Building Deep Syntax Structures for User-Generated Content: Annotation Challenges in *Extreme Syntax* Scenarios

Djamé Seddah Marie Candito

Univ. Paris Sorbonne Univ. Paris Diderot
Inria Paris CNRS
Joint work with Benoit Sagot, Hector Martinez-Alonso, Vanessa Combet

CLOS Meeting, March 2018

Before gory technico-linguistics details..

Statistical Parsing of English (*our own 100 metres sprint*) has long been perceived as..

- being a very specific game played on a very specific play field

⇒ *very little lexical variation*

⇒ *very specific text genre*

⇒ *with (most often) small incremental improvement in face of the amazingly complicated technology being deployed*

- With 93% of F-score, a soon to be solved problem?

Before gory technico-linguistics details..

Statistical Parsing of English (*our own 100 metres sprint*) has long been perceived as..

- being a very specific game played on a very specific play field

⇒ *very little lexical variation*

⇒ *very specific text genre*

⇒ *with (most often) small incremental improvement in face of the amazingly complicated technology being deployed*

- With 93% of F-score, a soon to be solved problem?

The Parsing tree which hides the NLP forest

Unfortunately, that level of performance does not mean anything when it comes to Noisy User-Generated Content -or real world English, cf. #ParsingTragedy's results

Parsing UGC = Dealing with the Jabberwocky Syndrom

Lewis Caroll's Jabberworky (1872)

'Twas brillig, and the slithy toves
Did gyre and gimble in the wabe;
All mimsy were the borogoves,
And the mome raths outgrabe.

Il était **grilheure**; les **slictueux toves**
Gyraient sur l'**alloinde** et **vriblaient**:
Tout **flivoreux** allaient les **borogoves**;
Les **verchons fourgus bourniflaient**.

A mandatory deciphering exercise for most linguistics students

Parsing UGC = Dealing with the Jabberwocky Syndrom

Lewis Caroll's Jabberworky (1872)

'Twas brillig, and the slithy toves
Did gyre and gimble in the wabe;
All mimsy were the borogoves,
And the mome raths outgrabe.

- At first glance, very little in common between Lewis Caroll and User Generated Content

Parsing UGC = Dealing with the Jabberwocky Syndrom

Lewis Caroll's Jabberworky (1872)

'Twas brillig, and the slithy toves
Did gyre and gimble in the wabe;
All mimsy were the borogoves,
And the mome raths outgrabe.

- At first glance, very little in common between Lewis Caroll and User Generated Content

Sample of Google web-answer (Bies et al, 2012)

maybe they like u or they just r weird Im sorry to the person I called A freaazoid?
it is allright i guess you cooled down now, wan na be friends ??

Parsing UGC = Dealing with the Jabberwocky Syndrom

- At first glance, very little in common between Lewis Caroll and User Generated Content besides:

⇒ **out of vocabulary words** *(typos, capitalization, lexical creativity, new domains, new words)*

Jabberworky

'Twas **brillig**, *and the* **slithy toves**
Did **gyre** *and* **gimble** *in the* **wabe**;
All **mimsy** *were the* **borogoves**,
And the **mome raths outgrabe**.

Sample of Google web-answer (Bies et al, 2012)

maybe they like **u** *or they just* **r** *weird* **Im** *sorry to the person I called A* **freaazoid***?*
it is allright i guess you cooled down now, **wanna** *be friends ??*

Parsing UGC = Dealing with the Jabberwocky Syndrom

- At first glance, very little in common between Lewis Caroll and User Generated Content besides

⇒ *out of vocabulary words*

⇒ **Tokenization**

Jabberworky

'Twas **brillig**, *and the* **slithy toves**
Did **gyre** *and* **gimble** *in the* **wabe**;
All **mimsy** *were the* **borogoves**,
And the **mome raths outgrabe**.

Sample of Google web-answer (Bies et al, 2012)

maybe they like **u** *or they just* **r** *weird* **Im** *sorry to the person I called A* **freaazoid**?
it is **allright** *i guess you cooled down now,* **wanna** *be friends ??*

Parsing UGC = Dealing with the Jabberwocky Syndrom

- At first glance, very little in common between Lewis Caroll and User Generated Content besides

⇒ *out of vocabulary words, Tokenization*

⇒ **Sentence splitting**

Jabberworky

'Twas **brillig**, *and the* **slithy toves**
Did **gyre** *and* **gimble** *in the* **wabe**;**n?**
All **mimsy** *were the* **borogoves**,
And the **mome raths outgrabe**.**n**

Sample of Google web-answer (Bies et al, 2012)

maybe they like **u** *or they just* **r** *weird* **nIm** *sorry to the person I called A* **freaazoid***?*
it is **allright****n** *i guess you cooled down now,***n?wanna** *be friends ??***n**

Dealing with the Jabberwocky Syndrom

In short, parsing UGC involves working on 3 levels

- the base unit level: Tokenization
- the lexical level: Out Of Vocabulary words(OOVs) handling
- the phrase structure level: New syntactic structures

While having to cope with "some" troubling phenomena

⇒ *Crippled syntax (ie. noisy input* **Best. Workshop. Ever.**, **www.idontknow.com**, **@John seriously, dude..."***)*

