Acquiring Verb Frames for a Text Simplification Lexicon in the Medical Domain

Ornella Wandji Tchami
TKE 2016, Copenhagen

STL UMR 8163 CNRS, Université Lille 3, 59653 Villeneuve d'Ascq, France
IwiSt, University of Hildesheim, Universitätsplatz 1, 31141 Hildesheim, Germany

Table of Contents

Context and Problem

- Standard medical language is sometimes hard to understand for non-expert users [McCray, 2005; Zeng-Treiler et al., 2007]
  - linguistic complexity [Putz, 2008]
  - specific terminology
  - specialised phraseology [Wandji et al., 2014, 2015]

  $\Rightarrow$ communication problems between medical experts and non-experts

$\Rightarrow$ Text simplification:

a means for making specialised medical texts accessible to the non-experts [Zeng-Treiler, 2006; Chmielik Grabar, 2011; Siddharthan, 2014]

- **Need of resources**

Research Goal

- Aim of our PhD project: Creation of a resource of verb frames and cooccurrences for the simplification of specialised medical texts
  - $\Rightarrow$ medical frames of verbs aligned with their lay equivalents

Expert	**(specialised)** verb usages	Lay equivalent usages
Le PATIENT (S)	**subit** une MALADIE (D) **développe** **relève** d'	**a** **fait** **souffre** d'
PATIENT	develops/suffers (from) DISEASE	has
Le MEDECIN	**diagnostique** le PATIENT	**dépiste** (D chez S)
DOCTOR	diagnoses PATIENT	detects (D on S)

- Method inspired by Frame Semantics (Framenet [Ruppenhofer et al., 2006])
- But different approaches:
  - bottom-up vs. top-down
  - Snomed semantic categories vs. semantic roles

Research Goal

- Aim of this study: proposing a method
  for the acquisition of specialised medical frames of verbs
  - Selection of frames based on:
    a syntactico-semantic classfication of verbs [Le Pesant, 2007]
  - Syntactic analysis of sentences from
    three medical sub-corpora of non-aligned texts
  - Semantic annotation using an existing med. termin. [Côté, 1996]

Material

Corpus

- Three medical sub-corpora differentiated according to the level of expertise of their authors and intended readership
- Similar sizes

Subcorpus	Size (words)	Description
expert to expert texts	1,785,665	scientific publications, reports
expert to student texts	1,755,497	didactic supports for students
expert to lay texts	1,627,466	documentation, brochures

- Source: CISMEF portal
  - indexing of French medical documents
  - according to different categories:
    for patients, for students, for medical professionals

Material

Verb Resource

- An electronic version of *Les Verbes français by Jean Dubois and Françoise Dubois-Charlier* [Dubois, 1991]:
  - A syntactico-semantic classification of French verbs based on their valency patterns: verb classes are defined by syntax
  - 25 610 entries:
    12 310 single entries
    4118 polysemous
  - different types of information separated by a tabulation:
    entry: subir 03
    domain: MED
    syntactic class: D3e
    valency pattern+sem restrictions:
    T1300 (transitive: subject=human, object=thing)
    meaning (synonym, definition, or explanation):
    supporter, se soumettre à
    sentence: On subit une intervention, des tests

Material

Medical Terminology

- Snomed International terminology [Côté, 1996]

  11 semantic classes, 9 used:

$\mathcal{T}$: Topography or anatomical locations (*e.g.*, *coeur, cardiaque, nez*);

$\mathcal{S}$: Social status (*e.g.*, *mari, soeur, enfant*);

$\mathcal{P}$: Procedures (*e.g.*, *césarienne, télé-expertise*);

$\mathcal{L}$: Living organisms:

- bacteries (*e.g.*, *bacillus, enterobacter*),
- animals (*e.g.*, *chien, porc, chat*);
- plants (*e.g.*, *aristide, herbe à épée*);

$\mathcal{J}$: Professional occupations (*e.g.*, *équipe de SAMU, anesthésiste*);

$\mathcal{F}$: Functions of the organism (*e.g.*, *pression artérielle, détresse*);

$\mathcal{D}$: Disorders and pathologies (*e.g.*, *obésité, cancer, maladie*);

$\mathcal{C}$: Chemical products (*e.g.*, *médicament, sodium*);

$\mathcal{A}$: Physical agents and artefacts (*e.g.*, *tubes*, *prothèses*).

