PARAPHRASTIC REFORMULATIONS IN SPOKEN CORPORA

IRIS ESHKOL-TARAVELLA ET NATALIA GRABAR

CNRS UMR 7270 LLL, UNIVERSITÉ D'ORLÉANS

CNRS UMR 8163 STL, UNIVERSITÉ LILLE 3

PLAN

1. Existing work
2. Work description
   - Methodology
   - Reference corpus creation
   - Evaluation
   - Analysis and results
3. Conclusion

EXISTING WORK

Paraphrase (types of transformations)

- Paraphrase refers to the utterance situation and receives contextual values (Culioli, 1976; Martin,1976; Vezin, 1976; Fuchs, 1994)
  - *two year ago/ in 2012*
- Linguistic paraphrase (Melˇcuk, 1988; Vila et al., 2011; Bhagat & Hovy, 2013)

- Morphological paraphrase
  - *We need an improvement of recycling system / We need an improved recycling system*
- Lexical paraphrase
  - There's a risk of receiving a severe wound / There's a possibility of receiving serious injure
- Semantic paraphrase
  - Emma burst into tears / Emma cried
- Syntactic paraphrase
  - The riddle is solved by him and He solved the riddle
- Mixed paraphrase

Paraphrase (size of linguistic units)

(Flottum,1995; Fujita, 2010; Bouamor, 2012)

- Lexical paraphrase
- Sub-phrastic paraphrase
- Sentence paraphrase

Paraphrase in Natural Language Processing

- *Monolingual corpora*

  (Malakasiotis & Androutsopoulos 2007, Lin & Pantel 2001, Pasça & Dienes 2005)

- *Monolingual parallel corpora*

  (Och & Ney 2000, Barzilay & McKeown 2001, Ibrahim et al. 2003, Quirk et al. 2004)

- *Monolingual comparable corpora*

  (Shinyama et al. 2002, Sekine 2005, Shen et al 2006)

- *Bilingual parallel corpora*

  (Bannard & Callison-Burch 2005, Madnani et al. 2008, Callison-Burch et al. 2008, Kok & Brockett 2010)

Paraphrastic reformulation in spoken language

- Reformulation (Gülich & Kotschi 1987, 1983; Rossari 1990, 1993)
  - written / spoken
  - different functions
  - types of markers
    - Markers of non-paraphrastic reformulation
      - *en tout cas (anyway), enfin (well)*
    - Markers of paraphrastic reformulation (MPRs)
      - *ça veut dire (which means), en d'autres mots (in other words)*

WORK DESCRIPTION

Difficulty of the task

- Spoken corpora
  - absence of typographic markers of segmentation in the transcriptions files
  - oral disfluencies
- Ambiguity of studied MPRs
- Discontinuity of paraphrastic relation (PR)
  - it cat appear in more than one turn of speech (ToS)
  - segments in PR can be adjacent or distant

MPRs :

