Analyse outillée d'un corpus de spécifications techniques en vue de la création d'une langue contrôlée pour la rédaction des exigences

Maxime Warnier
maxime.warnier@univ-tlse2.fr

CLLE-ERSS (CNRS & Université Toulouse - Jean Jaurès)
Centre National d'Études Spatiales
France

Thématiques Actuelles de la Recherche en TAL
Toulouse, le 7 décembre 2015

Plan

Plan

Thèse débutée en octobre 2013, CLLE-ERSS :

Analyse linguistique comparée des guides de rédaction technique et des usages réels dans les spécifications de systèmes spatiaux au CNES en vue d'améliorer la rédaction et la compréhension des exigences

Directrice : Anne Condamines
Responsables CNES : Daniel Galarreta, Jean-François Gory

Financement : CNES & Conseil régional de Midi-Pyrénées

Plan

Demande de la sous-direction Assurance Qualité du CNES : comment améliorer la rédaction des **exigences** ?

"conditions or capabilities that must be met or possessed by a system or system component to satisfy a contract, standard, specification, or other formally imposed document" [IEEE 1990]

"expression d'une condition ou d'une fonctionnalité à laquelle doit répondre un système ou un logiciel" [specief.org]

Exemple : « ORAMIC doit pour chaque session de mesure restituer finement l'attitude du satellite Microscope »

Demande de la sous-direction Assurance Qualité du CNES : comment améliorer la rédaction des **exigences** ?

"conditions or capabilities that must be met or possessed by a system or system component to satisfy a contract, standard, specification, or other formally imposed document" [IEEE 1990]

"expression d'une condition ou d'une fonctionnalité à laquelle doit répondre un système ou un logiciel" [specief.org]

Exemple : « ORAMIC doit pour chaque session de mesure restituer finement l'attitude du satellite Microscope »

Exigences (**spécifications**), rédigées en langue naturelle (**français**, anglais) :

- logiciels de gestion (DOORS, Reqtify, etc.)
- mais aucune règle imposée, aucune aide à la rédaction

Risques liés à la langue naturelle (oral et écrit) :

- ambiguïté (lexicale, syntaxique, référentielle)
- flou, imprécision
- incomplétude, implicite

Risques liés à la langue naturelle (oral et écrit) :

- ambiguïté (lexicale, syntaxique, référentielle)
- flou, imprécision
- incomplétude, implicite

Exigences critiques au CNES :

- projets de grande **envergure**
  (nombre d'exigences et d'intervenants, durée)
- **valeur contractuelle** entre parties prenantes

Risques liés à la langue naturelle (oral et écrit) :

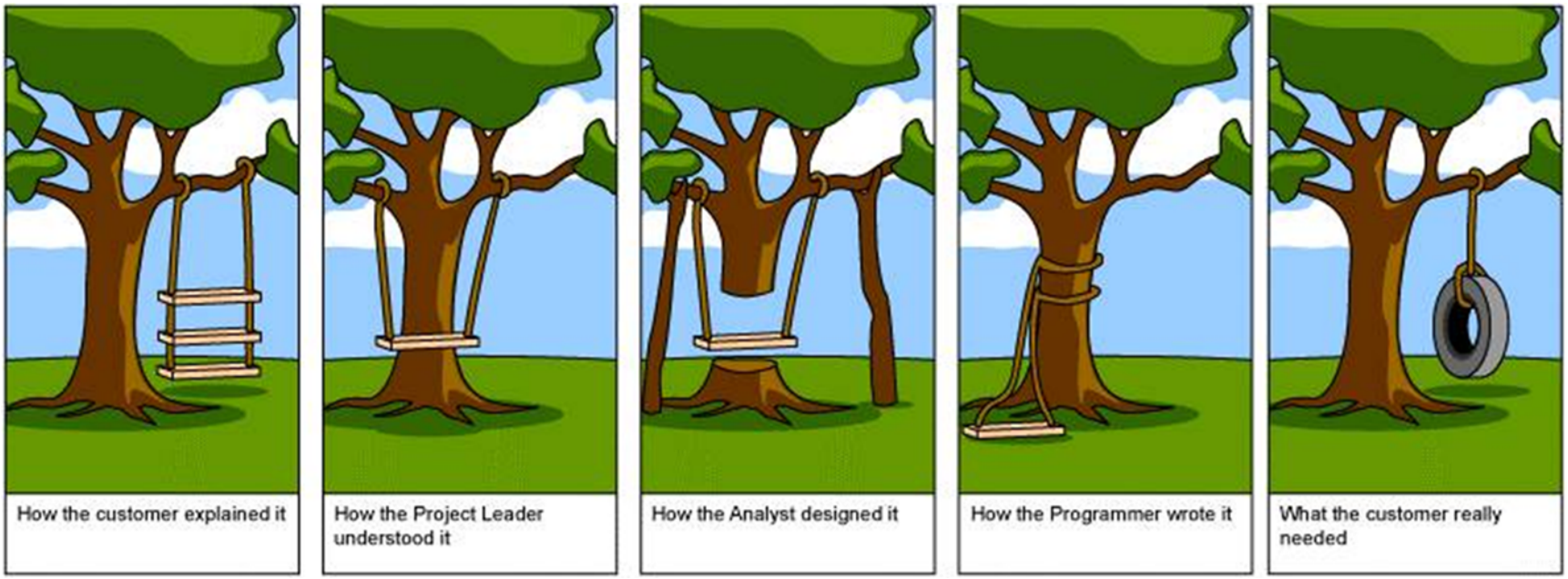
- ambiguïté (lexicale, syntaxique, référentielle)
- flou, imprécision
- incomplétude, implicite