⇒ *Emoticons: meta tokens or real words?*
Parsing is fun **:)** *vs* **:)** *doesn't mean it's funny*

⇒ *Not to mention mixed text encoding, multi-lingual sentences and of course, ascii art*

Dealing with the Jabberwocky Syndrom

In short, parsing UGC involves working on 3 levels

- the base unit level: Tokenization
- the lexical level: Out Of Vocabulary words(OOVs) handling
- the phrase structure level: New syntactic structures

Core principles for **annotating** such data

- Hardcore pre processing to ease the pre-annotation
- Extension of the existing annotation guidelines
- Heavy phase of multi-layer correction (sentence segmentation, MWEs, tokenization, morphology, syntax..) *some segmentation appears only on last reading*

Dealing with the Jabberwocky Syndrom

In short, parsing UGC involves working on 3 levels

- the base unit level: Tokenization
- the lexical level: Out Of Vocabulary words(OOVs) handling
- the phrase structure level: New syntactic structures

Core principles for **processing** such data

- Intensive automatic cleaning phase
- Thorough POS tagging
- The most robust parsing we can get

Is such a machinery necessary?

Main issues with most statistical parsers

- Systems with the best coverage, best overall performance **BUT**

⇒ *Extremely tied to the training material "context"*

  - genre, domain, sentence splitting and tokenization must be pretty much the same as the training corpus
  - Strong lexical sensitivity
  - *Lower out-of-domain performance*

Problems are accentuated in the case of UGC

- How to quantify them?
- Evaluate them?
- Overcome them?

Is such a machinery necessary?

Main issues with most statistical parsers

- Systems with the best coverage, best overall performance **BUT**

⇒ *Extremely tied to the training material "context"*

  - genre, domain, sentence splitting and tokenization must be pretty much the same as the training corpus
  - Strong lexical sensitivity
  - *Lower out-of-domain performance*

Toward a stress test for stat. parsing of UGC

- A new source of linguistics data
  - with a panel of attested examples
  - coming from diverse sources and the most common
  - allowing a fine grained evaluation of our tool chain

The French Social Media Banks: A set of treebanks of French as it is used in UGC

French Social Media Bank: Data Selection

Selection criteria : Doctissimo.fr

- very debatable presupposition: written fluency level is probably tied to the age of the speaker

⇒ *We wanted a large overview (including well edited text)*

- 1st Topic: Problems affecting first time pregnant women

⇒ *language level:medium*

- 2nd Topic: Birth control issues for young adolescent girls

⇒ *language level : noisy*

French Social Media Bank: Data Selection

Selection criteria: Doctissimo.fr (Exemples)

(3) a. pt que les choses ont changé depuis ?
Peut-être que les choses ont changé depuis ?
Maybe things have changed since then? **Topic 1**

b. lol vu que 2-3 smaine apres qd jai su que j'etai enceinte jetai de 3 semaine.....
Rires, vu que 2-3 semaines après, quand j'ai su que j'étais enceinte, je l'étais de 3 semaines....
Lol, given that 2-3 weeks later, when I learned I was pregnant, I was for 3 weeks... **Topic 2**

French Social Media Bank: Data Selection

Selection criteria: Doctissimo.fr (2)

- Problem: Selected texts did not contain extreme cases
- Solution: Choose texts produced without much control from the author

⇒ *Texts loaded with emotional charge*

- Subpart from sentimental and sexual distress forums

⇒ *Content extremely noisy*

French Social Media Bank: Data Selection

Exemple (suite) : Doctissimo.fr

(8) a. car je ne me senté pa desiré, pa aimé, pa bel du cou, g t pa grd chose en fet.
Car je ne me sentais pas désirée, pas aimée, pas belle du coup, je n'étais pas grand chose en fait.
Because I didn't feel desired, nor loved, thus not beautiful, I wasn't much actually

French Social Media Bank (5)

Selection criteria: JeuxVidéos.com

- Objective: Corpus with very specialized lexicon, many borrowing, many "anglicism" and a very rich vocabulary. Includes its own gestures: smileys over-presented "+1", meta-discursive elements (quote, inserted images, etc.)
- Topic : most popular threads (Call of Duty, Linux, hardware and software issues)

French Social Media Bank (5)

Example (suite) : JeuxVidéos.com

(10) a. Ces pas possible déjà que battelfield a un passe online
Ce n'est pas possible, Battlefield a déjà un pass en ligne
it's not possible, since Battlefield already has an online pass

b. je suis lvl 56
Je suis au niveau 56
I'm at level 56

c. Si y'a que Juliet &Zayn qui sont co' sur le RPG,et qui font leur vie tranquilles
Si, il n'y a que Juliet et Zayn qui sont connectés sur le jeux de rôle, makeet qui vaquent à leurs occupations
yes, There's only Juliet and Zayn connection on the role playing game and go on with their lives

French Social Media Bank (6)

Selection Criteria: Twitter

- Context: Real Time Social Media **Temps réel** archetype. Twitter does not allow a free access to its archives. Content evolves with current news, affairs, global event
- Themes : Key words linked to current events (Nov. 2011, Mars. 2014)

⇒ *At that time, difficulty to find "natural" French texts: Most of prominent tweets were from authors, bloggers or semi-professions (as opposed to the US then)*

⇒ *Difficulty to identify the informational content retweets, follow-up, hashtag being part or not of the tweet content tweet (***I love #football these days*** vs ***ManU lost!!! #football #BBC4 thesundaytimes***)*

French Social Media Bank (6)

Selection criteria (suite): Twitter

- How to find non edited tweet without biases?
- no specific thematics: random keywords (daily life objects, slang words,..)