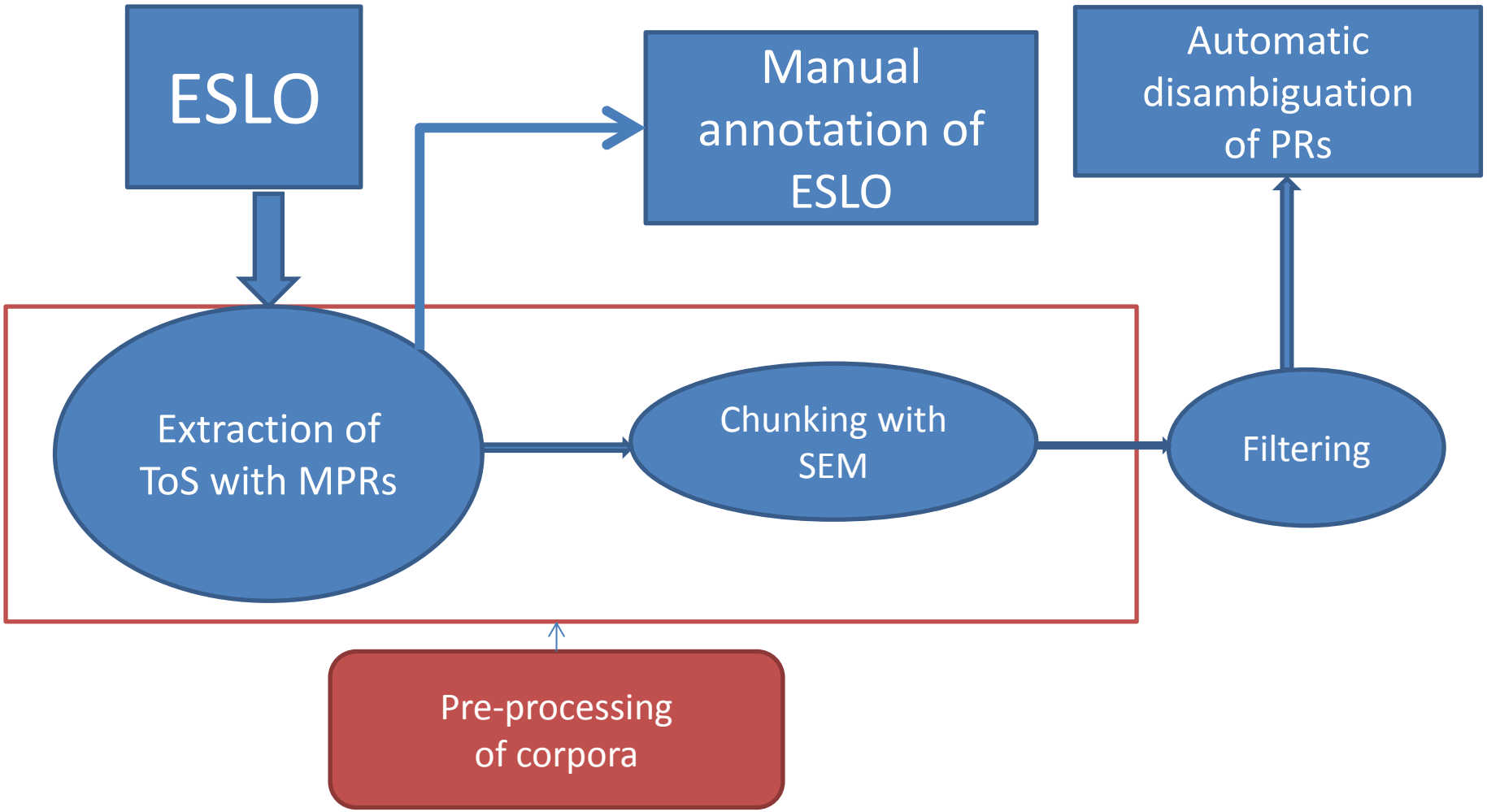
c'est-à-dire (in other words), *je veux dire* (that is to say, I mean), *disons* (let's say)

Segment1 MPR Segment2

Corpus	original ESLO1	original ESLO2
Nb interviews	260	308
Nb words	2 349 829	1 412 891

Corpus	processed ESLO1	processed ESLO2
Nb interviews	54	30
Nb ToS	476	394

Methodology



Guidelines for manual annotation

Multidimensional annotation

Syntactic information (*N, NP, VP, ADV*, etc.)

Attributes

- *rel-lex, modif-lex, modif-morph, modif-synt,*
- *rel-pragm*
  - addition of new information : explic, préc, exempl, déf
  - equivalence
  - result (summary)

*euh <VP1>**démocratiser l'enseignement**</VP1> <MRP>**c'est-à-dire**</MRP> <VP2 rel-lex="syn(démocratiser/permettre à tout le monde) syn(enseignement/faculté)" modif-lex="ajout(rentrer à)" rel-pragm="explic">**permettre à tout le monde de rentrer en faculté**</VP2>*

[ESLO1_ENT_121_C]

*démocratiser l'enseignement (***democratize the education***),*

*permettre à tout le monde de rentrer en faculté (***allow everybody to enter the university***)*

Rules for automatic disambiguation

- Position of MPRs in ToS
- Presence of oral disfluencies
- Specific lexical contexts

  nous disons (we say), *par contre / mais je veux dire* (but I want to say)
- MPR occurs within existing expression or phrase

est-ce que vous remarquez une différence sensible entre vos différents clients dans leur façon de choisir la viande dans ce qu'ils achètent et caetera ***indépendamment*** *<MRP>****disons****</MRP>* ***de leurs origines*** *de classe*