Exigences critiques au CNES :

- projets de grande **envergure**
  (nombre d'exigences et d'intervenants, durée)
- **valeur contractuelle** entre parties prenantes

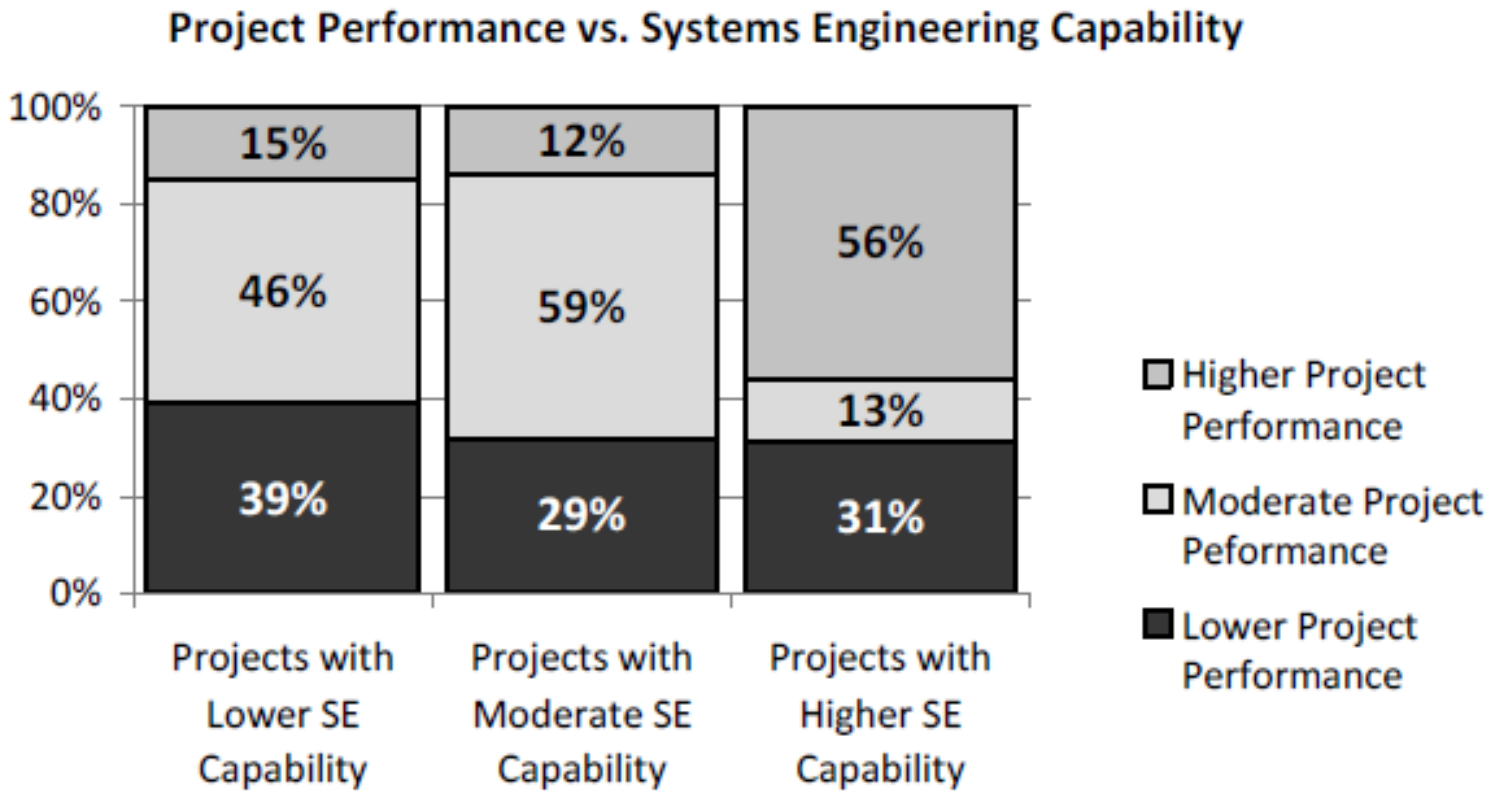
Conséquences possibles :

- Retards
- Augmentation des coûts
- Litiges
- Accidents



"Tree Swing Cartoon"

L'ingénierie des exigences est un facteur de réussite...



http://www.sei.cmu.edu/reports/08sr034.pdf

... la qualité rédactionnelle des exigences aussi !
(travail « en amont »)

Avantages de la langue naturelle :

- expressivité maximale
- simplicité
- inévitable

Avantages de la langue naturelle :

- expressivité maximale
- simplicité
- inévitable

Objectifs généraux (et évaluation) :

- une solution
  - qui réduise au maximum les risques langagiers [= efficace]
  - qui reste aussi simple et naturelle que possible [= utilisée]
- un apport théorique

Plan

Solution : une **langue contrôlée** (*Controlled Natural Language*) [Kuhn 2014]

Historique :

- Basic English (1930)
- Caterpillar Fundamental English (1971)
- ...

