UCLouvain

Corpora and resources for automatic text simplification:

The role of the target audience

Rémi Cardon
CENTAL – UCLouvain

CENTAL
centre de traitement automatique du langage

Outline

- Automatic Text simplification
  - Objectives
  - Methods
  - Evaluation
  - Resources
  - Example
  - Perspectives
  - Conclusion

Outline

- Automatic Text simplification (ATS)
  - Objectives
  - Methods
  - Evaluation
  - Resources
  - Example
  - Perspectives
  - Conclusion

Objectives

- First ATS works : pre-processing step for other natural language processing tasks
  - Part-of-Speech tagging (Chandrasekar et al., 1996)
  - Summarization (Vale et al., 2020)
  - Machine translation (Štajner et al., 2019)
  - Information extraction / retrieval (Evans & Orasan, 2019)

Objectives

- Make information more accessible to humans
  - Language learners (Tack et al., 2016)
  - Adults with neurocognitive disorders (Carroll et al., 1999)
  - Children or adults with reading difficulties (De Belder & Moens, 2010)
  - General public – specialized texts (Cardon & Grabar, 2020)

Simplification

- Guide for human writers:
  - FALC *facile à lire et à comprendre – easy to read and understand* (Audiau, 2009)
  - OCDE (OCDE, 2015)
  - Haute Autorité de Santé [1]
- Simplification initiatives:
  - UNAPEI (https://www.unapei.org/)
  - Cochrane Foundation (https://france.cochrane.org/)
- Writing recommendations:
  - Short sentences
  - No passive voice
  - Accessible words
  - Layout
  - ...

[1] https://www.has-sante.fr/jcms/c_430286/fr/elaboration-d-un-document-ecrit-d-information-a-

Outline

- Automatic Text simplification
  - Objectives
  - Methods
  - Evaluation
  - Resources
  - Example
  - Perspectives
  - Conclusion

Methods

- Historically, rule-based systems for syntactic simplification
- Manually design rules for syntactic simplification
  - Delete specific clause types
  - Split complex sentences into several simple sentences
  - Pronoun resolution
  - ...
- Requires efficient syntactic parsers
- Produces a lot of rules
- Costly to make and hard to maintain

Methods

- Lexical simplification:
  1. Complex word identification / Lexical complexity prediction: what to simplify
  2. Candidates generation: list of potential substitutes
  3. Candidate selection: keep relevant substitutes (part-of-speech, meaning in context...)
  4. Candidate ranking: identify the simplest adequate substitute

Methods

- Currently, neural methods
- End-to-end: lexical and syntactic simplification in one pass
- Let the machine learn from the data
- Require parallel corpora
    - Corpora with aligned complex / simple sentences
    - Importance of the quality of the data

Outline

- Automatic Text simplification
  - Objectives
  - Methods
  - Evaluation
  - Resources
  - Example
  - Perspectives
  - Conclusion

Evaluation

- Three criteria are considered:
  - Grammaticality: is the output grammatically correct?
  - Meaning preservation: does the output express the same meaning as the input?
  - Simplicity: is the output simpler than the input?

Evaluation

- BLEU: metric from machine translation (n-gram overlap between output and reference(s))
- SARI: metric designed for text simplification (n-gram overlap between input, output and reference(s))
- Readability measures
  - Commonly, Flesch Reading Ease and FKGL
- In the past few years, all have been shown to have poor correlation with simplicity

Evaluation

- Manual evaluation
- 5-point Likert scales for the three criteria
  - Grammaticality
  - Meaning preservation
  - Simplicity
- No consistency / consensus from one study to another
- Difficult to interpret
- Performed by researchers or crowdworkers
- Less frequently: eye-tracking, reading comprehension evaluation

Outline

- Automatic Text simplification
  - Objectives
  - Methods
  - Evaluation
  - Resources
  - Example
  - Perspectives
  - Conclusion

Resources

- Neural methods are data-hungry
- Building parallel corpora:
  - Manual alignment
  - Automatic alignment
  - Machine translation of existing corpora in other languages
  - Crowdsourcing