[ESLO1_ENT_001_C]

indépendamment de (independently of)

Evaluation

- Manual annotation

	ESLO1					ESLO2				
	A1		A2		Agr.	A1		A2		Agr.
	yes (%)	no (%)	yes (%)	no (%)		yes (%)	no (%)	yes (%)	no (%)	
c'est-à-dire	96 (**33**)	193 (**67**)	66 (**23**)	223 (**77**)	249	74 (37)	124 (63)	65 (32)	137 (68)	162
je veux dire	16 (25)	49 (75)	8 (12)	57 (88)	57	47 (34)	91 (66)	27 (20)	110 (80)	107
disons	18 (15)	104 (85)	8 (7)	115 (93)	106	10 (18)	45 (82)	9 (16)	46 (84)	46
Total nb of ToSs	130 (27)	346 (73)	82 (17)	395 (83)	412	131 (33)	260 (67)	101 (26)	293 (74)	315

- *C'est-à-dire* is the most grammaticalized
- The inter-annotator agreement is substantial in ESLO1 (0,617) and moderate (0,526) in ESLO2.

Analysis and results (1)

- average size of utterances with MPRs :
  - 62,88 in *ESLO1* and 86,34 in *ESLO2*
- frequency of MPRs
  - *c'est-à-dire* is the most frequent
  - *disons* is very frequent in *ESLO1* but less frequent in *ESLO2*
- MPRs grammaticalization

Analysis and results (2)

- In over 70% of contexts, there is **no syntactic equivalence** between the entities in PR
- Very **few formal cues** are available for the detection of paraphrases
  - **morphological modications**
  - **syntactic modifications**

*<P1>**le mouvement scout de France a décidé de euh diviser les tranches d'âges**</P1> <MDR>**c'est-à-dire**</MDR> que moi je vous ai dit tout à l'heure nos nos enfants ont entre huit et douze ans et douze à dix-sept ans <P2 **modif_synt**="passif(a décidé de euh diviser les tranches d'âges/cette tranche a été divisée)" **modif_morph**="flex(diviser/a été divisée)" rel_pragm="exempl">**cette tranche douze dix-sept a été divisée en deux par les scouts U-les scouts de France en douze quinze et quinze dix-sept**</P2>*

[ESLO2_ENT_5_C]

le mouvement scout de France a décidé de diviser les tranches d'âges **(Scouts in France have decided to distinguish age groups)**

cette tranche douze dix-sept a été divisée en deux par les scouts de France en douze quinze et quinze dix-sept **(the ten seventeen group has been devided in two by the Scouts in France in twelve fifteen and fifteen seventeen)**

- Among the **lexical relations**, synonymy and hyperonymy are the most frequent, followed by instances with named entities.

Analysis and results (3)

- we can find several paraphrastic reformulations in which entities have no semantic relation except the one marked by the MPR,

des conférences *y en a assez souvent sur France culture enfin <MPR> disons </MPR>* ***des causeries***

[ESLO1_ENT_121_C]

*des conférences (***conferences***)*

*des causeries (***chat, natter***)*

Evaluation

- Automatic detection

	ESLO1		ESLO2	
	A1	A2	A1	A2
lexical and discursive filters	40,5	40,5	37,7	37,8
lexical and discursive filters + frequency (>6000)	25,8	25,9	18,7	18,9
lexical and discursive filters + priority frequency (>6000)	63,0	63,0	66,4	66,3

CONCLUSION

Some observations

- reformulations are not always paraphrastic
- MPRs
  - do not always introduce PRs
  - can create PR between entities that do not show semantic equivalence otherwise
- syntactic equivalence: 30 % of contexts
- very few formal cues are available for the detection of paraphrases

Originality of our approach

- detection of paraphrastic reformulations in monolingual spoken corpora in French
- taking into account the specificity of the spoken data
- syntagmatic approach using
- manual multidimensional annotation
- large acceptance of paraphrase

In perspective

- Enrich with new data
  - involve additional human annotators
  - study other MPRs
  - use paralinguistic information
  - add sociological criteria about speakers
- Improve automatic detection
  - use the machine learning
  - detect PR between different ToSs
- Perform new analysis
  - compare PRs in spoken and written corpora
  - analyse MPRs from diachronic point of view

THANK YOU !!!