Method

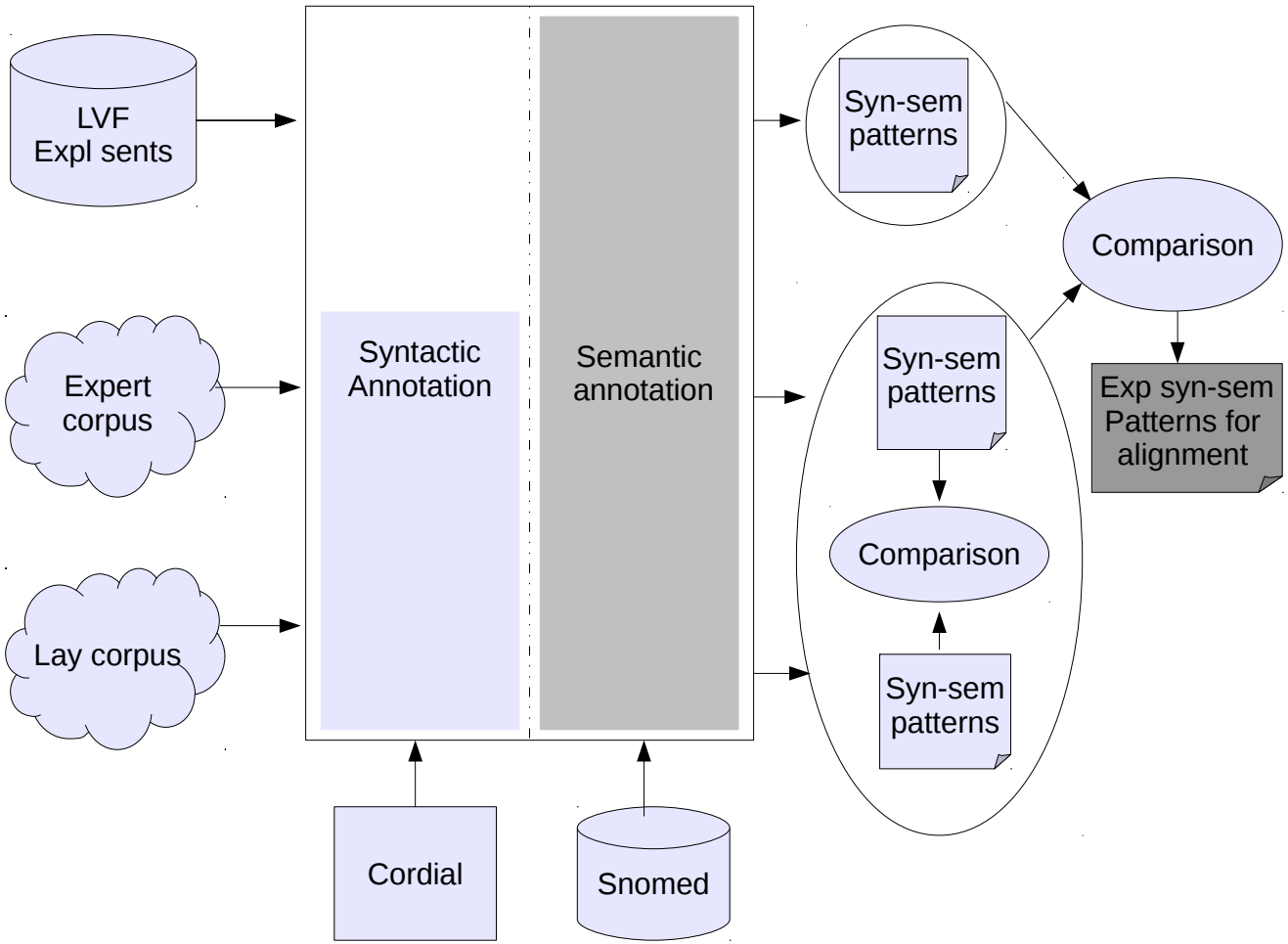
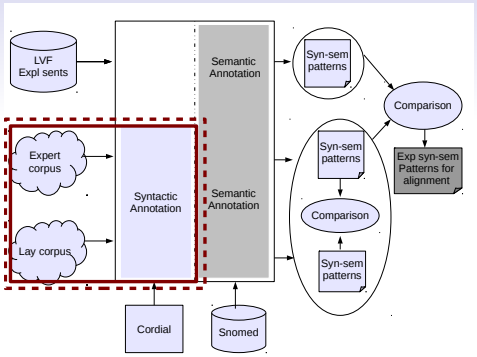