Resources

Manual approach			
Language	**Textual genre**	**Target**	**Dimension**
ENG (Pellow and Eskenazi, 2014)	Everyday documents	GP	200 sentence pairs
ENG (Xu et al., 2015)	Newspapers	CHI	56,037 original sentences
ENG (Barzilay and Elhadad, 2003)	Encyclopedia Britannica	CHI	2,600 easy-to-read documents
ENG (Allen, 2009)	Classroom materials	LL	178,967 of simplified words
ENG (Petersen and Ostendorf, 2007)	Newspapers	LL	2,539 original sentences
ENG (Xu et al., 2016)	Wikipedia	CS	2,359 original sentences
ENG (Alva-Manchego et al., 2020a)	Wikipedia	CS	2,359 original sentences
Many (Orasan et al., 2013)	Miscellanea	PLI	320 original sentences
SPA (Bott and Saggion, 2014)	Newspapers	PLI	145 simplified sentences
SPA (Collados, 2013)	Newspapers	NLP	300 simplified sentences
FRE (Brouwers et al., 2014)	Narrative texts	L2LL	83 original sentences
FRE (Grabar and Cardon, 2018)	Encyclopedic, scientific, clinical texts	GP	4,596 sentence pairs
FRE (Gala et al., 2020)	L1 student materials	PLI	52,704 tokens
DAN (Klerke and Søgaard, 2012)	Newspapers	L2LL	3,701 document pairs
POR (Caseli et al., 2009)	Newspapers	PLL	2,116 original sentences
POR (Aluísio et al., 2008)	Popular science articles	PLL	882 original sentences
GER (Klaper et al., 2013)	Websites	PLI	7,755 original sentences
GER (Sauberli et al., 2020)	Newspapers	L2LL	3,616 sentence pairs
JPN Goto et al. (2015)	Newspapers	L2LL	2,885 sentence pairs
EUS Gonzalez-Dios et al. (2017)	Popular science articles	L2LL	227 original sentences
RUS Dmitrieva and Tiedemann (2021)	Literary texts	L2LL	69,737 original sentences
ITA Tonelli et al. (2016)	Administrative texts	GP	157 original sentences
ITA Brunato et al. (2015)	Children's literature	PLI	1,060 sentence pairs
ITA Brunato et al. (2015)	Educational material	L2LL	1,356 original pairs

Resources

(Semi)Automatic Approach			
ENG Kauchak (2013)	Wikipedia	GP	167K sentence pairs
ENG Kajiwara and Komachi (2016)	Wikipedia	GP	492,993 sentence pairs
ENG Zhu et al. (2010)	Wikipedia	GP	108,016 sentence pairs
ENG Narayan et al. (2017)	Wikipedia	GP	5,546 original sentences
ENG Woodsend and Lapata (2011)	Wikipedia	GP	14,831 sentence pairs
ENG Botha et al. (2018)	Wikipedia	GP	1,004,944 original sentences
ENG Pavlick and Callison-Burch (2016)	Miscellanea	CS	4.5 million of simplifying paraphrase rules
ITA Tonelli et al. (2016)	Wikipedia	GP	530 original sentences
FRE Brouwers et al. (2014)	Wikipedia	L2LL	72 original sentences
FRE Cardon and Grabar (2020)	Wikipedia	GP	297,494 sentence pairs
ITA Brunato et al. (2016)	Web corpus	GP	63,000 sentence pairs

Brunato, D., Dell'Orletta, F., & Venturi, G. (2022). Linguistically-Based Comparison of Different Approaches to Building Corpora for Text Simplification: A Case Study on Italian. Frontiers in Psychology, 13

Resources

- Training corpora : scrapped from the Internet (mostly Wikipedia)
- Evaluation corpora:
  - Subsets of training corpora
  - Crowdsourced
    - End-to-end: ask annotators to manually simplify a relatively small number of sentences
    - Lexical simplification: ask annotators to propose substitutes

Resources

- Existing but not yet taken into account: typologies of transformations
- Two types of transformations:
  - Linguistic transformations
  - String edits

Resources

	Lexical Simpification	Syntactic Simplification	Discourse-level Simplification
Dyslexia	Long words Less frequent words Homophones Words that are orthographically similar New words Non-words		
Aphasia	Information density Noun compounds	Long sentences Long sequences of adjectives Passive voice Object relative clauses Comparison of word meaning	
Intellectual Disability (ID)	Limited vocabulary		
Deaf and Hard-of-Hearing	Limited vocabulary	Complex sentences Morphology Syntax	
Austism Spectrum Disorder	Words related to emotions		Figurative language Texts that require little social knowledge
Second Language Learners	Limited vocabulary		Tight text structure
Children	New words Limited vocabulary		