Solution : une **langue contrôlée** (*Controlled Natural Language*) [Kuhn 2014]

Historique :

- Basic English (1930)
- Caterpillar Fundamental English (1971)
- ...

Définition :

1. Basée sur une langue naturelle
2. **Pose des restrictions sur le vocabulaire, la syntaxe et/ou la sémantique**
3. Reste suffisamment proche de la langue source
4. Est un langage consciemment défini

Critique des langues contrôlées existantes :

- parfois inapplicables (trop contraignantes)
- parfois inefficaces (pas assez strictes)
- manque de rigueur (incohérentes)

$\Rightarrow$ pas adaptées aux pratiques réelles

Critique des langues contrôlées existantes :

- parfois inapplicables (trop contraignantes)
- parfois inefficaces (pas assez strictes)
- manque de rigueur (incohérentes)

$\Rightarrow$ pas adaptées aux pratiques réelles

Établies par des experts du domaine (ingénieurs), pas des experts de la langue (linguistes)

- connaissances complémentaires

Critique des langues contrôlées existantes :

- parfois inapplicables (trop contraignantes)
- parfois inefficaces (pas assez strictes)
- manque de rigueur (incohérentes)

$\Rightarrow$ pas adaptées aux pratiques réelles

Établies par des experts du domaine (ingénieurs), pas des experts de la langue (linguistes)

- connaissances complémentaires

Objectif concret : proposer une langue contrôlée pour la rédaction des exigences au CNES basée sur des spécifications authentiques (c'est-à-dire proche des pratiques réelles)

Plan

Problème : peut-on mettre au jour des **régularités linguistiques** si les rédacteurs écrivent en langue non contrôlée ?

Problème : peut-on mettre au jour des **régularités linguistiques** si les rédacteurs écrivent en langue non contrôlée ?

Nous supposons l'existence d'un **genre textuel*** ou d'un **sous-langage**** propre à la rédaction des exigences au CNES, dont nous voudrions déterminer la grammaire.

* *"a recognizable communicative event characterized by a set of communicative purpose(s) identified and mutually understood by the members of the professional or academic community in which it regularly occurs"* [Bhatia 1993]

** *"an identifiable genre or text-type in a given subject field, with a relatively or even absolutely closed set of **syntactic structures** and **vocabulary**"* [Somers 1998]

La distinction langue contrôlée *vs* sous-langage peut être rapprochée de l'opposition **normaison** *vs* **normalisation** [Guespin 1993] :
« [...] deux procès normatifs : la normaison, relevant de l'activité spontanée à l'œuvre dans tout échange, et la normalisation, domaine des interventions conscientes et planifiées » [Gaudin 1993]

La distinction langue contrôlée *vs* sous-langage peut être rapprochée de l'opposition **normaison** *vs* **normalisation** [Guespin 1993] :
« [...] deux procès normatifs : la normaison, relevant de l'activité spontanée à l'œuvre dans tout échange, et la normalisation, domaine des interventions conscientes et planifiées » [Gaudin 1993]

Nous voudrions donc proposer

- une langue contrôlée basée sur un genre textuel
- une normalisation basée sur une normaison

Plan

Extrait d'une spécification (document Word)

3.1. Paramètre TM

N£ PHR-IF-1/6-30-CNES_100 £N
T£ A un paramètre TM décommuté pourra être associé :
* un calculé sol de condition de significativité,
* une surveillances sol. £T
La condition de significativité permet de connaître l'état du paramètre TM (significatif ou non).

N£ PHR-IF-1/6-30-CNES_200 £N
T£ Une seule surveillance à un moment donné est applicable au paramètre TM. £T

Pré-traitements :

1. Suppression des tableaux et images
   - algorithmes
2. Exportation au format texte
3. Concaténation
4. Extraction des exigences
   - problèmes de balisage
   - mélange anglais/français

Corpus d'exigences en français extraites des spécifications :

- Pleiades : 2 500 exigences (120 000 mots)
- Microscope : 1 000 exigences (44 000 mots)

$\Rightarrow$ mêmes profils de rédacteurs
$\Rightarrow$ mêmes niveaux de spécification (système, segment, interface)

$\Rightarrow$ différence d'envergure
$\Rightarrow$ différence d'objectif et de domaine

Corpus d'exigences en français extraites des spécifications :

- Pleiades : 2 500 exigences (120 000 mots)
- Microscope : 1 000 exigences (44 000 mots)

⇒ mêmes profils de rédacteurs
⇒ mêmes niveaux de spécification (système, segment, interface)

⇒ différence d'envergure
⇒ différence d'objectif et de domaine

Corpus de comparaison :

