Searching for parallel sentences in comparable corpora for French biomedical text simplification

Rémi Cardon, Natalia Grabar

CNRS, Université de Lille, UMR 8163 - STL - Savoirs, Textes, Langage
{remi.cardon, natalia.grabar}@univ-lille.fr

29/05/21

Text simplification

Transform texts :
- to make them more accessible
- preserve meaning

- Pre-processing step for NLP
  - Syntactic parsing (Chandrasekar et *al.*, 1996)
  - Automatic summarization (Vale et *al.*, 2020)
  - Machine translation (Štajner et *al.*, 2019)
  - Information retrieval / extraction (Evans & Orasan, 2019)
  - ...
- Make information accessible to humans
  - Children or adults with reading difficulties (De Belder & Moens, 2010)
  - Second language learners (Tack et *al.*, 2016)
  - Adults with neurocognitive disorders (Carroll et *al.*, 1999)
  - General public (specialized texts)
  - ...

Healthcare issues

- Health literacy of patients (Sørensen, 1996 ; Berkman et *al.*, 2011 ; Margat et *al.*, 2017) :
  - understanding of medical information and success of the treatment
  - communication between doctors and patients

  → confidence relationship
- Accessibility of medical information is important

Availability of medical information

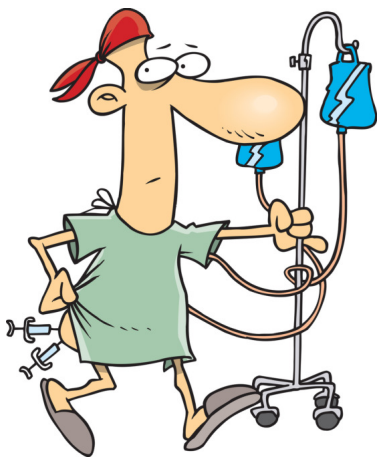
- Growing volume of written medical information online :
  - Bibliographic databases
  - Scholarly societies, associations
  - Compagnies
  - Encyclopedias
- Poor understanding by the general public (AMA, 1999)

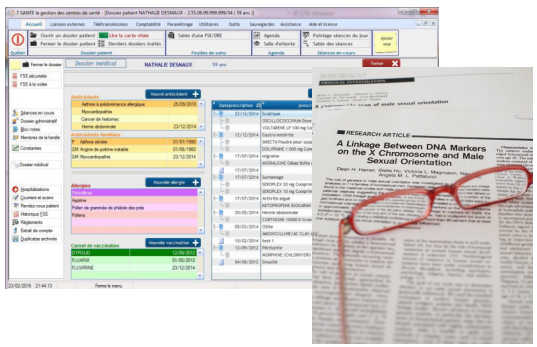
Le cholestéatome est une forme d'otite chronique avec présence d'épithélium pavimenteux stratifié dans l'oreille moyenne. Cet épithélium desquame et se kératinise (structure histologique de l'épiderme), et peut provoquer l'érosion voire la destruction des structures contenues dans et autour de l'oreille moyenne.
La forme la plus fréquente est le cholestéatome acquis par évolution terminale d'une otite chronique (poches de rétractions essentiellement).

Understanding of health documents

- Two aspects for document understanding :
  1. Patient literacy (his knowledge and needs)
     → Therapeutic Patient Education
  2. Document readability
     → Linguistics
     → Natural Language Processing (NLP)

Improve
the knowledge
and literacy
of people
↓
Therapeutic
Patient Education





Improve
the readability
of texts
↓
NLP, linguistics
methods

Simplification principles and initiatives

- Guides for human redactors :
  - FALC *facile à lire et à comprendre* (Audiau, 2009)
  - OCDE (OCDE, 2015)
  - Haute Autorité de Santé [1]
- Simplification initiatives :
  - UNAPEI (https://www.unapei.org/)
  - Cochrane Foundation (https://france.cochrane.org/)
- Simplification recommendations :
  - Short sentences
  - Active voice
  - Accessible vocabulary
  - Typesetting
  - ...