Rennes, E., Santini, M., & Jönsson, A. (2022, June). The Swedish Simplification Toolkit: Designed with Target Audiences in

Outline

- Automatic Text simplification
  - Objectives
  - Methods
  - Evaluation
  - Resources
  - Example
  - Perspectives
  - Conclusion

Experiment

- French biomedical text simplification
- Medical corpus (CLEAR) – 4,500 sentence pairs:
  - Wikipedia/Vikidia
  - Drug leaflets (for medical practitioners and patients)
  - Medical literature reviews, and manually simplified versions
- Additional data (WikiLarge) – 300,000 sentence pairs
- Paraphrases – 4,516 medical terms paraphrased
- Questions:
  - Do we really need large amounts of data ?
  - Do we need specialized data for specialized text simplification ?
  - Can a resource of a different linguistic nature (paraphrases) help ?

Experiment

- Algorithm
- OpenNMT-py
- Encoder-decoder with attention [2]:
  - Two bi-LSTM layers of 500 cells for encoder and decoder
  - ADAM optimizer
  - learning rate 0,001
  - dropout probability 0.3
  - attention dropout probability 0,2
- Use of the paraphrase table:
  - SL and LE: `–replace_unk`
  - LS : `–phrase_table` with the paraphrase table

[2] Hyperparameters inspired by Abdul Rauf et al., (2020)

No paraphrase table (SL)