⇒ *Here again, presupposition (prejudice?) on the expected noise level*

French Social Media Bank (6)

Example (suite) : Twitter

(13) a. Je soupçonnes que "l'enfarineuse" était en faite une cocaineuse vu la pêche de #Hollande ce soir à #Rouen.
Je soupçonne que l'enfarineuse était en fait une cocaïneuse vu la pêche de #Hollande ce soir à #Rouen.
I suspect that the "flouring-lady" was actually a cocaine-lady given the energy of #Hollande that night at #Rouen **news-based (relatively édited)**

b. @IziiBabe C mm pa élégant wsh tpx mm pa marshé a coté dsa d meufs ki fnt les thugs c mm pa leur rôle wsh
Ce n'est même pas élégant quoi, tu peux même pas marcher. à coté de sa il y a des filles qui jouent les voyous, c'est même pas leur rôle quoi. (bad translation)
It is not even elegant. One cannot even walk. Besides girls act as bullies. It is not even their role.

French Social Media Bank (7)

Selection criteria: Facebook

- Context: Social Network with controlled-broadcast. Facebook doesn't allow any access to private content.
- Goal: to Focus on open "walls" (political people, brands, celebrities) Collect various forms of French noisy text.

⇒ *Difficulties: Informative content is somewhat hidden under the mass of information of a page (status, login name, shared contents..)*

⇒ *This content is somewhat expressed graphically (**J'♡ ma 6t - Votez → (:Hollande:)**)*

- the sentence segmentation notion has sometimes very little sense. Structures close to spoken language (speech-turn, interruption, "noding"..)

French Social Media Bank (1)

Specifications

- Representative of the phenomena commonly found in NMC
- Significant size (v1: 1700 sent, v2 +3600)
- Covering almost of NMCs usages and constraints
  - Small message size: profusion of ellipsis, abbreviations, apocopes, lack of ponctuation
  - Use of a specialized lexicon: technical jargon, high unknown word rate
  - non canonical spelling (to say the least)
- Arbitrary choice of sentences: consequence of our will to depict a usage of French now common but non canonical

⇒ **the FSMB is thus not a balanced corpus.**

French Social Media Bank (2)

Data Source

- **Asynchronous: Forums, web 2.0**
  - **Doctissimo.fr**: general health forum (one of the biggest audience in France)
  - **JeuxVidéos.com**: Videos games web forums (games, platforms, general assistance). 1st in its category
- **Real Time: micro-blogging platform**
  - **Twitter**: Widely popular - (still) 140 characters limit
  - **Facebook**: *omnipresent* Social network

Noisy sub parts

- All data but JeuxVideos.com are available in 2 forms: *regularly* noisy and *super* noisy
- *noisiness* evaluated with a variant of the Kullback-Leibler divergence calculated on trigram of characters.

Linguistics of User Generated Content (3)

Lexical Phenomenon

- **non standard contractions**: **Jme** (je me/I myself-REFLX)), **lapa** (elle n'a pas/she has not..), **atu** as-tu/has-yiu, **kil** (qu'il/that he), **ct** (c'était/it was)

⇒ *cover diverse actions: bad punctuation, typographic errors, brevity oriented (apocope, abbreviation, vowel removing, etc..) or SMS language transfer (dem1 for demain/tomorrow)*

- **Lexical creativity and specialized lexicon**: Very domain dependent and very socio-demographically biased (is slang creative for its speakers?)

⇒ *Video games domain: the richest in term of creativity (borrowing + domain specific denominal verbs (lagger, fragger, headshoter, rebooter, etc..). Facebook and Twitter (noisy): most extreme cases* **(until now)** *of variance from canonical forms.*

Linguistics of User Generated Content (3)

Syntactical Phenomenon

- **oversplitting** (morpho-syntax): very frequent (quoique -> **koi ke**) especially after a contraction (c'était -> ct -> **c t** ; il a raison -> ila **ré zon** ; parce qu'il -> parcekil -> **parcek y**) ou lack of dash for MWEs (rendez-vous -> **rendez vous**)
- Prevalence of **ellipsis** on UGC, linked to the formal limit (Twitter), visual(Facebook: message display windows size) or platform media (short chat sessions between respawn).
- **Dislocated-phrases** in forums: (*le paracetamol, moi, on m'a dit que..*, **it-cleft constructions** (*c'est le samedi que ça se passe*), **imperative mood** (*redis-le doucement ?*)

⇒ *All of those are not present in our training corpora and cannot be analyzed properly*

Linguistics of User-generated Content (4)

Prevalent phenomena are characterized on two axis:

- **a encoding simplification axis**
  - **Ergographic phenomena**, whose purpose is to reduce the writing effort by diacritic removal, phonetization, spelling simplification (=? genuine typos), ellipsis (no subject, *pro-dropification*?)
  - **Transverse phenomena** such as contractions (gonna = go to), typographic diaresis (oversplitting, often after contraction))
- **Sentiment expression axis**
  - **Emulation of mark of expressiveness** via graphemic stretching, smileys, inclusion of pictures (url), capitalisation, etc..