1. https://www.has-sante.fr/jcms/c_430286/fr/elaboration-d-un-document-ecrit-d-information-a-l-intention-des-patien

Levels of automatic simplification

- Mainly two levels of simplification :
  - Lexical simplification :
    replace complex terms by more accessible equivalents
    - lexical substitution
    - anaphora replacement
  - Syntactic simplification :
    rework sentence structures
    - split, merge, reorganize sentences
    - insertion / deletion (clauses, phrases...)
    - modify verbal voice

Approaches for automatic simplification

- Rule-based approaches (Chandrasekar et *al.*, 1996 ; Carroll et *al.*, 1998 ; Devlin & Tait, 1998 ; Zhu & *al.*, 2010 ; Woodsend & Lapata, 2011 ; Brouwers et *al.*, 2014 ; Evans & Orăsan, 2019)
- Deep learning models (*seq2seq*) (Nisioi et *al.*, 2017, Zhang & Lapata, 2017 ; Sulem et *al.*, 2018 ; Shardlow & Nawaz, 2019 ; Abdul Rauf *al.*, 2020 ; Cooper & Shardlow, 2020)
- Parallel corpora with different degrees of complexity are needed

Simplification corpora

- Comparable corpora (general language) :
  same topics, different content
  - Revisions in Simple English Wikipedia (Yatskar et *al.*, 2010)
  - Scientific articles (Kim & Hullman, 2016)
- Parallel corpora (general language) :
  aligned sentences
  - English, Spanish, Italian, Brazilian Portuguese, Danish
    (Chandrasekar et Srinivas, 1997 ; Saggion et *al.*, 2012 ; Brunato *al.*, 2014 ; Caseli et *al.*, 2009 ; Klerke et Søgaard, 2012)
  - Not always freely available
  - French : Alector (Gala et *al.*, 2020)
- French : no available corpus at the time of our work
- Medical domain : no available corpus

Objective

- Build a corpus for automatic simplification
  - in French
  - for the medical domain

Complex : *Les médicaments inhibant le péristaltisme sont contre-indiqués dans cette situation*
Simple : *Dans ce cas, ne prenez pas de médicaments destinés à bloquer ou ralentir le transit intestinal*

Building the comparable corpus

- French biomedical corpus
- Document pairs
  - Same topic
  - Different levels of complexity (more or less technical)
- Three subcorpora :
  - Encyclopedia articles
  - Drug information
  - Scientific literature

Encyclopedia articles

- Wikipedia : targets the general population

  (2 186 891 tokens, 19 287 lemmas)
- Vikidia : targets children from 8 to 13

  (183 051 tokens, 3 117 lemmas)
- Written independently
- Medecine portal : 2 $\times$ 575 articles
- Examples :

  Complex : La luette ou uvule est un appendice conique situé au fond de la cavité buccale. La luette est un organe de 10 à 15 millimètres de long. Elle est constituée d'un tissu membraneux et musculaire.

  Simple : La luette ou uvule est un appendice conique situé au fond de la bouche. C'est un organe fait de tissus membraneux et musculaires, d'environ 10 à 15mm de long, qui pend à la partie moyenne du voile du palais.

Drug information

- Published by the French Ministry of Health
- *Résumés des caractéristiques produit (RCP)*, for medical practicioners

  (51 705 111 tokens, 43 515 lemmas)
- Leaflets, for patients (drug boxes)

  (33 116 119 tokens, 25 725 lemmas)
- $2 \times 11\,800$ RCP / leaflets pairs
- Examples :
  Complex : hypersensibilité à l'huile de paraffine. / - ne pas utiliser chez les personnes présentant des difficultés de déglutition en raison du risque d'inhalation bronchique et de pneumopathie lipoïde.
  Simple : si vous avez une allergie à l'huile de paraffine. / - ne pas utiliser chez les personnes présentant des difficultés pour avaler en raison du risque d'inhalation de la paraffine liquide qui entraîne une pneumopathie lipoïde.

Scientific summaries

- Published by the Cochrane Foundation
- Summaries of systematic reviews, for practicioners

  (2 804 335 tokens, 11 558 lemmas)
- Simplified versions of the summaries, for the general public

  (1 491 243 tokens, 7 567 lemmas)
- $2 \times 3\,815$ summaries
- Examples :

  Complex : L'hématome aigu de l'oreille est une affection qui se caractérise par la formation d'une collection sanguine sous le périchondre du pavillon. Il est souvent provoqué par un traumatisme contondant.

  Simple : L'hématome aigu de l'oreille est une affection qui se caractérise par la formation d'une collection sanguine dans le pavillon (oreille externe), souvent à la suite d'un traumatisme contondant.

