infinitives:

- tough adjectives: ***these*** *are easy to* ***draw.***

etc ...

Distributing shared arguments

“Subjects” of non finite verbs: **cases not fully determined by syntax**

- **Example: infinitive adverbial clauses**

Paul$_{\mathbf{j}}$ *mangera avant de jouer*$_{\mathbf{j}}$
Paul$_{\mathbf{j}}$ *will-eat before to play*$_{\mathbf{j}}$
« Paul will eat before playing »

Not fully determined by syntax, but strong heuristics, e.g.:

Distributing shared arguments

“Subjects” of non finite verbs: cases not fully determined by syntax

- **Example: infinitive adverbial clauses**

Paul$_{\mathbf{j}}$ *mangera avant de jouer*$_{\mathbf{j}}$
Paul$_{\mathbf{j}}$ *will-eat before to play*$_{\mathbf{j}}$
« Paul will eat before playing »

Not fully determined by syntax, but strong heuristics, e.g.:

- When main verb is active, with non expletive subject
  $\Rightarrow$ Subject of infinitive = Subject of main verb
  in most cases (83% on French Sequoia corpus)

Distributing shared arguments

"Subjects" of non finite verbs: cases not fully determined by syntax

- **Example: infinitive adverbial clauses**

Paul$_{\mathbf{j}}$ *mangera avant de jouer*$_{\mathbf{j}}$
Paul$_{\mathbf{j}}$ *will-eat before to play*$_{\mathbf{j}}$
« Paul will eat before playing »

Not fully determined by syntax, but strong heuristics, e.g.:

- When main verb is active, with non expletive subject
  ⇒ Subject of infinitive = Subject of main verb
  in most cases (83% on French Sequoia corpus)

Counter-example:
D'autres photos ont subi des retouches pour ***accentuer*** *le drame.*
'Other photos have undergone modifications to ***accentuate*** *the drama.'*

"Deep" syntax: recovering shared arguments

Example: Arguments shared by coordinated predicates

- **Paul** *is **starving** and **wants** to **eat***

- **Paul** *is **cooking** and **selling** pancakes*
- *Paul is **cooking** and **selling*** **pancakes**

- **Paul** *is **sleeping** and **selling** pancakes*

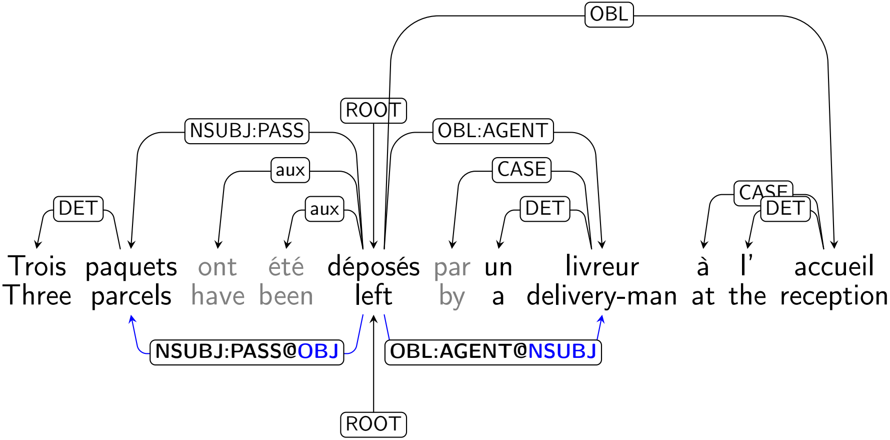
Neutralizing syntactic alternations

“Deep” syntax: Neutralizing syntactic alternations

- recover **canonical grammatical functions**
  - $\approx$ the function you would get in active personal voice
  - inspired by Relational Grammar (Perlmutter and Postal, 83)
- for French:
  - massive for passive

"Deep" syntax: Neutralizing syntactic alternations

- recover **canonical grammatical functions**
  - ≈ the function you would get in active personal voice
  - inspired by Relational Grammar (Perlmutter and Postal, 83)
- for French:
  - massive for passive


OBL
ROOT
NSUBJ:PASS
OBL:AGENT
aux
CASE
DET
aux
DET
CASE
DET
DET
Trois paquets ont été déposés par un livreur à l' accueil
Three parcels have been left by a delivery-man at the reception
NSUBJ:PASS@OBJ
OBL:AGENT@NSUBJ
ROOT