Linguistics of User Generated Content

A Threefold Categorisation for UGC Idiosyncrasies

- **Encoding simplification:** This axis covers ergographic phenomena, reduce the writing efforts (non standard spelling and contractions, ie **“iwuz” for “I was”**) and transverse phenomena (over-splitting, **“c t” for “c’était”/it was**)
- **Sentiment expression:** This axis corresponds to marks of expressiveness, e.g., graphical stretching, replication of punctuation marks such as ???, emoticons, sometimes used as a verb such as **Je t’$<$3** standing for **Je t’aime (I love you)**. Not to mentions emojis..
- **Context dependency:** amount of context needed to understand a post. The nature of different user platforms will influence the domain knowledge necessary to understand the specific terms, from ingredients in cooking recipes to weapon characteristics in video games.

Linguistics of UGC (suite)

Most frequent phenomena (from the French Social Media Bank (FSMB)

Phenomenon	Attested example	Std. counterpart	Gloss
	Ergographic phenomena		
Diacritic removal	*demain c'est l'***ete**	*demain c'est l'été*	'tomorrow is summer'
Phonetization	*je suis* **oqp**	*je suis occupé*	'I'm busy'
Simplification	*je* **sé**	*je sais*	'I know'
Spelling errors	*tous mes* **examen** **son** *normaux*	*tous mes examens sont normaux*	'All my examinations are normal'
	Transverse phen.		
Contraction	**nimp**	*n'importe quoi*	'rubbish'
	qil	*qu'il*	'that he'
Oversplitting	**c a dire**	*c'est-à-dire*	'namely'
	c t	*c'était*	'it was'
	Marks of expressiveness		
Punct. transgression	*Joli !!!!!!*	*Joli !*	'nice!'
Graphemic stretching	*superrrrrrrrr*	*super*	'great'
Emoticons/smileys	*:-)*, <3	–	–

Annotation Scheme

Phrase-based French Treebank (Abeillé et al, 2003)

- With some modifications to ease dependency extractions and undoing of regular MWEs (FTB-UC, Candito et Crabbé, 2009)
- Extended to cope with UGC idiosyncrasies
  - **Extended POS tagset:** productive contractions (**CLS+V**, **CS+CLS**, ...), Meta tokens (**META** for Twitter's RT, **HT** for #hashtag)
  - **New annotation scheme for typographic diaeresis**: first tag is **Y**, last one is the pos of the whole *word form* (manger/**VINF** –> man/**Y** ger/**VINF**)
  - **Extended non terminal labels: FRAG** for phrases that cannot be attached to the main clause of a syntactic unit (eg RT, salutations,@mentions, etc.)

Two pre-annotation phases

Standard pre-annotation for less noisy subcorpora

- segmentation tools from the Bonsai system (set of statistical parsers for French)
- Morfette tagger (Chrupała et al 2008)
  - state-of-the-art for French, best results on known words
  - FTB-CC tagset, “FTB-UC” version (Candito and Crabbé 2009)

→ pipeline used for pre-annotating sub-corpora with a noisiness score ≤ 1

Two pre-annotation phases

Pre-annotation for high-noisiness sub-corpora

- segmentation tools from the Bonsai system (Candito et al, 2010)
- **identification of several types of "named entities" using modules from the pre-processing chain SxPipe** (Sagot and Boullier 2008)
- **noisy text normalization module**
- MElt tagger (Denis et Sagot 2009) **used on the normalized text**
  - state-of-the-art for French, best results on unknown words
  - same tagset
- **de-normalization and tag dispatching on original (noisy) tokens**

→ pipeline used for pre-annotating sub-corpora with a noisiness score > 1

Annotation process for *noisy* text

sa fé o moin 6 mois qe les preliminaires sont "sauté" c a dire qil yen a presk pa

Source Tokens	Corrected " Tokens":	corrected " Tokens" Pos-tagged reference	Pos-tags sent back to source tokens	Manual correction on source tokens Pos-tags
sa	ça	ça/PRO	sa/PRO	sa/PRO
fé	fait	fait/V	fé/V	fé/V
o moin	au_moins	au/P+D moins/ADV	o/P+D moin/ADV	o/P+D moin/ADV
6	6	6/DET	6/DET	6/DET
mois	mois	mois/NC	mois/NC	mois/NC
qe	que	que/PROREL	qe/PROREL	qe/CS
les	les	les/DET	les/DET	les/DET
preliminaires	préliminaires	preliminaires/NC	preliminaires/NC	preliminaires/NC
sont	sont	sont/V	sont/V	sont/V
"	"	"/PONCT	"/PONCT	"/PONCT
sauté	sautés	sauté/VPP	sauté/VPP	sauté/VPP
"	"	"/PONCT	"/PONCT	"/PONCT
c a dire	c'est-à-dire	c'est-à-dire/CC	c/Y a/Y dire/Y	c/Y a/Y dire/Y
qil	qu' il	qu'/CS il/CLS	qil/X	qil/X
yen	y en	y/CLO en/CLO	yen/X	yen/X
a	a	a/V	a/V	a/V
presk	presque	presque/ADV	presk/ADV	presk/ADV
pa	pas	pas/ADV	pa/ADV	pa/ADV