Overview of the comparable corpus

	docs	$tokens_{tech}$	$tokens_{acc}$	*total tokens*	$lemmas_{tech}$	$lemmas_{acc}$
Encyclopedia	575 × 2	2,293,078	197,672	2,490,750	19 287	3,117
Drug Info.	11,800 × 2	52,313,126	33,682,889	85,996,015	43,515	25,725
Cochrane	3,815 × 2	2,840,003	1,515,051	4,355,054	11,558	7,567
Total	16,190 × 2	57,446,207	35,395,612	92,841,819		

- 16 190 document pairs
- Total : 92M tokens
  - Complex : 57M tokens
  - Simple : 35M tokens
- Available for research [2]

2. https ://natalia.grabar.free.fr/resources.php

Automatic sentence alignment

1. Reference data creation
2. Training and testing models

Reference data creation

- Training data for automatic alignment
- Processing unit : sentence pair
- Manual alignment of sentences
- 39 randomly selected document pairs
  - 2×14 encyclopedia articles
  - 2×12 drug information documents
  - 2×13 Cochrane summaries
- Two annotators
  1. Independent annotation
  2. Consensus

Reference data creation

Alignment criteria

- At least conjugated verb per sentence
- Equivalence – identical or almost identical meaning :
  Complex : Une gêne visuelle passagère peut être ressentie après instillation du collyre.
  Simple : Une gêne visuelle passagère peut être ressentie après l'administration du collyre.
- Inclusion – the meaning of one sentence is included in the other one (enables tracking cases of merging and splitting) :
  Complex : La maladie de Charcot est l'autre nom de la sclérose latérale amyotrophique.
  Simple : Il est le découvreur de la sclérose latérale amyotrophique, ou maladie de Charcot, une maladie neurodégénérative.

Reference data creation

Non-alignment criteria

- Identical sentences, or that differ only by ponctuation or grammatical words :
  Complex : Effets sur l'aptitude à conduire des véhicules ou à utiliser des machines
  Simple : Effets sur l'aptitude à conduire des véhicules et à utiliser des machines
- Intersection – Two sentences that share meaning but each one adds its own information :
  Complex : Une faiblesse musculaire (hypotonie axiale), des difficultés d'alimentation (troubles de la succion entrainant une faible prise de poids), une hyperexcitabilité, une agitation ou des tremblements peuvent survenir chez le nouveau-né, ces troubles étant réversibles.
  Simple : Un traitement en fin de grossesse par benzodiazépines même à faibles doses, peut être responsable chez le nouveau-né de signes d'imprégnation tels qu'hypotonie axiale, troubles de la succion entraînant une faible prise de poids.

Reference data creation

Overview

		Complex				*Simple*				*Alignment rate*	
		raw		*aligned*		*raw*		*aligned*			
Corpus	*doc.*	*sent.*	*tok.*	*sent.*	*tok.*	*sent.*	*tok.*	*sent.*	*tok.*	*tech.*	*simp.*
Encyclopedia	14×2	2,416	36,703	39	873	235	2,659	39	710	1,61	16,6
Drug info.	12×2	4,391	44,684	143	4,227	2,710	27,804	143	8,481	3,25	5,27
Cochrane	13×2	426	8,852	84	2,278	227	4,688	84	2,466	19,71	36,56
Total	39×2	7,233	90,239	266	7,378	3,172	35,131	266	11,657		

- Interannotator agreement before consensus (Cohen's κ) : 0,76
- 266 aligned sentence pairs
- Semantic relations in the sentence pairs :
  - 136 cases of equivalence
  - 130 cases of inclusion
- Substantial differences between the corpora