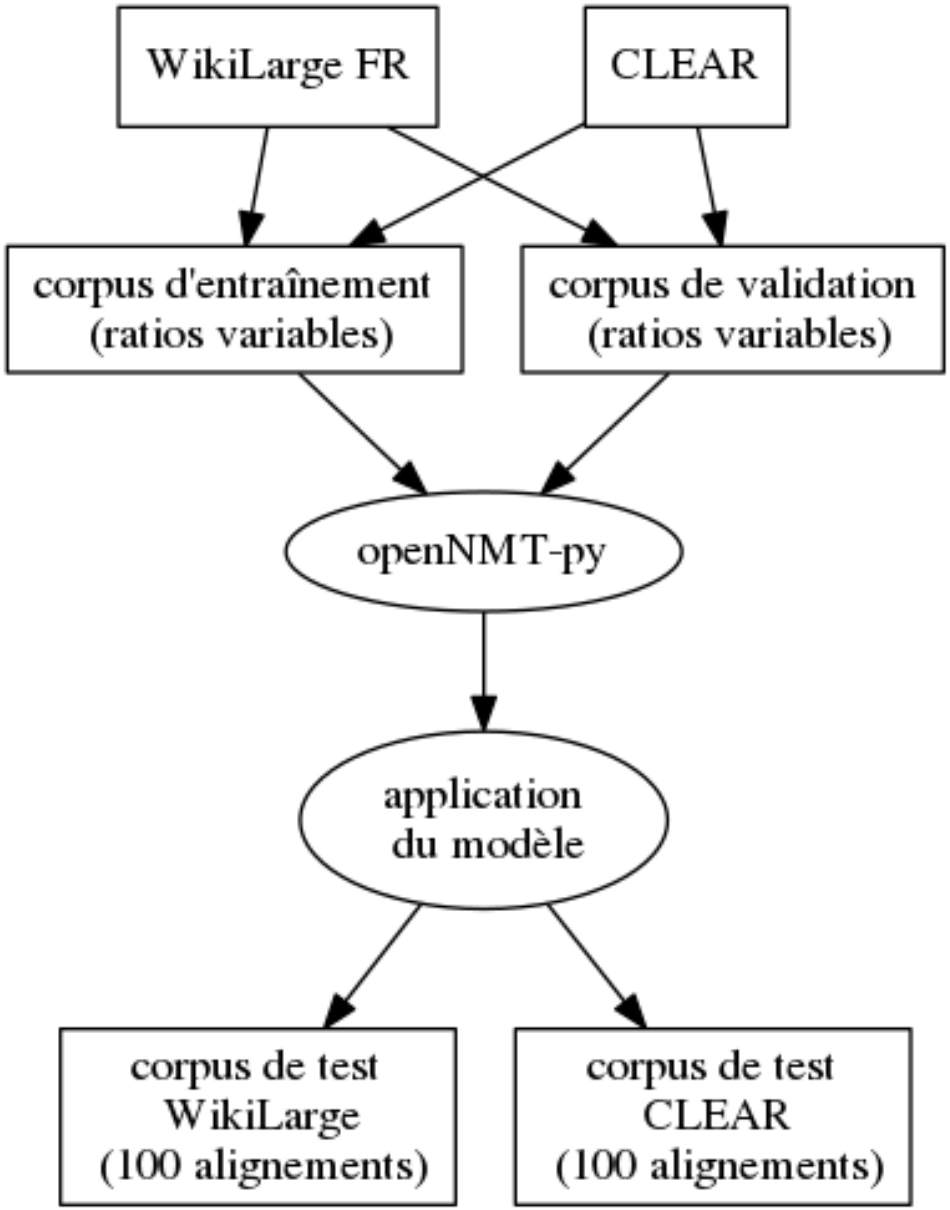


Table used during Simplification (LS)