Figure : General schema of the method

Method

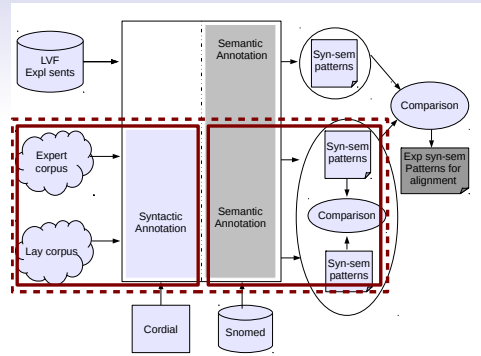
Corpus Pre-processing and Selection of Verbs



- Parsing with the Cordial dependency parser [Laurent et al., 2006]
  - tabulated format similar to the CoNLL format [Buchholz Marsi, 2006]
  - identification and retrieval of verbs and their arguments
- selection of verbs: 2 criteria
  - being part of the LVF's entries,
  - having
    at least 30 occurrences in the corpus (cumulated frequency)

Method

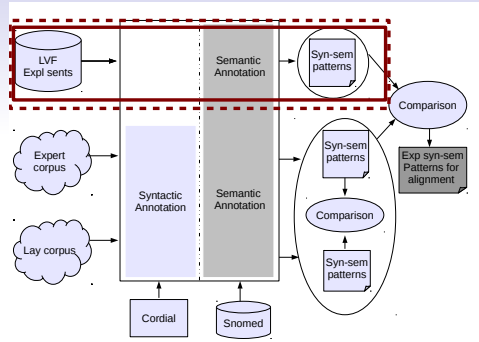
Semi-Automatic Acquisition of Frames from the Corpus



- Extraction of medical frames of the verbs using an automatic method [Wandji et al., 2015]
  - Snomed international terminology
    - projection of the terminology onto the corpus sentences (nominal chunks)
    - labelling of the terms with the corresponding Snomed categories
- Manual improvement of the semantic annotation
  - annotation of terms not covered by Snomed
  - correction of erroneous labels

Method

Semantic Annotation of Sentences and Acquisition of Frames from the LVF



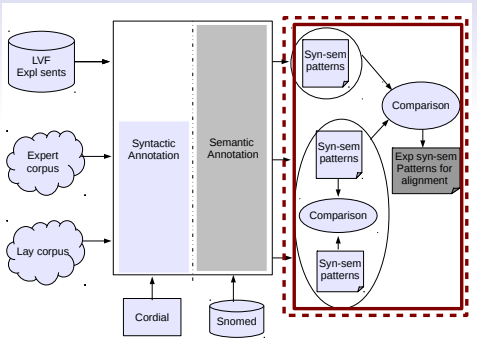
- Automatic retrieval of the sentences provided as examples in medical readings of LVF verbs
  - Argument positions instantiated by words that stand for the selectional restrictions of verbs
  - Generic medical terms, which match with Snomed entries
    e.g. *Le* **médecin** *administre un* **remède** *au* **malade**.
- Semantic annotation of the sentences and acquisition of frames
  - Same method as with the corpus (using our automatic system [Wandji et al., 2015])

    e.g. *Le médecin administre un remède au malade*
    (The doctor administers a medication to the patient.)