Canonical grammatical functions for passive: not so trivial

With UD labels nsubj:pass / obl:agent :
trivial replacement by obj / nsubj
(done in many works, e.g. by Reddy et al. 17)

But e.g. in French sequoia:
61% of passive forms have no direct subject (nsubj:pass/csubj:pass) in surface tree

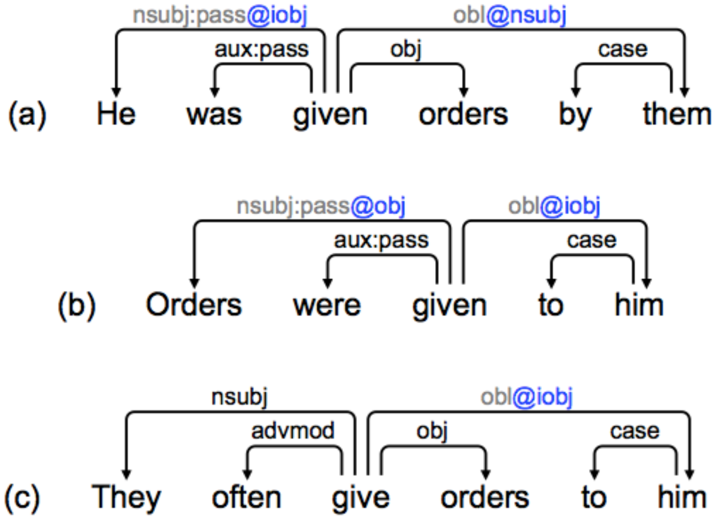
- passive reduced relative: *the dog* **chased** *by the cat*
- control/raising with passive infinitives: *they seem to be* **posted** *at fairly regular intervals*
- coordination: *I was called and* **informed** *that ...*
- coordination: *I called them and was* **informed** *that ...*

Deep syntax: Neutralizing syntactic alternations

Handled alternations for French:

- passive
  - massive (18.3% of (non auxiliary) verbs are passives in Sequoia corpus)
  - unambiguous marking
- other alternations with morpho-syntactic marking
  - marking is in general ambiguous
  - but much rarer:
    - mediopassive (0.7% of non aux verbs in Sequoia)
    - impersonal active (1%)
    - impersonal passive (0.27%)
    - causative (0.37%)
    - causative mediopassive (absent)

Detour: passive for English ditransitives



Why syntactic labels ?

- Syntactic versus semantic labels
  - semantic roles
    - patient, addressee, beneficiary ...
    - (tectogrammatical structures in Prague DT)
  - numbered arguments
    - arg0, arg1, arg2...
    - MTT: deep syntactic arguments I, II, III ...

Why syntactic labels ?

Semantic labels are sound iff linked to a semantic lexicon

- but often not the case
    - recall propbank SRL : e.g. the set of annotated ARG2 is not coherent
- oblicity hierarchy is insufficient to decide numbering
    - cf. omission
    - *Anna talked about Spinoza to her friend.*
    - *Anna talked mainly about Spinoza.*
    - *Anna talked mainly to her friend.*
    - or polysemy
    - *Anna parle italien. (A. speaks italian)*
    - *Anna parle de l'Italie à Kim. (A. talks about Italy to K.)*

Why **canonical** syntactic labels ?

Key choice: use **canonical** grammatical functions

- As a way to limit **argument linking diversity**
- Syntactic alternations
  - known to reflect semantic characteristics (cf. Levin's classes)
  - but often have strong syntactic constraints
  - exhibit regularities independently of underlying semantic roles
- $\longrightarrow$ cope with the most syntactic alternations at the syntactic level

Well-known nice regularity: “subject” of non finite verb

- Noun-modifying participle/gerund clauses: the modified noun is the (final) subject of the non finite verb
  - *people arriving late*
  - *things (being) said*
  - *people born in 2001*
  - *FR: the animals remained behind were caught*
  - *EN: the room entered into*