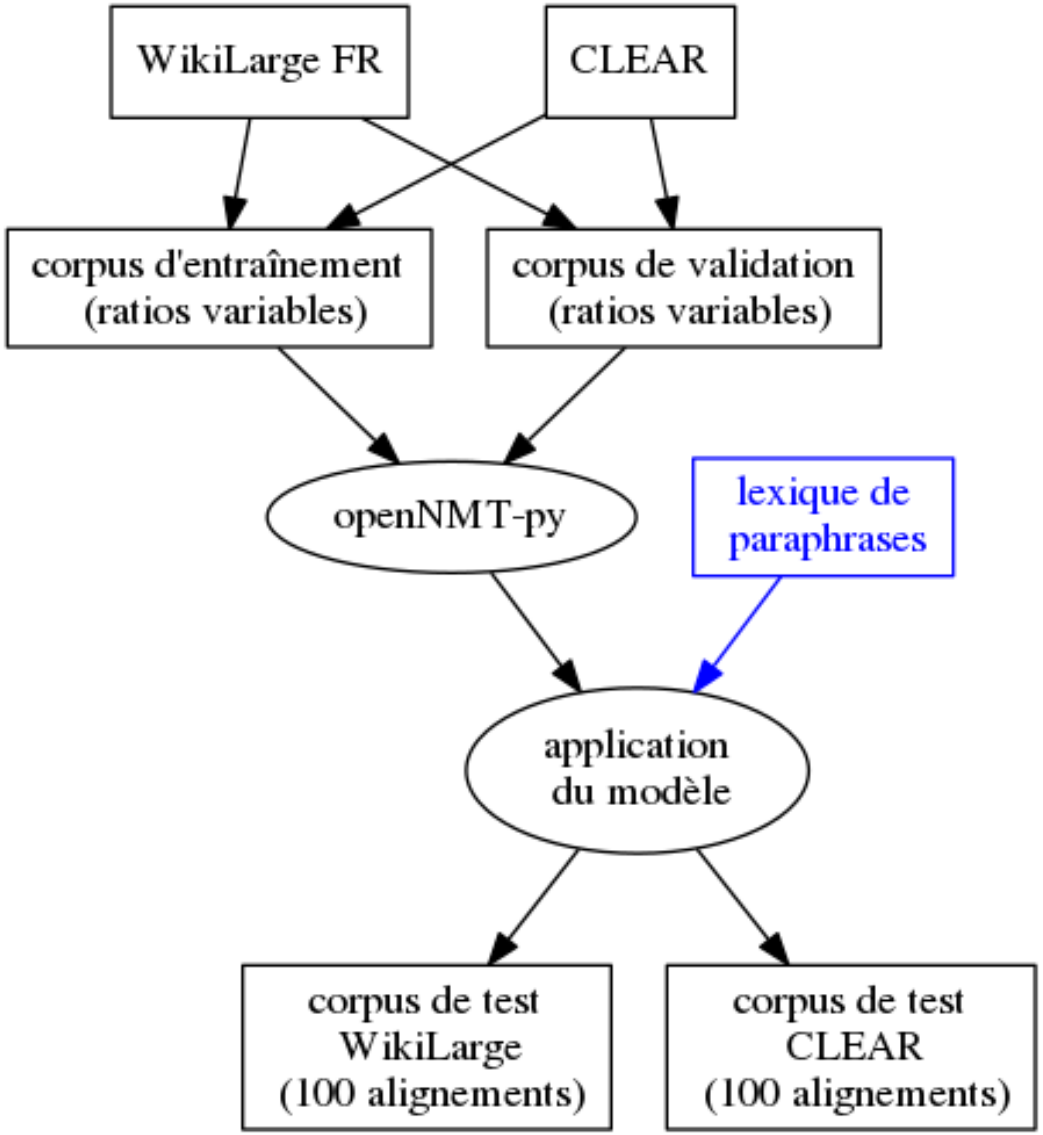
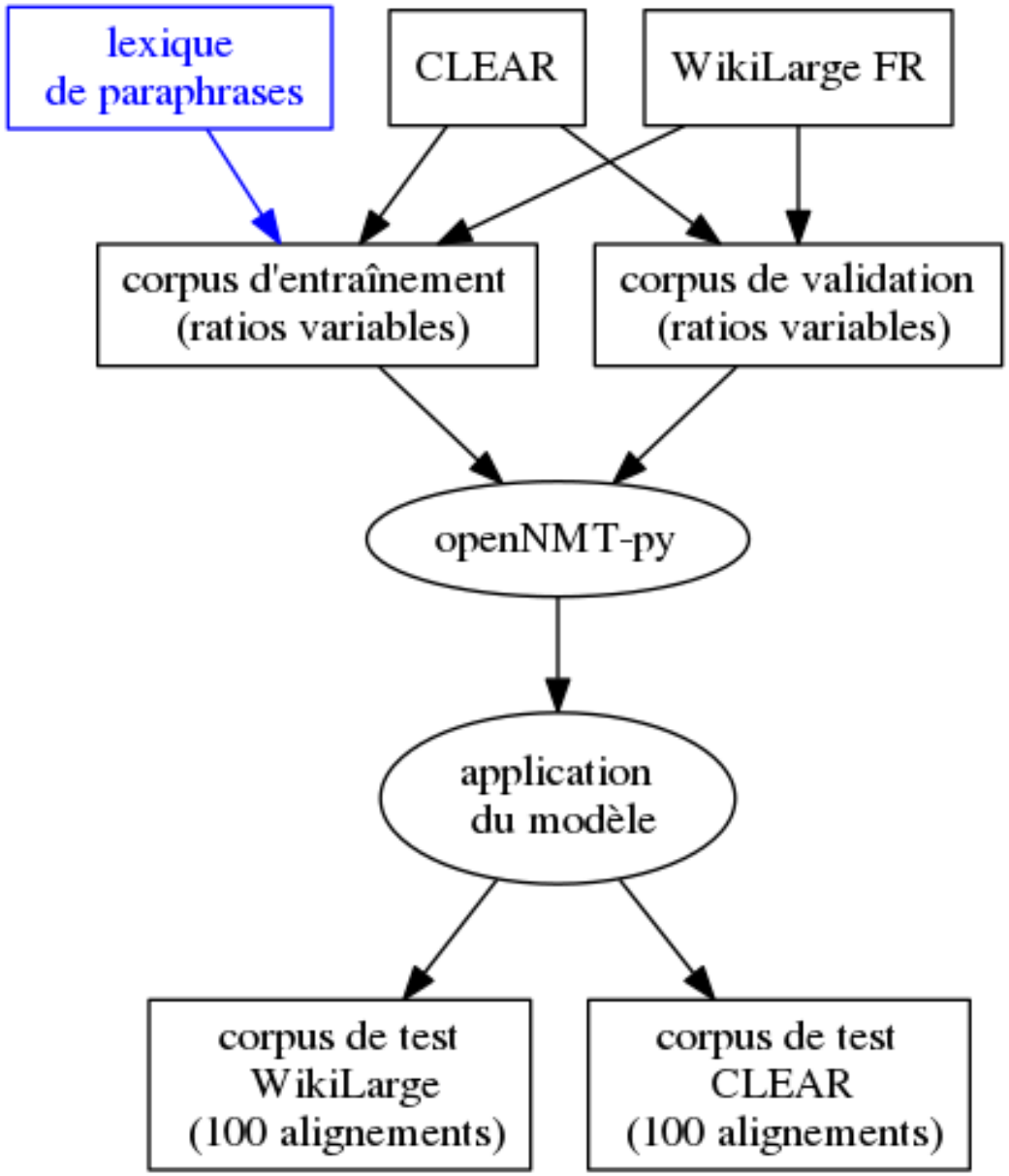