    $\Rightarrow$ s_J administrer cod_C coi_S
    s_J administer cod_C coi_S

Method

Comparison of Frames and Selection of Entries for the Text Simplification Resource

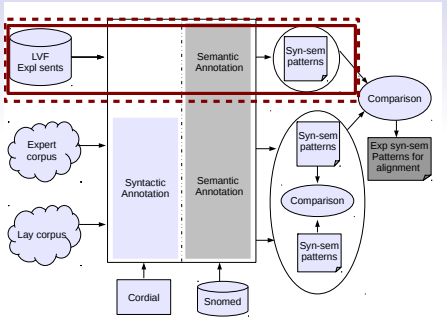


- For each verb: comparison of the corpus frames with those acquired from the LVF sentences
  - Qualitative comparison:
    - The number of corpus frames vs LVF frames
    - The number of common frames vs specific frames
      Frames covered only by the LVF
      Frames covered only by the corpus (corpus-specific)
  - Quantitative comparison
    - Frequency of each frame in the different sub-corpora

    $\Rightarrow$ On the basis of these figures, we select verb readings whose frames can constitute entries for our final resource

Results and discussion

Sentence Extraction and Automatic Acquisition of LVF Frames



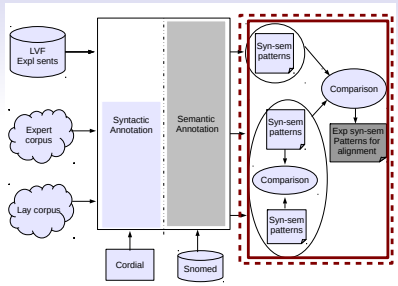
Step	Number
0. Sentence extraction	318
1. Generation of frames	420
2. Verbs covered by the corpus	288
3. Verbs selected for detailed evaluation	11

- Detection of verb polysemy based on the acquired frames: 1 LVF sentence example ⇒ 2 verb readings

  e.g. *On apaise un malade, la douleur*
  On apaise un malade ⇒ On apaise S Someone **relieves/appeases** a patient
  On apaise la douleur ⇒ *On apaise F* Someone **alleviates/eases** the pain

Results and discussion

Qualitative Comparison of Frames: Corpus Frames vs LVF Frames

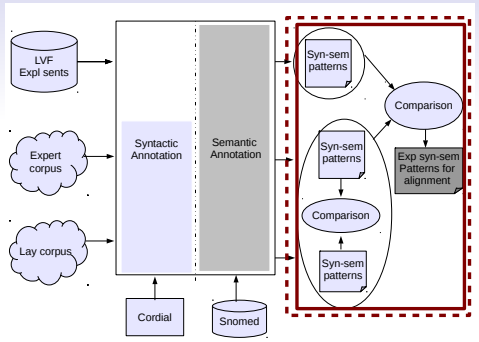


Selected verbs	LVF fram	Nb com_fram	Occ. com_fram	Corpus fram	Occ. corpus fram
abaisser	1	0	0	19	55
admettre	1	1	1	33	74
diagnostiquer	1	1	16	30	101
imposer	1	0	0	66	388
indiquer	1	1	3	**193**	768
relever	1	1	2	78	183
subir	1	1	38	43	92
suivre	1	0	0	142	488
survivre	2	1	3	**14**	31
stimuler	1	1	5	39	76
traiter	1	1	1	107	297