Annotation process for *noisy* text

sa fé o moin 6 mois qe les preliminaires sont "sauté" c a dire qil yen a presk pa

Source Tokens	Corrected " Tokens":	corrected " Tokens" Pos-tagged reference	Pos-tags sent back to source tokens	Manual correction on source tokens Pos-tags
sa	ça	ça/PRO	sa/PRO	sa/PRO
fé	fait	fait/V	fé/V	fé/V
o moin	au_moins	au/P+D moins/ADV	o/P+D moin/ADV	o/P+D moin/ADV
6	6	6/DET	6/DET	6/DET
mois	mois	mois/NC	mois/NC	mois/NC
qe	que	que/PROREL	qe/PROREL	qe/CS
les	les	les/DET	les/DET	les/DET
preliminaires	préliminaires	preliminaires/NC	preliminaires/NC	preliminaires/NC
sont	sont	sont/V	sont/V	sont/V
"	"	"/PONCT	"/PONCT	"/PONCT
sauté	sautés	sauté/VPP	sauté/VPP	sauté/VPP
"	"	"/PONCT	"/PONCT	"/PONCT
c a dire	c'est-à-dire	c'est-à-dire/CC	c/Y a/Y dire/Y	c/Y a/Y dire/CC
qil	**qu' il**	**qu'/CS il/CLS**	**qil/CS+CLS**	**qil/CS+CLS**
yen	**y en**	**y/CLO en/CLO**	**yen/CLO+CLO**	**yen/CLO+CLO**
a	a	a/V	a/V	a/V
presk	presque	presque/ADV	presk/ADV	presk/ADV
pa	pas	pas/ADV	pa/ADV	pa/ADV

Annotation process for *noisy* text

sa fé o moin 6 mois qe les preliminaires sont "sauté" c a dire qil yen a presk pa

Source Tokens	Corrected " Tokens":	corrected " Tokens" Pos-tagged reference	Pos-tags sent back to source tokens	Manual correction on source tokens Pos-tags
sa	ça	ça/PRO	sa/PRO	sa/PRO
fé	fait	fait/V	fé/V	fé/V
o moin	au_moins	au/P+D moins/ADV	o/P+D moin/ADV	o/P+D moin/ADV
6	6	6/DET	6/DET	6/DET
mois	mois	mois/NC	mois/NC	mois/NC
qe	que	que/PROREL	qe/PROREL	qe/CS
les	les	les/DET	les/DET	les/DET
preliminaires	préliminaires	preliminaires/NC	preliminaires/NC	preliminaires/NC
sont	sont	sont/V	sont/V	sont/V
"	"	"/PONCT	"/PONCT	"/PONCT
sauté	sautés	sauté/VPP	sauté/VPP	sauté/VPP
"	"	"/PONCT	"/PONCT	"/PONCT
c a dire	**c'est-à-dire**	**c'est-à-dire/CC**	**c/Y a/Y dire/Y**	**c/Y a/Y dire/CC**
qil	qu' il	qu'/CS il/CLS	qil/X	qil/CS+CLS
yen	y en	y/CLO en/CLO	yen/X	yen/CLO+CLO
a	a	a/V	a/V	a/V
presk	presque	presque/ADV	presk/ADV	presk/ADV
pa	pas	pas/ADV	pa/ADV	pa/ADV

Syntactic annotations

Classical Treebanking Architecture

- Constituent parsing done with the Berkeley parser and the Charniak parser, with gold POS supplied
- Corrected by 2 annotators (+adjudication phase)
- Followed by a functional labelling phase + correction and adjudication

Inter-annotator agreement

Doctissimo	95.05	JeuxVideos.com	97.44
Twitter	95.40	Facebook	93.40
Dcu's TwitterBank	95.8	-	-

- High agreement: Annotators were highly trained on the Sequoia Treebank (3k out-of-domain sentences, Candito & Seddah, 2012)
- In par with Dcu's TwitterBank (Foster et al, 2011) agreement score

Dependency Conversion (first results)

Classical pipeline

- Based on Candito et al (2010)'s Constituent tree to dependency conversion.
- Rely on highly optimized head-rules and an extensive knowledge of the original scheme
- Produces a native scheme (functional heads, pre-UD, relatively parsable)

Used to produce 3 dependency treebanks

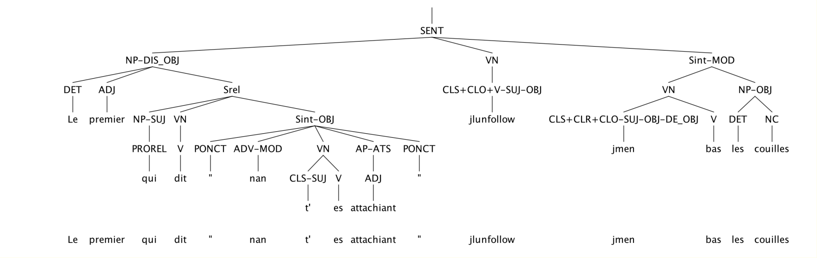
- The Sequoia treebank (Candito et Seddah, 2012)
- The FTB (Candito et al, 2010)
- The French Question Bank (Seddah et Candito, 2016)

All of them were then converted to Deep syntax graphs (cf. Marie's talk).