- un manuel du CNES
  (Techniques et Technologies des Véhicules Spatiaux)
- des articles du *Monde*
- (à venir :) des exigences d'un autre domaine

Plan

Linguistique de corpus

Linguistique de corpus

Quelques contraintes et difficultés :

- Confidentialité
- Taille (et nature) du corpus
- Langue
- Complexité des analyses

Combiner analyses quantitatives et qualitatives

Combiner analyses quantitatives et qualitatives

Combiner approches **corpus-based*** et **corpus-driven****
[Tognini-Bonelli 2001]
* *"assumes the validity of linguistic forms and structures derived from linguistic theory"* [Biber 2009]
** *"is more inductive, so that the linguistic constructs themselves emerge from analysis of a corpus"* [Biber 2009]

Phénomènes étudiés :

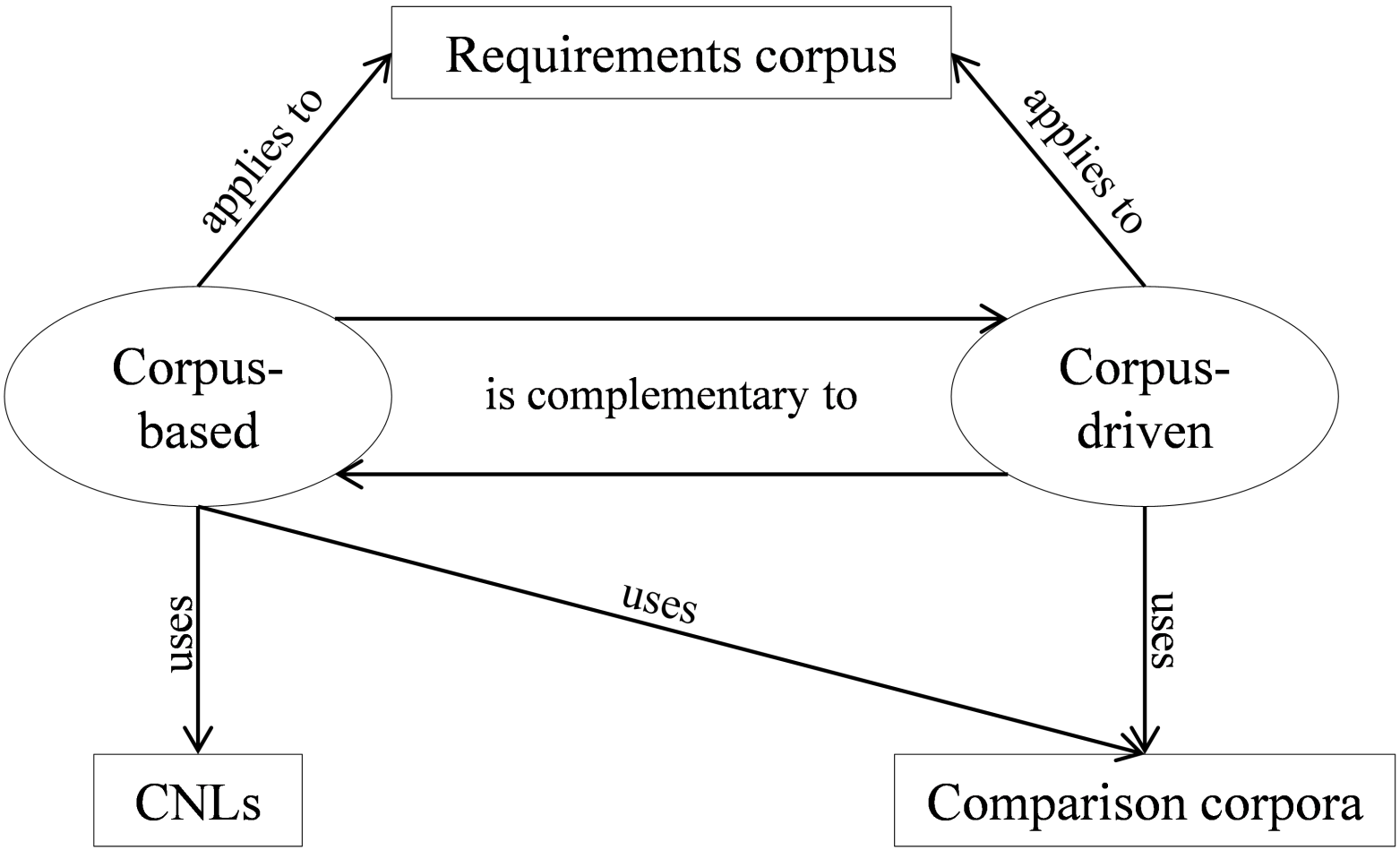
- Phénomènes visés par les règles des langues contrôlées (ou extrapolés)
- Phénomènes émergeant du corpus

Phénomènes étudiés :

- Phénomènes visés par les règles des langues contrôlées (ou extrapolés)
- Phénomènes émergeant du corpus

Langues contrôlées envisagées :

- ASD *Simplified Technical English* (ASD-STE), 2007 : standard de l'aérospatiale (documentation de maintenance)
- *Guide for Writing Requirements* d'INCOSE, 2011 : multi-disciplinaire, largement inspiré de règles antérieures


