Checking coverage against corpora

Eva Dauphin, Veronika Lux, Sylvie Regnier-Prost,
Doug Arnold, Lorna Balkan, Frederik Fouvry,
Judith Klein, Klaus Netter, Stephan Oepen,
Dominique Estival, Kirsten Falkedal, Sabine Lehmann

Report to LRE 62-089 (D-WP3.2)
Test Suites for Natural Language Processing (TSNLP)
University of Essex
April 1995

Summary

Title	The Construction of Test Material - Checking against Corpora
Project	LRE-62-089
Document ID	D-WP3.2
WP	The Construction of Test Material
Date	April 1995
Principal Authors	Eva Dauphin, Veronika Lux
Authors	Doug Arnold, Lorna Balkan, Frederik Fouvry, Eva Dauphin, Veronika Lux, Sylvie Régnier-Prost, Judith Klein, Klaus Netter, Stephan Oepen, Dominique Estival, Kirsten Falkedal, Sabine Lehmann
Reviewers	The consortium
Status	public

Checking coverage against corpora

Introduction

TSNLP WP#3.2 task is to check the TSNLP Test Suite against a corpus, so as to report on the possible differences in coverage or on major divergences in the most common phenomena illustrated.

The corpus on which we work is an example of technical French in the aerospace domain. It is a chapter from a helicopter Maintenance Manual, selected as representative of the manual by people in charge of the documentation[1]. A helicopter Maintenance Manual is writen by technical writers, in compliance with the ATA 100 standard[2]. It is read by technicians, or rather *used* by technicians for helicopter maintenance.

The following statistics on the corpus give indications on its size:
- number of sequences[3] : 1 275
- number of words: 8 767
- number of characters: 67 972

This report is divided in two main parts.

The first part is an attempt to look at a technical text with the eye of a TSNLPeer: to begin with (part 1), we used the TSNLP phenomena and domains to built a coarse grid. This grid was used as a means to get a general linguistic overview on the corpus, a rough idea of both the corpus characteristics included in general linguistics and its specific characteristics.

Linguists may object that such a grid is obviously inadequate for technical text description and speak in favour of a dedicated grid. But because our goal was precisely to „check against corpora", we needed a referent that be a representative of „general linguistics". A grid dedicated to technical text would rather be a result of this „checking" task than a starting point.

After this first part, we felt that the TSNLP grid helped us to examine and to present the corpus in a (linguistically) structured way but also confined us to a description level where the corpus particularities were invisible or even led us to conceal some of its characteristics.

Therefore, in a complementary second part (part 2)[4], we describe with more details the „linguistic" resources used in the corpus, point out the main structures used (quantitative data are given when possible) and propose a first analysis of technical writing specific phenomena (including „deviant" structures). This second part organization is guided by both the corpus and the ATA100 (distinction of three different levels[5]: text level, sentence level, nominal phrase level, distinction of descriptions and procedures). Its takes benefits from the work done in the context of her thesis by [*Dauphin, 1995*].

---

1 Because each chapter deals with a different class of operations that apply to the helicopter (chapter ten deals with the storage operations), this corpus has the advantage of being representative of the Maintenance Manual from a syntactical point of view though it is (relatively) restricted in vocabulary (lexically, it is not representative of the Maintenance Manual).

2 The American Transport Association elaborates standards on documentation. One of these standards (ATA100) rules the writing of an aircraft Maintenance Manual (form and content of each chapter, etc).

3 A sequence is a part of text limited by significant punctuation marks (full stop, colon, dash, exclamation mark...).

4 Initially, part 1 was to constitue the entire WP#3.2 report. Then part 2 was started with the aim of replacing part 1. But eventually, the two parts prove complementary.

5 These levels are distinguished by many linguists: for example by [Kocoureck, 1991]

In these two parts, our main goal is to evaluate the adequacy of the TSNLP Test Suite coverage to some real (technical) text. A secondary goal is to find out possible phenomena combinations, in order to add „complex“ examples to the TSNLP Test Suite.