Reference data creation

Linguistic data description

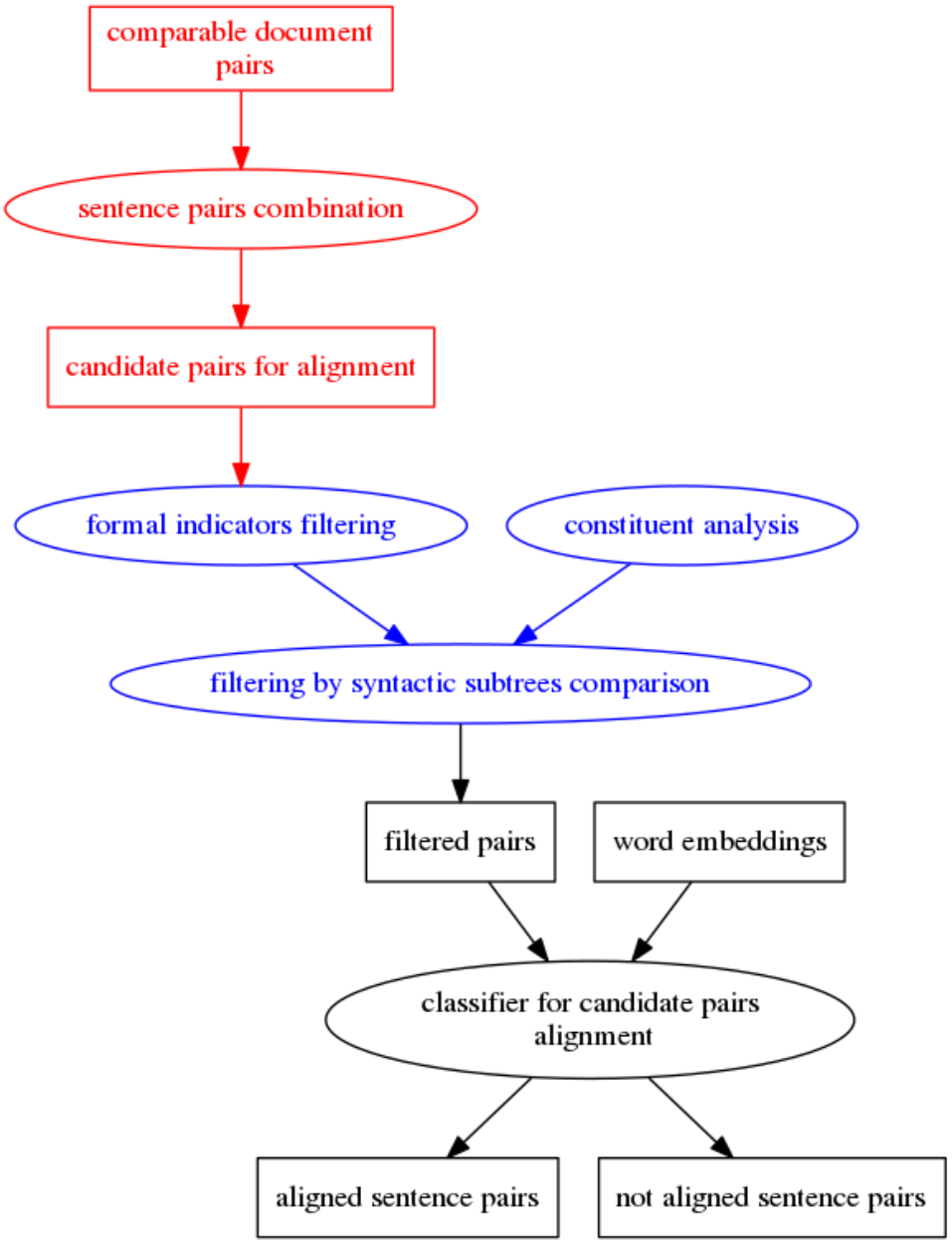
- Typology of lexical and syntactic transformations for simplification (Koptient et *al.*, 2019)
- Most frequent transformation : synonymy (41%)
- Comparison with existing typology for general language in Italian (Brunato et *al.*, 2014)
  $\rightarrow$ Medical language simplification has specific processed

Automatic sentence alignment

Approach

- Binary classification : sentences aligned or not aligned
- Sentence pairs features :
  - BL (*baseline*)
    - Words in common (Barzilay & Elhadad, 2003)
    - Percentage of words of a sentence found in the other
    - Length ratio between the two sentences (in words)
    - Average word length difference between the two sentences
  - N (n-grams)
    - Total number of common bigrams and trigrams (in characters)
  - S (similarity)
    - cosine similarity, Dice (Dice, 1945), Jaccard (Jaccard, 1912)
  - L (Levenshtein)
    - Levenshtein edit distance (Levenshtein, 1966) between the two sentences (characters, and words)
  - PL (*plongements lexicaux*, word embeddings)
    - *WAVG* (Štajner et *al.*, 2018)
    - *CWASA* (Franco-Salvador et *al.*, 2016)

Automatic sentence alignment



Automatic sentence alignment

Choice of the algorithm

- Positive class : equivalence
- Ratio 1 : 1 (136 positive, 136 negative)