- Almost all the LVF frames are covered by the corpus
  - Exception: 3 frames

Results and discussion

Qualitative Comparison of Frames: Corpus Frames vs LVF Frames



- LVF normalises verb usages to active form formulae
  ⇒doesn't keep track of passive form preferences

  s_J **suivre** *do_S* (*s_J monitor do_S*)
  Found only in the LVF appears several times (10) in the corpus, but in the passive voice

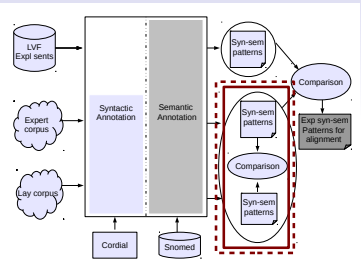
  e.g. *Les patients porteurs d' un défibrillateur doivent être suivis par les médecins du centre où a été implanté le défibrillateur.*
  (Patients with a defibrillator should be monitored by the doctors of the center where the defibrillator was implanted.)

- Corpus offers sets of specific frames: minimum 14 for individual verbs
  - Fine semantic granularity of the Snomed categories
  - Syntactic variation of frames

  ⇒ Potential candidate entries for the text simplification resource

Results and discussion

Quantitative Comparison of Frames across the Subcorpora



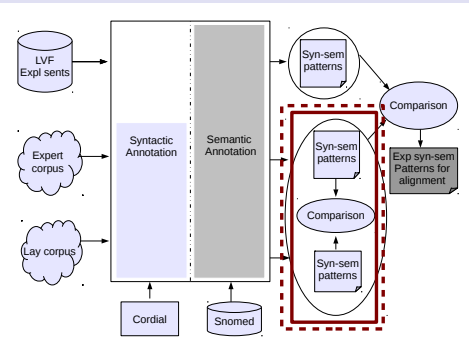
Verb frames	Corpus	Exp	Stu	Lay
s_J admettre cod_S coi_S	1	1	0	0
s_J diagnostiquer cod_D	**16**	**2**	**1**	**13**
s_T indiquer cod_F	3	1	2	0
s_On relever coi_D	2	1	1	0
s_F stimuler cod_F	5	1	2	2
s_On survivre coi_D	3	1	1	1
s_On subir cod_P	**38**	**4**	**17**	**17**
s_J traiter cod_S	1	0	1	0

- 8 LVF frames attested in the corpus, for the 11 sample verbs
- LVF frames with a low token frequency due to syntactic variation in the corpus

  s_J diagnostiquer cod_D: 3 variants
  s_J diagnostiquer cod_D chez coi_S
  s_D être-diagnostiqué par/chez coi_S (passive form)
  s_D être-diagnostiqué chez S (passive form with an omitted agent)

Results and discussion

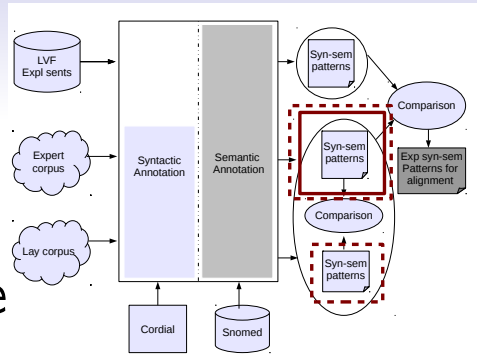
Qualitative Comparison of Frames across the Subcorpora



- Some of the LVF frames are frequent in the corpus while others are not
- The frequency variation (of frames across the subcorpora) is indicative (*diagnostiquer*, *subir*)
  - High frequency in the expert subcorpus:
    specialised meaning of the verb: experts' vocabulary
  - High frequency in the lay subcorpus:
    verb usage that is more common to the lay persons
  - Low frequency sometimes:
    - ⇒ hides highly specialised usages of the verbs (s_S relever coi_D)
      e.g. *[...]Lorsque le patient relève d' une affection de longue durée.*
      (when the patient suffers from a long-term illness.)
    - ⇒ indicates the corpus preference with respect to verb meanings (see *indiquer*, *suivre*, *traiter*)