1 - Corpus and TSNLP phenomena

The phenomena and domains distinguished in TSNLP and selected for the WP#3.1 data construction task give a general definition of the TSNLP Test Suite coverage. To compare this coverage with the one needed for corpus analysis, we fill in a grid with examples from the corpus:

- the ten main divisions of this grid correspond to the most common phenomena identified in TSNLP,
- the subdivisions of this grid correspond to the domains of TSNLP (constituents).

Naturally, both examples and the absence of examples are mentioned in each section/subsection of the grid.

As will be seen in part 2, the corpus specific phenomena are not among the phenomena that TSNLP partners set aside while looking for the ten most common phenomena (Ex. extraposition, dislocation, topicalisation...).

1.1. SENTENCES AND CLAUSES TYPES

Examples of positive and negative declarative sentences are:

• positive

(72) Le but des opérations de vérification et d'entretien d'un appareil stocké courte durée est de connaître l'état de l'appareil et la tenue du stockage afin de reconditionner l'appareil pour une durée de stockage.

• negative

(23) NOTA: La configuration „appareil entreposé non stocké“ n‘affecte pas la disponibilité de l‘appareil.

But there are no examples of interrogative sentences and no examples of subject inversion.

1.2. DIATHESIS

Examples for diathesis are:

• active

(2) Des obturateurs munis de flamme protègent les prises totale/statique et les tuyères GTM des agents extérieurs.

• passive

(17) Le choix du type de stockage à réaliser **est conditionné pa**r :
- la durée d'immobilisation,
- les conditions climatiques du lieu.

• reflexive

(28) Le standard de stockage longue durée **se différencie** du standard courte durée par le fait que tous les ensembles mécaniques sont déposés de l'appareil et stockés en containers.

• impersonal

(26) NOTA: Sous les climats maritimes ou tropicaux, **il est recommandé** d‘appliquer la méthode de stockage longue durée afin d‘assurer à l‘appareil une protection maximale.

1.3. MODALITY, TENSE, ASPECT

The corpus conjugated verbs are always in the indicative, and in present or future.

(2) Des obturateurs munis de flamme **protègent** les prises totale/statique et les tuyères GTM des agents extérieurs.

Examples of tensed imperative do not exist, while there are plenty of infinitive imperatives (75% of the verbal forms).

• infinitive

(101) **Essuyer** avec un chiffon imbibé de white spirit les tiges coulissantes des servocommandes.

Modals are found, mainly in passive forms.

(19) Seul un point fixe rotor tournant **doit être effectué** chaque semaine.

Conditional constructions of the corpus are „If Then constructions“, that do not illustrate the tense and modality concordancy phenomena in all its variety (the main clause verb is always in the infinitive, the subclause verb is in the indicative). No conditional construction shows inversion.

(75) **Si** la vérification du stockage correspond avec la périodicité de la pesée, **effectuer** la pesée.

1.4. COMPLEMENTATION

Verbal complementation

As for verbal complementation, various subcategorisation frames can be found in the corpus. To list them all, an unlimited patience or a sophisticated NLP tool would be necessary. Our impression is that the transitive verbs with a direct object only are the most frequent, while the three arguments subcategorisation frames mentioned in the TSNLP „verbal complementation“ section are not illustrated in the corpus.

(45) S‘assurer qu‘il n‘y a aucune fuite sur les circuits.
(79) Faire chuter la pression des circuits.
(32) Débrancher les batteries pour toute la durée de l‘immobilisation de l‘appareil.
(33) Procéder au campement de l‘appareil suivant tâche 10.10.00.681.
(22) Toutefois dans les pays à climat chaud et humide, ou chaud à atmosphère saline, il y a lieu d'éviter l'immobilisation prolongée des pales en plein soleil.

Examples of predicative constructions are:

• AP predicate

(89) Pour réduire l'encombrement et faciliter l'expédition d'un hélicoptère par voie maritime ou par aérotransport la dépose de certains éléments (pales, moyeu rotor, cloison pare-feu, poutre de queue, etc.) **s'avère nécessaire**.

• PP predicate

(112) Rotation des ensembles mécaniques par point fixe si les GTM sont **en place** ou en rotation de 4 à 5 tours à la main.