Classifier	*P*	*R*	*F1*	*MSE*	*VP*
`Perceptron`	0,90	0,93	0,92	0,08	28
`MLP`	0,93	0,93	0,93	0,06	28
`RF`	**1,00**	**0,97**	**0,98**	**0,02**	**29**
`LDA`	0,93	0,87	0,90	0,09	26
`QDA`	0,96	0,90	0,93	0,06	27
`LogReg`	0,97	**0,97**	0,97	0,03	**29**
`SGD`	0,90	0,93	0,92	0,08	28
`LinSVM`	0,97	0,93	0,95	0,04	28

Automatic sentence alignment

Feature types behaviour

Descripteurs	*R*	*P*	*F1*	*MSE*	*VP*	*Descripteurs*	*R*	*P*	*F1*	*MSE*	*VP*
BL	0,97	0,93	0,95	0,05	28	*BL + L + S*	1,00	0,97	0,98	0,02	29
S	0,97	0,97	0,97	0,03	29	*BL + L + N*	1,00	0,97	0,98	0,02	29
L	0,90	0,93	0,92	0,09	28	*BL + L + PL*	1,00	0,97	0,98	0,02	29
N	0,97	0,93	0,95	0,05	28	*BL + S + N*	1,00	0,97	0,98	0,02	29
PL	0,97	0,97	0,97	0,03	29	*BL + S + PL*	1,00	0,97	0,98	0,02	29
L + S	1,00	0,93	0,97	0,03	28	*BL + N + PL*	1,00	0,97	0,98	0,02	29
L + N	1,00	0,97	0,98	0,02	29	*L + S + N*	1,00	0,97	0,98	0,02	29
L + PL	0,97	0,97	0,97	0,03	29	*L + S + PL*	1,00	0,97	0,98	0,02	29
S + N	1,00	0,97	0,98	0,02	29	*L + N + PL*	1,00	0,97	0,98	0,02	29
S + PL	1,00	0,97	0,98	0,02	29	*BL + L + S + N*	1,00	0,97	0,98	0,02	29
BL + L	1,00	0,97	0,98	0,02	29	*BL + L + S + PL*	1,00	0,97	0,98	0,02	29
BL + S	1,00	0,97	0,98	0,02	29	*BL + L + N + PL*	1,00	0,97	0,98	0,02	29
BL + N	1,00	0,97	0,98	0,02	29	*BL + S + N + PL*	1,00	0,97	0,98	0,02	29
BL + PL	1,00	0,97	0,98	0,02	29	*L + S + N + PL*	1,00	0,97	0,98	0,02	29
N + PL	1,00	0,97	0,98	0,02	29	*BL + L + S + N + PL*	1,00	0,97	0,98	0,02	29

- Positive class : equivalence
- Balanced data : ratio 1 : 1
- All feature types are useful
- Better results for any possible combination

Automatic sentence alignment

Equivalence – inclusion

- Ratio 1 : 1
- All features

Set	*P*	*R*	*F1*	*MSE*
Equivalence	1,00	0,97	0,98	0,02
Inclusion	1,00	0,94	0,97	0,03

- Inclusion slightly harder to classify

Automatic sentence alignment

Tackling data imbalance

- Search space : all possible sentence pairs
  $\rightarrow$ Equivalence ratio : 8000 : 1
- Objective : imbalance reduction
  - *FI* : formal indicators
    - Reject sentences with less than five words
    - Reject pairs where the sentences are identical
  - Independent application of two syntactic filters
    - Comparison of the syntactic trees of the two sentences
      $\rightarrow$ constituency parser : `benepar` (Kitaev et *al.*, 2018)
    - Depth 1 : word in common in node with identical label
    - Depth 3 : word in common and if node with identical label among the three superior levels

Automatic sentence alignment

Tackling imbalance – Results

Remaining pairs	*Original*	*FI*	*Depth 1*	*Depth 3*
All pairs	1 164 407	409 530	16 879	21 428
Equivalence	136	136	94	94
Inclusion	130	130	94	100

- Important reduction of negative examples
- Limited reduction of positive examples
- Depth 3 kept
- Imbalance ratio reduced from 8000 : 1 to 200 : 1