Dependency Conversion (The good)

(14) a. Le premier qui dit “nan, t’ es attachiant” **jlunfollow** **jmen** bas les couilles

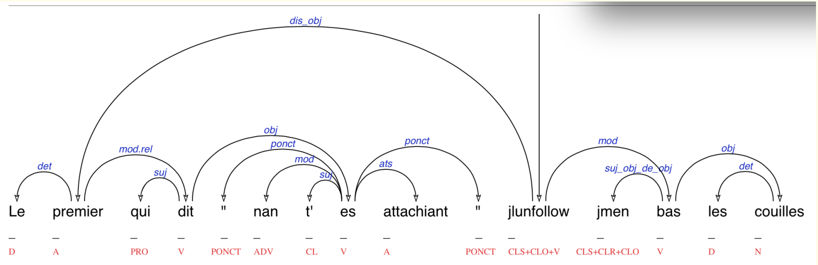
b. the first who says ”na, you’re attachnoying” **iunfollowhim** **idon - (I me it)** giv a damn



Dependency Conversion (The good)

(15) a. Le premier qui dit “nan, t’ es attachiant” **jlunfollow** **jmen** bas les couilles

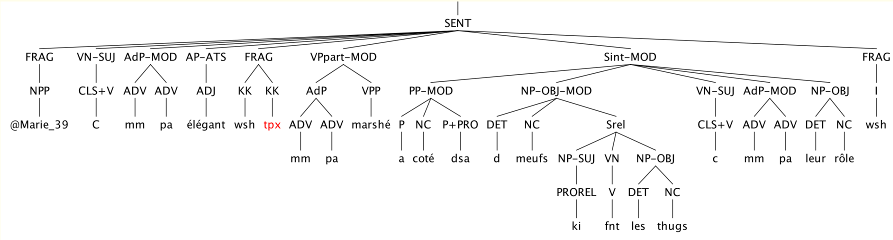
b. the first who says ”na, you’re attachnoying” **iunfollowhim** **idon - (I me it)** giv a damn



Deep Dependency Conversion (The bad)

(16) a. @IziiBabe C mm pa élégant wsh tpx mm pa marshé a coté dsa d meufs ki fnt les thugs c mm pa leur rôle wsh

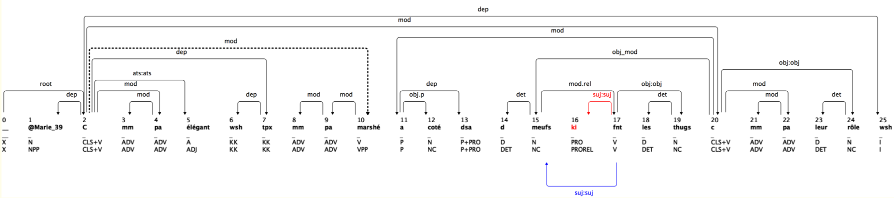
b. It is not even elegant. One cannot even walk. Besides girls act as bullies. It is not even their role.



Deep Dependency Conversion (The bad)

(17) a. @IziiBabe C mm pa élégant wsh tpx mm pa marshé a coté dsa d meufs ki fnt les thugs c mm pa leur rôle wsh

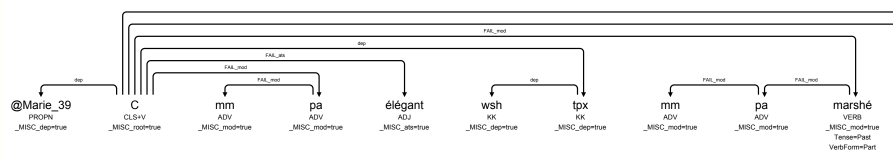
b. It is not even elegant. One cannot even walk. Besides girls act as bullies. It is not even their role.



UD Dependency Conversion (The Ugly)

(18) a. @IziiBabe C mm pa élégant wsh tpx mm pa marshé a coté dsa d meufs ki fnt les thugs c mm pa leur rôle wsh

b. It is not even elegant. One cannot even walk. Besides girls act as bullies. It is not even their role.



Dependency Conversion (Early views)

- Marie's conversion surprisingly robust (even in case of non-canonical contraction)
- Even the Deep-Syntax conversion works in some extent
- Lack of punctuations leads to no "apposition" (or "parataxis" in the UD terminology)
- problem with ellipsis or verb-less sentence
- Need a real gold standard on both native and UD scheme.