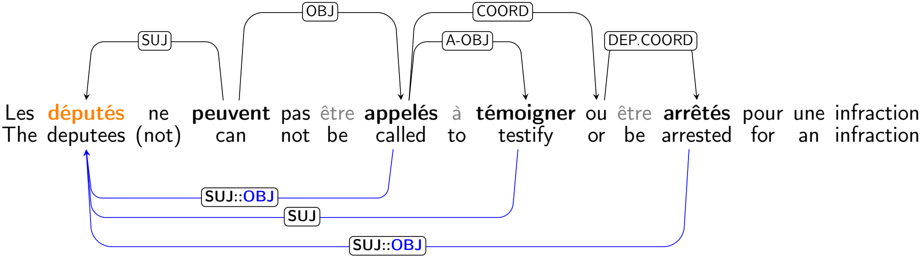
- Control phenomena: over the (final) subject of the infinitival verb
  - *He wants to be heard.*
  - rem: the controller is fixed at the semantic level
    - *He asked Paul to get up at 6*
    - *Paul was asked to get up at 6*

- Coordination of VPs:
  - *I went to this urgent care center and was blown away with their service.*

Deep syntax: interaction of phenomena

Interaction:

He wants to hear and be heard.
The room was entered into and cleaned in our absence.


Les députés ne peuvent pas être appelés à témoigner ou être arrêtés pour une infraction
The deputees (not) can not be called to testify or be arrested for an infraction
SUJ
OBJ
COORD
A-OBJ
DEP.COORD
SUJ::OBJ
SUJ
SUJ::OBJ

Note that alternation neutralization can concern deep edges!

Quantitative analysis

Quantitative analysis

Sequoia Deep syntactic graphbank

Nb Sentences	3099
Nb tokens	68802
Nb tokens ignored in deep (aux, empty preps, empty complementizers)	9338 (13.6%)
Nb full verbs (incl. cop)	6400
Nb copulas	621

Quantitative analysis

Verbs in Sequoia Deep syntactic graphbank:

	Verb **deep** mood			
	all	finite	infinitive	participle
Nb full verbs	6400	3884	1370	1146

Nb of argumental arcs				
	all	finite	infinitive	participle
all arcs	10849	5084	3606	2159
arcs only in deep	2386 (22)	693 (14)	933 (26)	760 (35)
arcs with normalized label	1942 (19)	901 (18)	360 (10)	680 (31)
arcs with normalized label only in deep	932 (39)	268 (39)	221 (24)	443 (58)

⟶ **22%** of arguments of verbs were not in surface trees

⟶ **19%** of arguments of verbs have a normalized canonical label

⟶ union of the two sets = **31%** of arguments of verbs

Obtaining deep syntactic graphs

Gold data: Sequoia corpus (3099 sentences)

- boostrapping using deterministic graph-rewriting rules applied to dependency trees
  - Grew system (Guillaume et al. 2012)
  - OGRE system (Ribeyre et al. 2012)
- adjudication of conflicts between the two systems
- plus manual checking of all non finite verbs and all coordinations
- and further tuning of the graph-rewriting rules

Obtaining deep syntactic graphs

Pseudo-gold data: deterministic rules applied to French Treebank (18 k sentences)

- Evaluation on 200 sentences shows quality is quite good (Fscore=97.7)

Obtaining deep syntactic graphs

Pseudo-gold data: deterministic rules applied to French Treebank (18 k sentences)

- Evaluation on 200 sentences shows quality is quite good (Fscore=97.7)

Annotation schemes:

- initial work on FTBdep annotation scheme
- now available for (French) UD scheme also

Obtaining deep syntactic graphs

Pseudo-gold data: deterministic rules applied to French Treebank (18 k sentences)

- Evaluation on 200 sentences shows quality is quite good (Fscore=97.7)

Annotation schemes:

- initial work on FTBdep annotation scheme
- now available for (French) UD scheme also

Deep syntax parsing:

Gold + pseudo-gold data (21 000 sent) usable as training data

- pipeline surface parsing + deterministic rules
- or direct learning of graph parser (Ribeyre, de la Clergerie and Seddah, 15)

Deep syntax for FrameNet parsing

Deep syntax for FrameNet parsing

Joint work **Olivier Michalon**, Corentin Ribeyre, Alexis Nasr
(Michalon et al. Coling 2016)

FrameNet parsing = 2 tasks (sometimes joint):