Table used during training (LE)



Results

Model	Test on WikiLarge FR
Source	Le 14 octobre 1960, le candidat à la présidence John F. Kennedy a proposé le concept de ce qui est devenu le Peace Corps sur les marches de l'Union du Michigan.
Référence	John F. Kennedy, un candidat à la présidence, a proposé l'idée de ce qui devint le Peace Corps sur les marches de l'Union du Michigan le 14 octobre 1960.
Wikilarge	*No change*
CLEAR	le cancer est de la médecine
SL 1:25	En 1960, le candidat du président John F. Kennedy a suggéré le Peace Corps sur les marches de l'Union du Michigan.
SL 1:50	En 1960, le candidat à la présidence John F. Kennedy a proposé l'idée du Peace Corps sur les marches de l'Union du Michigan.
SL 1:75	*Aucun changement*
LE 1:25	Le 14 octobre 1960, le candidat au président John F. Kennedy a proposé le concept de ce qui est devenu le Peace Corps sur les marches de l'Union du Michigan.
LE 1:50	En 1960, le candidat à la présidence John F. Kennedy a proposé l'idée de ce qui est devenu le Peace Corps sur les marches de l'Union du Michigan.
LE 1:75	John F. Kennedy a proposé le concept de ce qui est devenu le Peace Corps sur les marches de l'Union du Michigan.

Results – Exemple CLEAR

Model	Test on CLEAR
Source	une hypotension artérielle peut être observée en cas d'administration intraveineuse trop rapide, inférieure à 60 minutes (voir rubrique 4.2)
Référence	une hypotension artérielle peut être observée en cas d'administration intraveineuse trop rapide, inférieure à 60 min
WikiLarge	Une artérielle artérielle peut être observée en cas de crise intraveineuse trop rapide et inférieure à 60 minutes
CLEAR	le traitement de la naissance de la naissance de l' repos [...] de l' repos de la médecine [...] de la médecine de la peau [...] de la peau de la
SL 1:50 & SL 1 : 75	une hypotension artérielle peut être observée en cas d'administration intraveineuse trop rapide, inférieure à 60 minutes
LS 1:75	une tension inférieure à la normale artérielle peut être observée en cas d' administration intraveineuse trop rapide, inférieure à 60 minutes
LE 1:50	une hypotension artérielle peut être observée en cas d'administration intraveineuse trop rapide, inférieure à 60 minutes (voir rubrique 3)
LE 1:75	une diminution de la tension artérielle peut être observée en cas d' administration intraveineuse trop rapide, inférieure à 60 minutes

Outline

- Automatic Text simplification
  - Objectives
  - Methods
  - Evaluation
  - Resources
  - Example
  - Perspectives
  - Conclusion

Perspectives – Objectives

- Go beyond the sentence level
  - Some works have begun exploring paragraph simplification
- Personalized text simplification

Perspectives – Methods

- Design less data-hungry methods (unsupervised)
- Add constraints to the text generation step
  - Sentence length
  - Vocabulary used
  - Syntactic constraints
  - ...

Perspectives

- Design better evaluation metrics
  - Adequacy with the target audience
  - Adequacy with the text genre
- Involve members of the target audience in the process

Perspectives

- Build resources with members of the target audiences
- Build resources with experts (writers) of accessible language for a given audience

Outline

- Automatic Text simplification
  - Objectives
  - Methods
  - Evaluation
  - Resources
  - Example
  - Perspectives
  - Conclusion

Conclusion

- Current methods begin to be able to produce good paraphrases according to some surface criteria...
  - ... but which criteria for which audience ?
  - ... but evaluating simplicity is still an open question
- Need for more inclusion of members of target audiences, starting at the design step
- Need for input from other fields (psycholinguistics, sociolinguistics...)