This is tough but how about the context?

The context: the most crucial aspect of social media

- Social medias broadcast conversations, reactions to events
- Analyzing posts without contexts leads to a crucial information loss
- For example, for MT, NL understanding, context-unawareness is like blind working

IMHO One of the most important point in NLP. 2 ERCs on the subject, both in MT, one tied to connected objects), many papers coming out

Symptomatic example

(@rigolboche)

ORIGINAL SOURCE	BING© TRANSLATION
→ T'as vu il l'a bien cherché wsh #AperoChezRicard	→ You have seen sought it wsh #AperoChezRicard
→ +10000, shah!	→ +10000, shah!
→ tabuz, lavé rien fé	→ tabuz, washed anything fe
→ ki ca ? le mec ou son chien ?	→ ki ca? the guy or his dog?
→ Wtf is wrong with him ? #PETA4EVER	→ Wtf is wrong with him ? #PETA4EVER
→ ki ca ? le chien ? looool	→ ki ca? the dog? looool

What kind of context would we need?

Ideally, all of it..

- The thread source (image, url, vidéo,..)

⇒ *automatic captioning*

- @mentions, entity linking, anaphora, time marks

⇒ *discourse analaysis* , *co-reference solving*

- hashtags (that can bring on another structure to the current thread)

⇒ *goto 1*

In Real Life

- we would be extremely dependent: on automatic captioning quality,
- on discourse analyze module (far, far from being solved),
- on semantic "stuff" (all of it)

Getting started: Video games live chat session

Starting small

Let's see how it works in semi-closed world scenario..

Minecraft and League of Legends

- Extremely popular video games
- Allow in-game chat sessions and of course large amount of around-the-game forums discussions are available
- LoL is a massively multi player "arena" game
- Minecraft is a sand-box game that allows players to interact in their "own world" (or to kill each other with Lego-like weapons)

the idea is to study how the language at play interacts with the surrounding context. (Highly ongoing work)

League of Legends

Minecraft


Serveur RP avec chat vocal ingame - Minecraft (w/ FireCommander & GamEdzod)
Fps 36
alpha0107
Total salary: 500.00 euros Tax Free
Civil Chuck >> LA PAYE
Civil Chuck >> x)
Civil Dorian >> YAHOU
TimFox2014 died
Civil George >> Ninja PlayerLight Toujous A l' Action !
Civil Chuck >> maisser moi sortir !!
gordon !!
Maire Gregoire >> je reviendrai vous attaquer avec mes shuri !!
[Astuce] Pour obtenir le permis d arme, il faut se rendre au
3:19 / 14:47

What kind of data are we talking about?

Corpus Properties

	# of sentences	# of tokens	Av. lenght	Std deviation	noisiness level (KL)
Marmitton	285	2080	7.30	2.57	3.43
League of Legends	453	5106	11.27	12.55	3.48
in-game	254	961	3.78	2.95	2.98
outside	199	4145	20.82	13.57	3.46
Minecraft	236	913	3.87	3.94	3.10
all	974	8099	8.32	9.38	3.58

- (Marmitton is a *noisy* part from the French QuestionBank (used as a *control* dataset)
- Huge variation in length, size, etc..
- Obviously in-game interactions are way more shorter

What kind of data are we talking about? (2)

Is it "taggable"?

		Baseline (FTB trained)		FTB trained+ Normalisation	
	OOV(%)	All	Unseen	All	Unseen
Marmitton	27.29	81.84	70.82	83.15	75.44
League of Legends	29.21	80.02	**52.92**	80.35	**45.77**
in-game chat	61.81	58.79	47.46	55.25	40.40
off-game session	21.64	84.95	56.41	86.13	60.42
Minecraft	52.57	**53.12**	**28.13**	58.27	36.04
all	31.36	77.44	52.19	78.62	45.42
FSMB (dev)	23.40	80.64	-	84.72	-
FTB (dev)	5.20	97.42	-	97.42	-

Barely. So no. Not yet.

For the record, this is where we dropped hybrid rule-based normalization. It's a dead-end when facing new domain.

What kind of data are we talking about? (3)

Can we annotate it?

- at the morphological level yes we can but very domain specific
- at the syntactic level: too many ellipsis, too many interpretations

Typology crucial problematic cases

- missing verbs (NoVerbs), conflicting predicates (Pred), parataxis (Parat)
- Code Switching (CoSwi), harmful missing punctuation (Punct)
- typographic diaeresis (Tok), non standard contraction (Cont)

What kind of data are we talking about? (3)

Qualitative analysis (random sample of 100, each)

Domain	NoVerb	Pred.	Parat	CoSwi	Punct	Tok	Cont
Lol	3	3	17	39	10	8	0
Marmitton	42	7	2	0	0	2	11
Minecraft	16	1	14	17	15	8	31

Typology crucial problematic cases

- missing verbs (NoVerbs), conflicting predicates (Pred), parataxis (Parat)
- Code Switching (CoSwi), harmful missing punctuation (Punct)
- typographic diaeresis (Tok), non standard contraction (Cont)

Pathological case

(19) a. A chaque fois des 3VS1 et du couˆ -2 P4

b. A chaque fois **il y a** des 3VS1 et du couˆ **on a** **-2** P4

c. Each time **there are** 3VS1s and then **we get** **-2** of P4

here **-2** can be **less of** or **minus 2**. P4 is an level 4 shield protection

what ?