Requirements corpus
applies to
applies to
Corpus-based
is complementary to
Corpus-driven
uses
uses
uses
CNLs
Comparison corpora

Plan

Conjonctions : “Avoid combinators” [INCOSE]

Combinators are words that join clauses together, such as 'and', 'or', 'then', 'unless'. Their presence in a requirement usually indicates that multiple requirements should be written

Conjonctions : "Avoid combinators" [INCOSE]
Combinators are words that join clauses together, such as 'and', 'or', 'then', 'unless'. Their presence in a requirement usually indicates that multiple requirements should be written

Pronoms : "Repeat nouns in full instead of using pronouns to refer to nouns in other requirement statements" [INCOSE]
Pronouns are words such as 'it', 'this', 'that', 'he', 'she', 'they', 'them'. When writing stories, they (sic.) are a useful device for avoiding the repetition of words ; but when writing requirements, pronouns should be avoided, and the proper nouns repeated where necessary

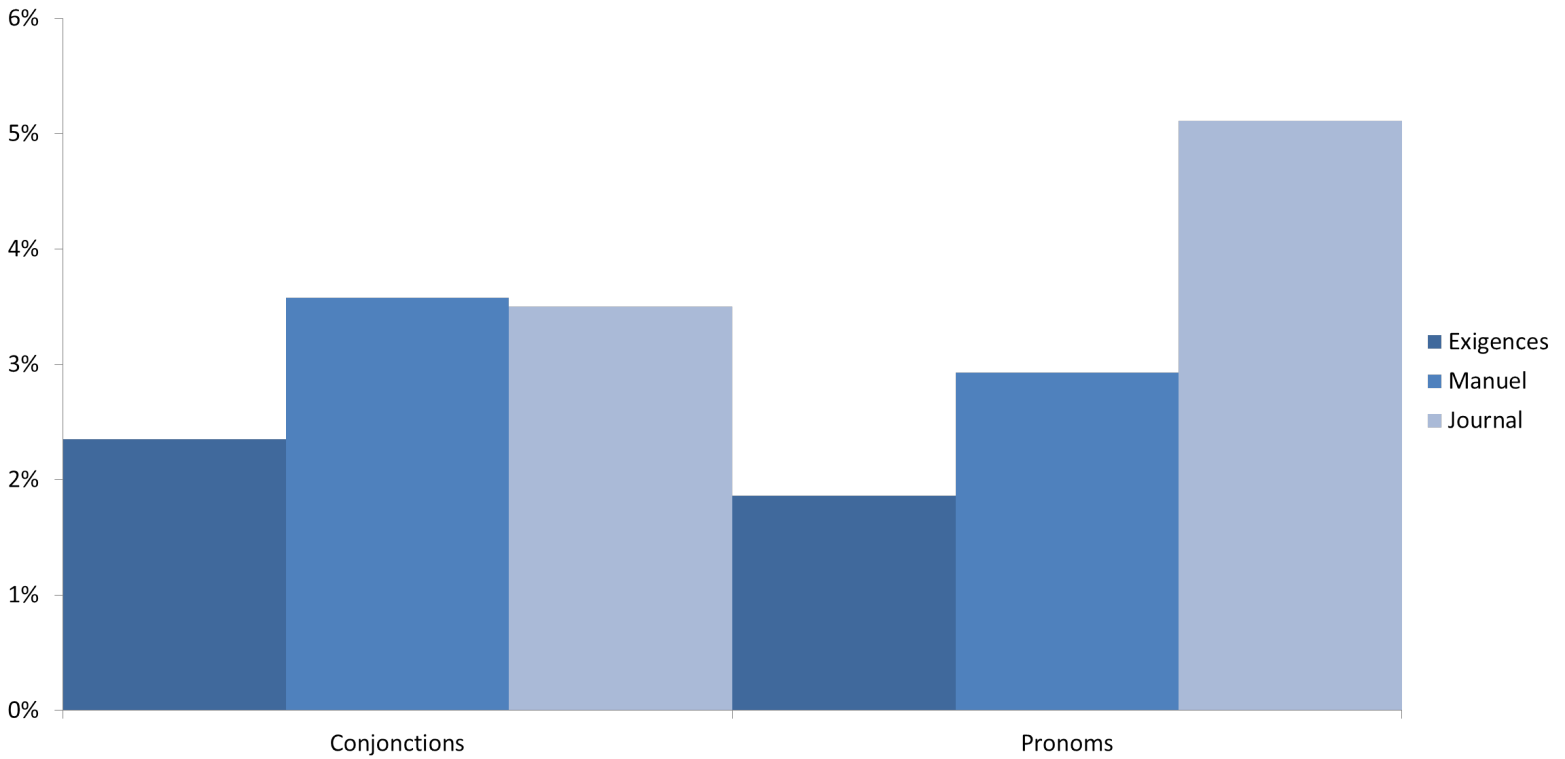
Conjonctions : "Avoid combinators" [INCOSE]
Combinators are words that join clauses together, such as 'and', 'or', 'then', 'unless'. Their presence in a requirement usually indicates that multiple requirements should be written