Pronominal constructions are found.

• reflexive

(102) Vérifier la cinématique des commandes de vol et **s‘assurer** de leur libre débattement.

• reciprocal: no example

Examples of pronominal complements never include several pronominal complements one after the other (the linear order/clitization point mentioned in TSNLP cannot be illustrated).

(110) Débrancher les prises des étoupilles des extincteurs et **les** protéger (toile thermosoudable).

NP complementation

For NP complementation, PPs (argument) and infinitive complements are used.

• PP (argument)

(109) Rotation **des roues.**

• infinitive complement

(17) Le choix du type de stockage **à réaliser** est conditionné par:
- la durée d‘immobilisation,
- les conditions climatiques du lieu.

One example of nominalization that results in a clausal complement (?) is:

(28) Le standard de stockage longue durée se différencie du standard courte durée par **le fait que tous les ensembles mécaniques sont déposés de l‘appareil et stockés en containers.**

AP complementation

The examples of AP complementation that we have found only use PP complements or infinitive complements. There is no case of clausal complement.

• **PP**

(88) les containers **étanches à la vapeur d'eau et aux eaux de ruissellement**

• **Infinitive complements**

(29) Les appareils stockés longue durée ne sont pas **susceptibles d'être utilisés immédiatement** en raison des opérations de protection et de dépose qu'ils ont subies.

Examples of comparative complement are:

(18) Lorsque la durée d'immobilisation (T) est **inférieure ou égale à 2 mois**, aucune opération particulière de stockage n'est nécessaire.

(20) Lorsque la durée d'immobilisation (T) est **supérieure à 2 mois**, l'appareil doit être placé dans l'une des configurations suivantes.

1.5. MODIFICATION

Sentential and verbal modification

Sentential modification uses adverbs and participial constructions.

• adverbs

(67) Vérifier **visuellement** le gonflage des vérins et des pneumatiques.

• participial constructions

(30) **La configuration appareil entreposé non stocké n'affectant pas la disponibilité de l'appareil**, s'assurer que celui-ci est en permanence en état de vol et en parfait état de propreté.

There are many examples that include modifiers of verbs/sentences. We selected the following examples because they show a prepositional complement introduced by AVEC, that appear either after or before the object complement.

(41) Essuyer tous les éléments **avec un chiffon propre**.

(81) Obturer **avec de la bande adhésive** les mises à l'air libre des bâches hydrauliques.

NP modification

NP modification means used in the corpus include adjectives, PPs, relative clauses and participial clauses:

• adjective (in postposition, in anteposition)

(41) un chiffon **propre**

(102) leur **libre** débattement

• PP(specification)

(7) les rails **des sièges pilote et copilote**

• relative clause with extraction from subject or from object

(90) La cellule et les éléments **qui restent à poste sur celle-ci** font l'objet d'un stockage courte durée et sont soumis aux périodicité d'inspection du stockage suivant MMA 10.00.00 paragraphe 3.1.

(29) Les appareils stockés longue durée ne sont pas susceptibles d'être utilisés immédiatement en raison des opérations de protection et de dépose **qu'ils ont subies.**

(No examples of extraction from PP, extraction from NP complement, ... have been founded.)

• other cases of NP modification: participial clause

(76) l'étiquette **précisant le type de stockage**

(85) Renouveler les traitements anti-corrosifs et protections **ayant pu être endommagés au cours de la vérification.**

(100) Eliminer le produit anti-corrosif **protégeant les surfaces non peintes** avec du white spirit.

(93) Des tableaux **renseignés lors de l'expédition et joints au livret de l'appareil** répertorient les éléments déposés et indiquent leurs méthodes de stockage.

(94) Des anneaux et ferrures **fixés sur les cadres** permettent l'amarrage de l'appareil par l'intermédiaire d'amarres disposées de part et d'autre du fuselage.

AP modification

In the corpus, AP modification uses:

• adverbs

(16) une atmosphère **suffisamment** sèche

1.6.AGREEMENT

Agreement in sentences and VPs

Subject-verb and object-past-participle agreement are illustrated in the corpus.