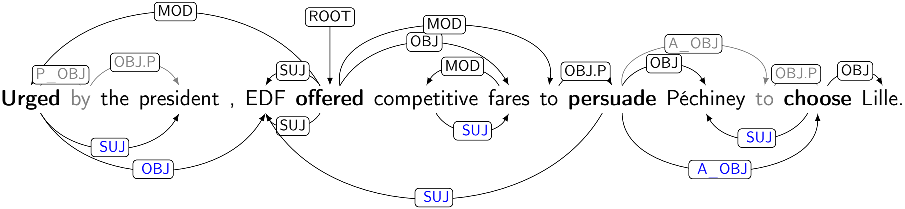
- WSD task: frame selection for an ambiguous trigger
- SRL task: role identification

- syntactic features known to be quite useful for SRL
  - since Gildea et Jurafsky, 2002
  - still true with neural networks approach
    (Hermann et al. 14; Yang and Mitchell 17)

- is it worth using deep syntax ?

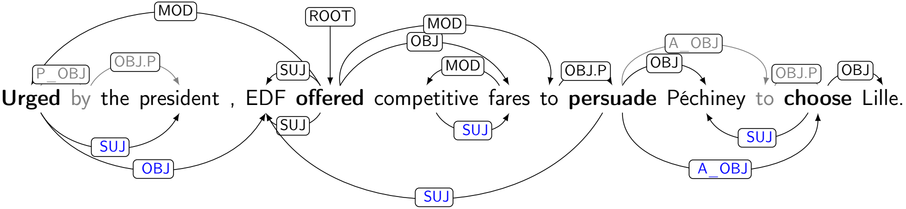
Deep syntax for FrameNet parsing

(arcs for determiners and punctuations not shown)

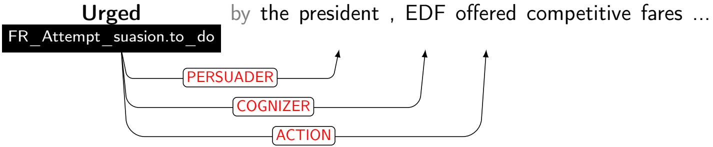

MOD
ROOT
MOD
OBJ
P_OBJ
OBJ.P
SUJ
MOD
OBJ.P
A_OBJ
OBJ
OBJ.P
OBJ
Urged by the president , EDF offered competitive fares to persuade Péchiney to choose Lille.
SUJ
SUJ
SUJ
OBJ
SUJ
A_OBJ
SUJ

Deep syntax for FrameNet parsing

(arcs for determiners and punctuations not shown)



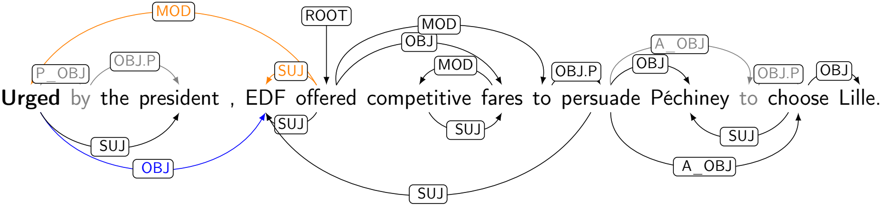
Urged by the president , EDF **offered** competitive fares to **persuade** Péchiney to **choose** Lille.

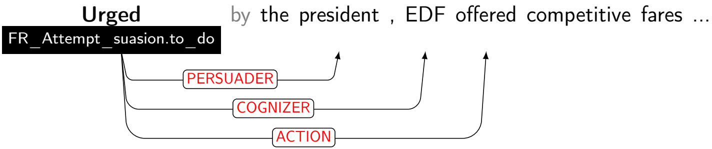


Urged by the president , EDF offered competitive fares ...

Deep syntax for FrameNet parsing

(arcs for determiners and punctuations not shown)


MOD
ROOT
MOD
OBJ
P_OBJ
OBJ.P
SUJ
MOD
OBJ.P
A_OBJ
OBJ
OBJ.P
OBJ
Urged by the president , EDF offered competitive fares to persuade Péchiney to choose Lille.
SUJ
SUJ
SUJ
OBJ
SUJ
A_OBJ
SUJ



Syntactic path between "Urged" et "EDF" :

- surface: -mod,+suj
- deep: +obj

Measuring the normalizing effect

Syntactic path between

- a predicate
- (the syntactic head) of a role filler

For a given role, deep syntactic paths are **more regular**:

Entropy of the distributions
P(path to role filler | frame-specific role)
averaged on all roles:

Measuring the normalizing effect

Syntactic path between

- a predicate
- (the syntactic head) of a role filler

For a given role, deep syntactic paths are **more regular**:

Entropy of the distributions
P(path to role filler | frame-specific role)
averaged on all roles:

- **1.65** with "surface" syntactic paths
- **1.32** with "deep" syntactic paths

⟶ **Distributions are less scattered when using deep syntax**

Measuring the normalizing effect

5 most frequent paths,
for the role fillers of verbal triggers

surface syntax		deep syntax	
(+suj)	25.0%	(+suj)	33.1%
(+obj)	17.0%	(+obj)	32.8%
(-mod)	8.0%	(+a_obj)	4.7%
(+obj,+obj.cpl)	4.4%	(-mod)	3.2%
(+a_obj,+obj.p)	4.1%	(+mod,+obj.p)	2.5%
Total	58.6 %	Total	76.2 %

Impact for FrameNet parsing

Very basic system (pipeline WSD + SRL, supervised linear classification)

- WSD : one classifier per ambiguous lemma
- SRL : one classifier per frame

Positive impact for FrameNet SRL, in particular for verbal triggers

	Prec.		Recall		F-measure	
Input syntax	surf	deep	surf	deep	surf	deep
WSD (gold frame $\neq$ Other_sense)	80.1	80.7	80.1	80.7	80.1	80.7
SRL (for gold role filler heads)	81.4	86.4	59.1	66.1	**68.5**	**74.9**

Prec.		Recall		F-measure	
surf	deep	surf	deep	surf	deep
80	80.5	80.8	80.9	80.4	80.7
75.7	80.3	51.6	59.0	**61.3**	**68.0**

Table: FastSem results for **verbs**, using **gold** (top) and **predicted** (bottom) surf and deep syntax.

Thank you

Bibliography I

Anne ABEILLE, Lionel CLEMENT et François TOUSSENEL : Building a treebank for french. In Treebanks : Building and Using Parsed Corpora, pages 165–188. Springer, 2003.

Collin F. BAKER, Charles J. FILLMORE et John B. LOWE : The berkeley framenet project. ACL '98, pages 86–90, Stroudsburg, PA, USA, 1998.

Hans C. BOAS : Semantic frames as interlingual representations for multilingual lexical databases. In Hans Christian BOAS, ed. : Multilingual FrameNets in computational lexicography : methods and applications. Trends in linguistics. Mouton de Gruyter, Berlin, New York, 2009.

Aljoscha BURCHARDT, Katrin ERK, Anette FRANK, Andrea KOWALSKI et Sebastian PADO : SALTO : A versatile multi-level annotation tool. LREC 2006, Genoa, Italy, 2006.

Marie CANDITO , Bruno GUILLAUME, Guy PERRIER and Djamé SEDDAH : Enhanced UD Dependencies with Neutralized Diathesis Alternation, DEPLING 2017, Pisa, Italy, 2017.

Marie CANDITO, Guy PERRIER, Bruno GUILLAUME, Corentin RIBEYRE, Karën FORT, Djamé SEDDAH et Eric VILLEMONTE DE LA CLERGERIE : Deep syntax annotation of the sequoia french treebank. LREC, Reykjavik, Iceland, 2014.

Marie CANDITO et Djamé SEDDAH : Le corpus sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. TALN'2012, pages 321–334, Grenoble, France, 2012.

Marianne DJEMAA, Marie CANDITO, Philippe MULLER and Laure VIEU : Corpus annotation within the French Framenet: methodology and results, LREC 2016, Portoroz, Slovenia, 2016.

Bibliography II

Charles J. FILLMORE : Valency issues in framenet. In Thomas HERBST et Katrin GOTZ-VOTTELER, eds : Valency : Theoretical, descriptive and cognitive issues, volume 187 de Trends in Linguistics. Studies and Monographs, pages 129–160. Walter de Gruyter, 2007.

Dan FLICKINGER, Yi ZHANG and Valia KORDONI : DeepBank. A dynamically annotated treebank of the Wall Street Journal. TLT'11, Lisbon, Portugal, 2012.