Pronoms : "Repeat nouns in full instead of using pronouns to refer to nouns in other requirement statements" [INCOSE]
Pronouns are words such as 'it', 'this', 'that', 'he', 'she', 'they', 'them'. When writing stories, they (sic.) are a useful device for avoiding the repetition of words ; but when writing requirements, pronouns should be avoided, and the proper nouns repeated where necessary

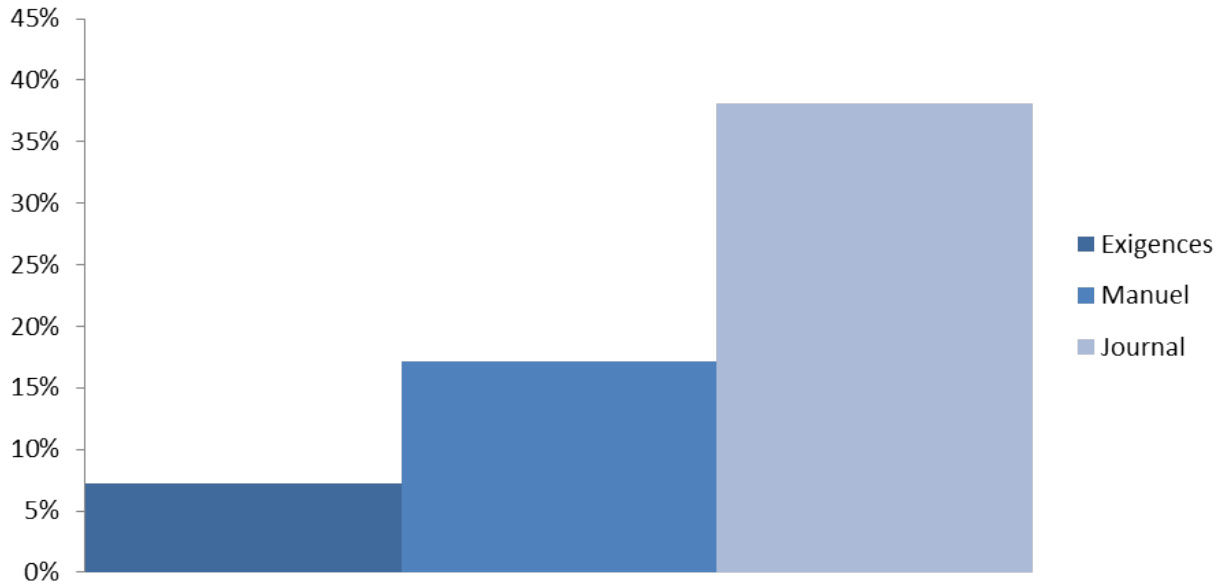
Longueur des phrases : "Keep procedural sentences as short as possible (20 words maximum)" [ASD-STE]

Plan

	Coordination	Subordination	**Conjonctions**	**Pronoms**
Exigences	1,66 %	0,69 %	2,35 %	1,86 %
Manuel	2,75 %	0,83 %	3,58 %	2,93 %
Journal	2,40 %	1,09 %	3,50 %	5,11 %



	Phrases longues	Longueur moyenne
Exigences	7.20 %	11
Manuel	17.10 %	15
Journal	38.10 %	24



« Si la différence (en valeur absolue) entre les dates de fin de lecture de deux fichiers, lus sur tranche de COME M - canal TMI i et sur tranche de COME N - canal TMI j, est inférieure à OPS_DELAI_INTER_FIN_LEC secondes, alors il est interdit d'enchaîner (lecture enchaînée) par la lecture de la tranche de COME N sur le canal i et de la tranche de COME M sur le canal j. »

Plan

Conjonctions

Obligatoires

Le générateur de TCH vérifiera **que** la valeur du champ PHASE est comprise entre 0 **et** FREQ_DIV -1.

Conjonctions

Obligatoires

Le générateur de TCH vérifiera **que** la valeur du champ PHASE est comprise entre 0 **et** FREQ_DIV -1.

Utile

Les champs SM_ID **et** FM_ID seront extraits à partir de la BDS.

Conjonctions

Obligatoires

Le générateur de TCH vérifiera **que** la valeur du champ PHASE est comprise entre 0 **et** FREQ_DIV -1.

Utile

Les champs SM_ID **et** FM_ID seront extraits à partir de la BDS.

Problématique

Les demandes sont saisies sur le FOS **et** le logiciel ARPE gère les conflits entre les demandes Spot, Hélios et Pléïades.

Pronoms

Obligatoire

Il ne sera pas utile de vérifier ce paquet "vide".

Pronoms

Obligatoire

Il ne sera pas utile de vérifier ce paquet "vide".

Utile

Le paquet ne sera généré que s'**il** est activé par le LVC.

Pronoms

Obligatoire

Il ne sera pas utile de vérifier ce paquet "vide".

Utile

Le paquet ne sera généré que s'**il** est activé par le LVC.

Problématique

Il calculera aussi, a une fréquence paramétrable (ordre de grandeur 1 mois), la moyenne de mise en œuvre.