- Ellipsis: the verbs !
- Ambiguity: what is **-2**? phonetically: "moins de" (less than) or (minus two) –> ADV P or ADV DET
- very different interpretations (diff is even more visible in constituent trees)
- interpretation very tight to the context where the interaction took place

Pathological case

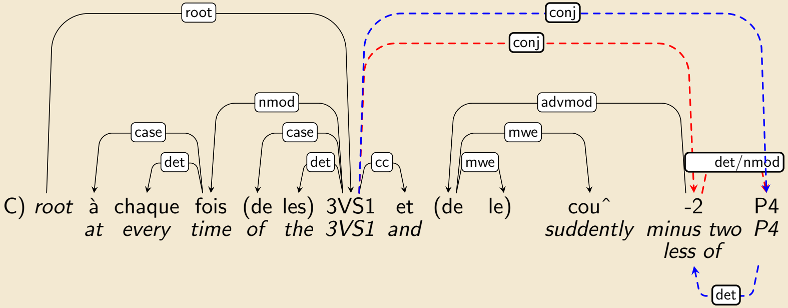
(20) a. A chaque fois des 3VS1 et du couˆ -2 P4

b. A chaque fois **il y a** des 3VS1 et du couˆ **on a -2** P4

c. Each time **there are** 3VS1s and then **we get -2** of P4

here **-2** can be **less of** or **minus 2**. P4 is an level 4 shield protection

UD analysis: two contesting structures from two different readings of the token "‘-2"



Pathological case

(21) a. A chaque fois des 3VS1 et du couˆ -2 P4

b. A chaque fois **il y a** des 3VS1 et du couˆ **on a -2** P4

c. Each time **there are** 3VS1s and then **we get -2** of P4

here **-2** can be **less of** or **minus 2**. P4 is an level 4 shield protection

So...

- The annotation scheme imposes its own view (almost normative in a sense).
- it forces us to disambiguate at all levels (tokenization, syntax)
- yet, there's no easy way to model what is crucially missing
- so, we're still working on it !

Treebanking: Let's talk about Money

Building annotated data is not only hard, it's costly

	start	Size *sent.*	morph *man/month*	syntax *man/month*	dep *man/month*	deep Synt *man/month*	cost *euros*
Sequoia	2011	3200	2	9	1	6	59k
FSMB 1	2012	1700	1	2	n/a	n/a	13k
FSMB 2	2014	2000	2	4	n/a	n/a	20k
FQB	2013	2600	2	4	1	4	36k
LoL	2015	450	3	-	-	-	3k
Minecraft	2016	230	0.5	-	-	-	2k
		10180					133k

a bit.. expensive

- 13 euros per sentence, 4 layers of annotations (so 3euros per layer per sentence. On par with LDC's costs and Fernando Perreira's experience at Google.)
- Core of the work was done by the same 2 annotators in many short terms contracts.
- We wrote guides, examples but when they left a lot of knowledge vanished. That's the most costly part. Training and getting up to speed

Thanks!

FSMB Preliminary evaluation: POS tagging

Large impact of pre-processing

	dev		test	
	MElt−corr	MElt+corr	MElt−corr	MElt+corr
Doctissimo				
high noisiness subc.	56.41	**80.78**	–	–
other subcorpora	86.57	88.42	87.78	89.18
JeuxVideos.com	81.20	82.41	82.64	83.63
Twitter				
high noisiness subc.	80.21	**84.51**	74.50	**81.65**
other subcorpora	84.09	**89.00**	86.23	**88.24**
Facebook				
high noisiness subc.	–	–	67.00	**70.75**
other subcorpora	71.75	**76.87**	78.66	**82.00**
all	*80.64*	**84.72**	*83.10*	**85.28**
Ftb (edited Text)	97.42	97.42	97.79	97.78

FSMB Preliminary evaluation: statistical parsing

Far below state-of-the-art PCFG-LA parsing on edited French

	Dev set				Test set			
	LR	LP	F1	OOVs	LR	LP	F1	OOVs
Doctissimo								
high noisiness	37.22	41.20	39.11	40.47	-	-	-	-
other	69.68	70.19	69.94	15.56	70.10	71.68	70.88	15.42
JeuxVideos.com	66.56	66.46	66.51	20.46	70.59	71.44	71.02	19.88
Twitter								
high noisiness	62.07	64.14	63.09	31.50	54.67	58.16	**56.36**	32.84
other	68.06	69.21	68.63	24.70	71.29	73.45	72.35	24.47
Facebook								
high noisiness	-	-	-	-	55.26	59.23	57.18	50.40
other	55.90	58.71	**57.27**	38.25	60.98	61.79	61.38	29.52
all	64.13	65.48	**64.80**	23.40	66.69	68.50	**67.58**	22.81
FTB ($\leq$ 40)	-	-	**86.06**	5.2	-	-	**86.16**	4.89