Automatic sentence alignement

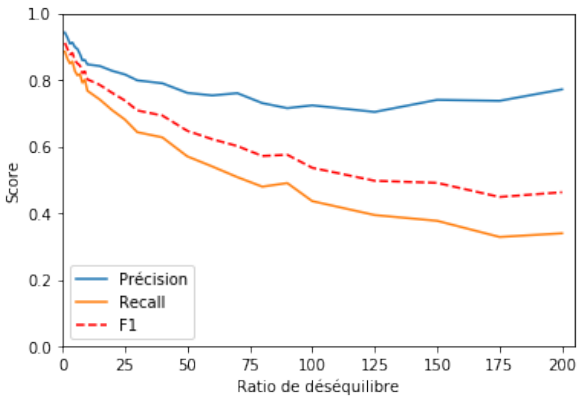
Experimental protocol

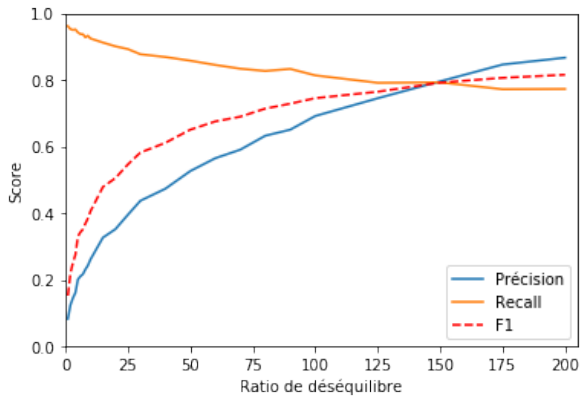
- Equivalence and inclusion
- Training with different ratios of imbalance
  - 1 : 1 to 200 : 1
- All positive examples
- Random sampling for negative examples
- All features
- 2/3 training, 1/3 test
- Each model is applied to the whole data
- 20 iterations for each configuration

Alignement automatique de phrases

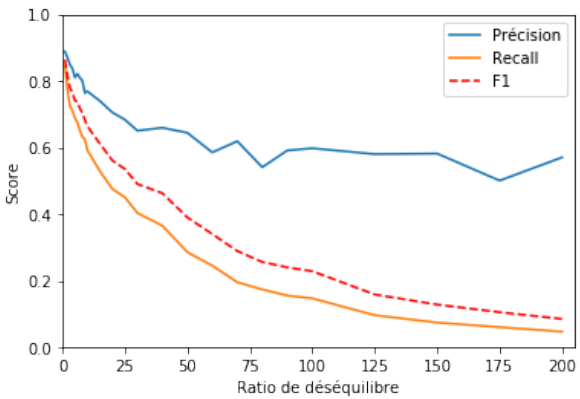
Results

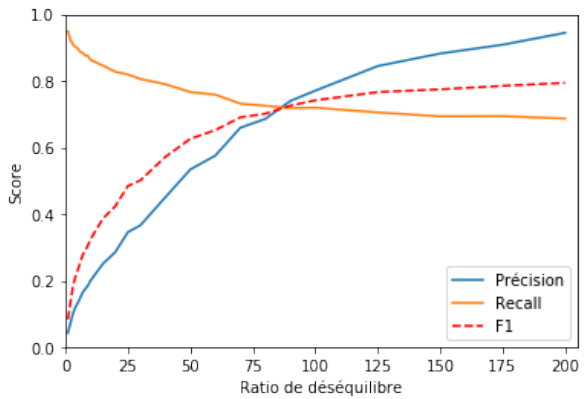
Equivalence





Inclusion





- Important effect of the imbalance ratio used in training

Automatic sentence alignment

Application to unseen data

	Equivalence	Inclusion	Intersection	False positives
Nb of alignments	75	15	2	8

- Analysis of 100 alignments proposed by the model
- Data not seen during the training phase

Automatic sentence alignment

Conclusion and perspectives

- High performance with balanced data
- Filters to fight imbalance
- Main perspective : work on the syntactic filter
  - Targets precise phenomena
  - Better knowledge of the obtained resource
  - Better explicability of the subsequent tasks' results

Conclusion

Resulting resources

- Main resource : parallel corpus (Cardon & Grabar, 2021)
  - 10,942 equivalent sentence pairs
- On the side : creation of a semantic similarity corpus (Cardon & Grabar, 2020)
  - 1,010 sentence pairs annotated from 0 to 5 by five persons
  - Sources : CLEAR and Wikipédia/Vikidia
  - Available for research
- Both resources used during the DEFT 2020 evaluation campaign (Cardon et *al.*, 2020)