Plan

Conclusions :

- Analyse d'exigences authentiques de projets spatiaux
- Des caractéristiques linguistiques (moins de pronoms et de conjonctions, phrases plus courtes, etc.) semblent indiquer l'existence d'un genre textuel
- Elles peuvent être utilisées pour améliorer des règles de rédaction ou en proposer de nouvelles

Conclusions :

- Analyse d'exigences authentiques de projets spatiaux
- Des caractéristiques linguistiques (moins de pronoms et de conjonctions, phrases plus courtes, etc.) semblent indiquer l'existence d'un genre textuel
- Elles peuvent être utilisées pour améliorer des règles de rédaction ou en proposer de nouvelles

Perspectives :

- Vérifier nos analyses sur des spécifications hors domaine spatial
- Vérifier la pertinence des règles par des tests de compréhension
- Vérifier que les règles soient réellement suivies

References (1/2)

- AeroSpace and Defence Industries Association of Europe (2007). "Simplified Technical English. Specification ASD-STE100. International specification for the preparation of maintenance documentation in a controlled language. Issue 4."
- Bhatia, V. K. (1993). Analysing genre : Language use in professional settings. London : Longman
- Biber, D. (2009). "Corpus-Based and Corpus-driven Analyses of Language Variation and Use," in The Oxford Handbook of Linguistic Analysis, 1st ed., B. Heine and H. Narrog, Eds. Oxford University Press
- Condamines, A., & Warnier, M. (2014). Linguistic Analysis of Requirements of a Space Project and Their Conformity with the Recommendations Proposed by a Controlled Natural Language. In B. Davis, K. Kaljurand, & T. Kuhn (Eds.), Controlled Natural Language (pp. 33–43). Springer International Publishing
- Gaudin, F. (1993). Pour une socioterminologie : Des problèmes sémantiques aux pratiques institutionnelles. Rouen : Publications de l'Université de Rouen
- Guespin, L. (1993). Normaliser ou standardiser ? Le langage et l'homme, 28(4), Bruxelles : De Boek Université, 213-222

References (2/2)

- IEEE Standard Glossary of Software Engineering Terminology. (1990). IEEE Std 610.12-1990, 1–84
- International Council on Systems Engineering (2011). "Guide for Writing Requirements." INCOSE
- Kuhn, T. (2014). A Survey and Classification of Controlled Natural Languages. Computational Linguistics, 40(1), 121–170
- Somers, H. (1998). An Attempt to Use Weighted Cusums to Identify Sublanguages. In D.M.W. Powers (Ed.), NeMLaP3/CoNLL 98 : New Methods in Language Processing and Computational Natural Language Learning (pp. 131–139). ACL
- specief.org. L'Ingénierie des Exigences. Site internet de la SPECIEF, [en ligne]. Disponible sur : http ://cms.specief.org/conc/index.php/ingenierie-des-exigences/
- Tognini-Bonelli, E. (2001). Corpus Linguistics at Work. John Benjamins Publishing
- Urieli, A. (2013). Robust French syntax analysis : reconciling statistical methods and linguistic knowledge in the Talismane toolkit. Université de Toulouse 2 - Le Mirail, Toulouse

Identification de structures :

Comment extraire uniquement des candidats termes et des structures (motifs) propres au **genre** des exigences - et non au domaine ?

- Cf. le lexique scientifique transdisciplinaire [Tutin 2007]

Les termes candidats peuvent appartenir au genre et/ou au domaine

- Les extracteurs de terminologie sont généralement prévus pour des corpus où le domaine est plus important
- Problèmes similaires avec d'autres types d'outils automatiques

Extraction terminologique : Talismane [Urieli 2013]

- Analyse syntaxique
- Noms
- Fréquence minimale = 5

Pleiades : 1 551 candidats termes
Microscope : 716 candidats termes

Pour retirer les candidats termes relevant du domaine, nous avons utilisé une "**stop list**" : une liste de termes utilisée par le CNES pour indexer des documents dans la base de connaissances (validés par des experts du domaine)

Pleiades : 1 355 (-196)
Microscope : 598 (-118)
Exemple (éliminé) : "satellite"

Pour retirer les candidats termes relevant du domaine, nous avons conservé seulement les entrées **présentes dans les deux listes**

Communs : 300 (-1 055 Pleiades ; -298 Microscope)
Exemples (éliminés) : “magnétomètre”, “masse interne”

Après révision manuelle, **267** candidats termes restants (-33)
(vs une extraction sur les deux corpus : presque 2 000 !)