Jan HAJIČ, Jarmila PANEVOVÁ, Eva HAJICOVÁ, Petr SGALL, Petr PAJAS, Jan ŠTEPÁNEK, Jiří HAVELKA, Marie MIKULOVÁ, M., Zdenek ZABOKRTSKỲ, and Magda Š. Razımová : Prague dependency treebank 2.0. CD-ROM, Linguistic Data Consortium, LDC Catalog No.: LDC2006T01, Philadelphia, 2006.

Karl Moritz HERMANN, Dipanjan DAS, Jason WESTON, and Kuzman GANCHEV : Semantic frame identification with distributed word representations. ACL 2014, Baltimore, USA, 2014.

Bruno GUILLAUME, Guillaume BONFANTE, MASSON, Mathieu MOREY, and Guy PERRIER : Grew : un outil de réécriture de graphes pour le TAL. TALN 2012, Grenoble, France, 2012.

Laura KALLMEYER and Rainer OSSWALD : Syntax-Driven Semantic Frame Composition in Lexicalized Tree Adjoining Grammars. Journal of Language Modelling 1(2):267-330, 2013.

Karin KIPPER, Anna KORHONEN, Neville RYANT et Martha PALMER : A large-scale classification of english verbs. Language Resources and Evaluation, 42(1):21–40, 2008.

Anne LACHERET, Sylvain KAHANE, Julie BELIAO, Anne DISTER, Kim GERDES, Jean-Philippe GOLDMAN, Nicolas OBIN, Paola PIETRANDREA and Atanas TCHOBANOV :

Bibliography III

Rhapsodie: un Treebank annoté pour l'étude de l'interface syntaxe-prosodie en français parlé. CMLF 2014, Berlin Germany, 2014.

Igor MEL'ČUK : Dependency Syntax: Theory and Practice. The SUNY Press, Albany, NY, 1988.

Olivier MICHALON, Corentin RIBEYRE, Marie CANDITO et Alexis NASR : Deeper syntax for better semantic parsing. Coling 2016, Osaka, Japan, 2016.

Simon MILLE, Alicia BURGA and Léo WANNER : AnCoraUPF: A Multi-Level Annotation of Spanish. DEPLING 2013, Prague, Czech Republic, 2013.

Ashutosh MODI, Ivan TITOV and Alexandre KLEMENTIEV : Unsupervised Induction of Frame-Semantic Representations. NAACL-HLT Workshop on the Induction of Linguistic Structure, Montréal, Canada, 2012.

Stephan OEPEN, Marco KUHLMANN, Yusuke MIYAO, Daniel ZEMAN, Dan FLICKINGER, Jan HAJIČ, Angelina IVANOVA, and Yi ZHANG. Semeval 2014 task 8: Broad-coverage semantic dependency parsing. SemEval 2014, Dublin, Ireland, 2014.

Martha PALMER, Daniel GILDEA et Paul KINGSBURY : The proposition bank : An annotated corpus of semantic roles. Computational linguistics, 31(1):71–106, 2005.

Daniel PERLMUTTER and Peter POSTAL (eds.) : Studies in Relational Grammar 1. University of Chicago Press, Chicago, 1983.

Corentin RIBEYRE, Marie CANDITO et Djamé SEDDAH : Semi-automatic deep syntactic annotations of the french treebank. TLT 13, Tubingen, Germany, 2014.

Bibliography IV

Sebastian SCHUSTER and Christopher D. MANNING : Enhanced English Universal Dependencies: An Improved Representation for Natural Language Understanding Tasks. LREC 2016, Portoroz, Slovenia, 2016.

Djamé SEDDAH, Benoit SAGOT, Marie CANDITO, Virginie MOUILLERON and COMBET Vanessa : The French Social Media Bank: a Treebank of Noisy User Generated Content, COLING 2012, Mumbay, India, 2012.

Djamé SEDDAH and Marie CANDITO : Hard Time Parsing Questions: Building a QuestionBank for French Djamé Seddah, LREC 2016, Portoroz, Slovenia, 2016.

Josef RUPPENHOFER, Michael ELLSWORTH, Miriam R.L. PETRUCK, Christopher R. JOHNSON et Jan SCHEFFCZYK : FrameNet II : Extended Theory and Practice. International Computer Science Institute, Berkeley, California, 2006. Distributed with the FrameNet data.

Bishan YANG and Tom Mitchell : A Joint Sequential and Relational Model for Frame-Semantic Parsing. EMNLP 2017, Copenhagen, Denmark, 2017.