Results and Discussion

Corpus-specific Frames of the Studied Verbs (expert subcorpus)



- Cases with f>10 occ: exploitable corpus-specific frame

Frames	Freq	Frames	Freq	Frames	Freq
s_F est abaissé	21	s_D imposer cod_P	49	**s_S subir cod_D**	**16**
s_S est admis coi_S	20	s_P imposer cod_P	38	**s suivre cod_P**	**42**
s_D est diagnostiqué	41	**s_F est indiqué**	**27**	s suivre cod_F	30
s_D diagnostiquer coi_S	11	**s_P est indiqué**	**73**	**s_On traiter cod_D**	**30**
s_P est imposé	94	s_P indiquer coi_D	16	**s_D est traité**	**23**

- Cases with f<10: still some relevant frames: *s_J diagnostiquer s_S* $\Rightarrow$ *dépister* (to detect)
  e.g. *[...] il faut entreprendre la médication contre le TDAH à la recommandation de la personne qui diagnostique et suit le patient[...]*
  (A treatment against the TDAH should be started with the recommendation of the person who diagnoses and monitors the patient.)

Evaluation of the Results

Verbs	Nb frames Corpus				LVF	Com	Precision full
	full	part	error	total			
abaisser	12	5	2	19	1	0	0.63
admettre	20	12	2	34	1	1	0.58
diagnostiquer	23	6	2	31	1	1	0.74
imposer	38	22	6	66	1	0	0.57
indiquer	110	82	2	194	1	1	0.56
relever	57	20	2	79	1	1	0.72
subir	33	9	3	45	1	1	0.73
suivre	74	68	0	142	1	0	0.52
survivre	8	6	0	14	2	1	0.57
stimuler	35	3	2	40	1	1	**0.87**
traiter	55	52	2	109	1	1	**0.50**
Total	465	285	23	773	11	8	0.601

- Variation of the precision score from one verb to the other: from 0.50 (*traiter*) to 0.87 (*stimuler*)
- Precision highly dependent on the quality of the semantic annotation of the subcorpora (see *traiter*)

Conclusion and perspectives

- Conclusion:
  - A method for the acquisition of medical frames of verbs, for the creation of a text simplification resource for medical texts
    - Three medical subcorpora differentiated according to the level of expertise of their author and intended audicience
    - A syntactico-semantic classification of verbs
    - An existing medical terminology
  - Promising evaluation results: precision from 0.50 to 0.87
- Future work:
  - Further development of the method
    - Improvement of the semantic annotation system
    - Finding a way for considering the frames of verbs not covered by the LVF while setting up the text simplification resource

Thanks for your attention !

Cependant, tous s'accordent à dire que la grande majorité des patients (environ 85 %) souffrant d'un SAOS ne seraient pas diagnostiqués durant leur vie [...]
However, all agree that the majority of patients (around 85%), suffering from OSA would not be diagnosed during the course of their life [...]
s_J diagnostiquer cod_D: Il est déterminant que les professionnels de la santé sachent diagnostiquer un AVC avec quelques outils simples.
It is crucial that health professionals know how to diagnose a stroke with a few simple tools.
s_On diagnostiquer cod_D coi_S: En 2011, on a diagnostiqué en Belgique 641 cas de ce type de cancer.
In 2011, 641 cases of this type of cancer were diagnosed in Belgium.
s_J diagnostiquer cod_D coi_S: Cette page s'adresse aux patients chez qui un médecin a diagnostiqué des extrasystoles ventriculaires.
This page is meant for patients on which a doctor has diagnosed ventricular extrasystoles.
s_D diagnostiquer coi_S: De nouveaux cas d'hypertension diagnostiquée chez les adultes.
New cases of hypertension diagnosed on adults.
s_D diagnostiquer: 9 millions d'angines sont annuellement diagnostiquées en France.(9 million cases of tonsillitis are diagnosed annually in France.)