Concerne à la fois les exigences fonctionnelles (“fonctionnalité”) et non fonctionnelles (“disponibilité”)

Pattern Mining : SDMC [Quiniou et al. 2012]

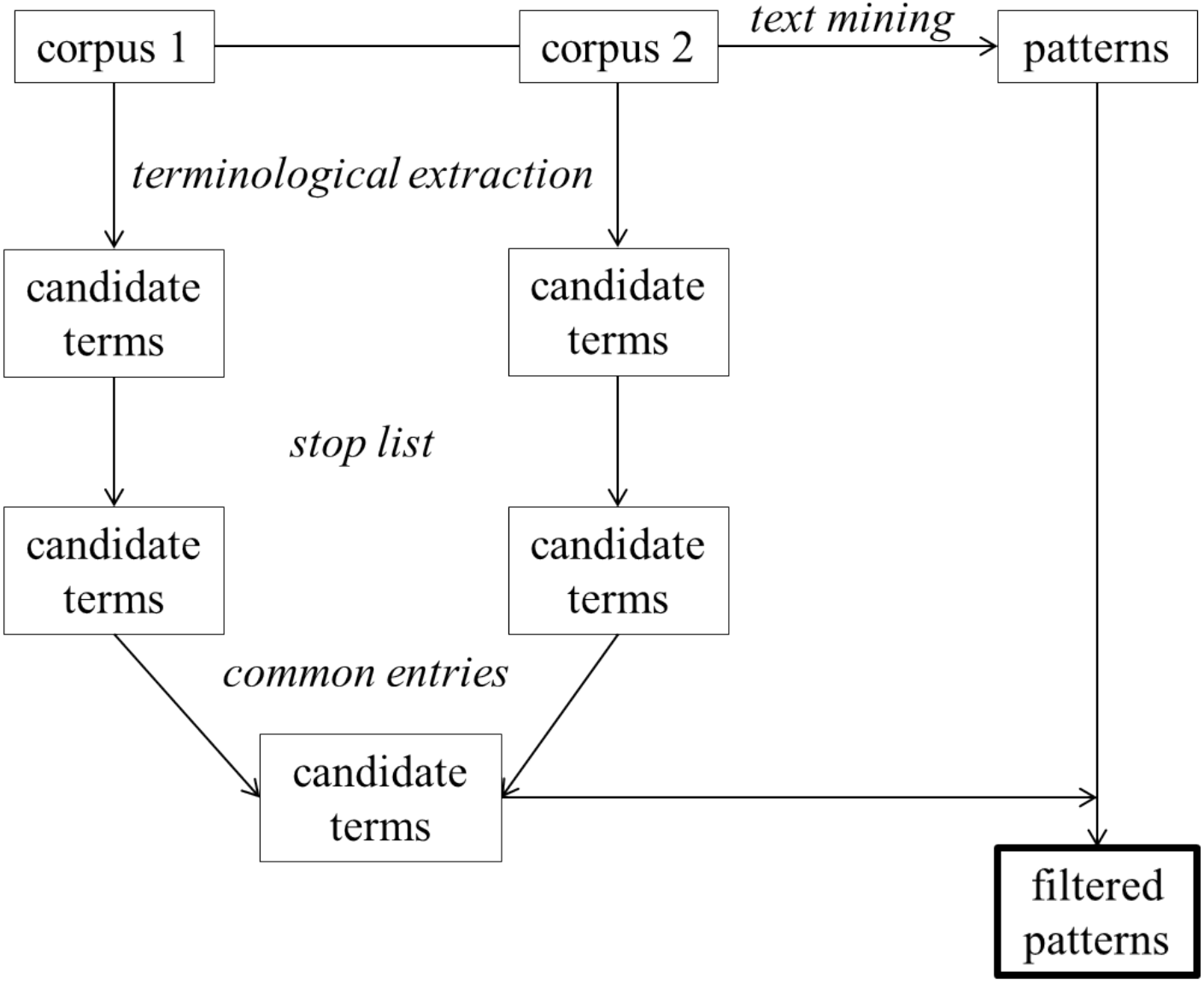
- Utilisé ici pour récupérer des motifs de *lemmes* fréquents (ex. “comme décrire dans le tableau”)
- Longueur variable

14 000 motifs

Pour réduire le nombre de motifs, nous n'avons conservé que ceux contenant au moins un des candidats termes restants
(Hypothèse : les structures basées sur un "terme du genre" sont plus susceptibles d'être elles-mêmes typiques du genre)

Motifs avec au moins un terme : **6 000** (- 8 000)
Exemples : “doit respecter la [contrainte]”, “être connaître avec une [précision] meilleure que (nombre)”

La liste peut être encore réduite en se concentrant sur les motifs contenant un verbe


corpus 1
corpus 2
text mining
patterns
terminological extraction
candidate terms
candidate terms
stop list
candidate terms
candidate terms
common entries
candidate terms
filtered patterns

- Det N1 permettre (de V+nominalisation déverbale)
  “le DUPC permettra de modifier localement les paramètres du calcul”
- Det N1 fournir Det N2 (à Det N3)
  “cette interface fournit les positions navigateur de l’instrument”
- Det N1 utiliser Det N2 (pour V)
  “le système GIDE utilisera le protocole FTP pour effectuer les transferts”
- Det nominalisation déverbale doit s’exécuter (conditions)
  “la consolidation du scénario de travail au CECT doit s’exécuter en moins de 15 secondes”
- sur réception de cette TC, le N1 exécute la procédure de mise ON+OFF de Det N2 (, par l’envoi de commandes (sur+vers+à Det N3)
  “sur réception de cette TC, le LVC exécute la procédure de mise ON de la carte IOT sélectionnée, par l’envoi de commandes discrètes sur l’OBMU”