• subject-verb agreement

(23) NOTA : **La configuration "appareil entreposé non stocké" n'affecte** pas la disponibilité de l'appareil.

• with object- past participle agreement

(29) Les appareils stockés longue durée ne sont pas susceptibles d'être utilisés immédiatement en raison **des opérations de protection et de dépose qu'ils ont subies.**

Agreement in NPs

Det-N and N-Adj agreement are shown in NPs. As modification by a participial clause is frequent, the N-past_participle or N-present _participle agreement is also illustrated.

• Det-N agreement

(99) la corde

• N-AP agreement (adjective in postposition or in anteposition)

(41) un chiffon propre

(102) leur libre débattement

• N-relative clause agreement

(90) **La cellule et les éléments qui restent** à poste sur celle-ci font l'objet d'un stockage courte durée et sont soumis aux périodicité d'inspection du stockage suivant MMA 10.00.00 paragraphe 3.1.

(29) Les appareils stockés longue durée ne sont pas susceptibles d'être utilisés immédiatement en raison des opérations de protection et de dépose qu'ils ont subies.

• other NP agreement

(76) l'étiquette **précisant le type de stockage**

(85) Renouveler les traitements anti-corrosifs et protections **ayant pu être endommagés au cours de la vérification.**

(100) Eliminer le produit anti-corrosif **protégeant les surfaces non peintes avec du white spirit**.

(93) Des tableaux **renseignés lors de l'expédition et joints au livret de l'appareil** répertorient les éléments déposés et indiquent leurs méthodes de stockage.

(94) Des anneaux et ferrures **fixés sur les cadres** permettent l'amarrage de l'appareil par l'intermédiaire d'amarres disposées de part et d'autre du fuselage.

Agreement with coordinated structures is shown in 1.8.Coordination.

1.7.WORD ORDER

As there are no examples of several pronominal complements one after the other in the corpus, the word order problems related to clitization are not illustrated by examples from the corpus

Some relevant examples for this „word order" chapter could be:

• NPs including APs (adjective/noun order)

(41) un chiffon propre

(102) leur libre débattement

(22) Toutefois dans les pays à **climat chaud et humide, ou chaud à atmosphère saline,** il y a lieu d'éviter l'**immobilisation prolongée** des pales en plein soleil.

• NPs with several complements and/or modifiers

(22) Toutefois dans les pays à climat chaud et humide, ou chaud à atmosphère saline, il y a lieu d'éviter **l'immobilisation prolongée des pales en plein soleil.**

• clauses including adverbs (verb/adverb order)

(16) La protection des appareils et de leurs différents équipements **est normalement obtenue** par l'un des trois procédés suivants: (...)

1.8.COORDINATION

An extensive use of coordination is made in the corpus.

Coordinated structures exist at most of the constituent levels (only ADVP coordination has not been observed).

- **Sentence coordination**
  (61) S‘assurer qu‘il n‘existe pas de fuite hydraulique et protéger les surfaces non peintes avec du produit anti-corrosion C.999.
- **VP coordination**
  (66) **Nettoyer et lubrifier** (fluide de fonctionnement) les tiges coulissantes des vérins et amortisseurs.
- **NP coordination**
  (83) Fermer **les trappes et portes d‘accès**.
  (65) Nettoyer **les blocs de frein, les pneumatiques**.
- **PP coordination**
  (88) l‘expédition d‘un hélicoptère **par voie maritime ou par aérotransport**
  (53) NOTA: Pour l‘installation „climatisation“ procéder **à la vérification de l‘installation et au nettoyage de la gaine textile de l‘extracteur d‘eau (voir documentation propre)**.
- **AP coordination**
  () les prises **totales et statiques.**
- **unlike constituent coordination**
  (64) les tiges des servocommandes **principales et arrière**
  (5) les roues **principales et avant.**

Extraction from a coordinated structure is illustrated by the following example:

(90) La cellule et les éléments **qui restent à poste sur celle-ci** font l‘objet d‘un stockage courte durée et sont soumis aux périodicité d‘inspection du stockage suivant MMA 10.00.00 paragraphe 3.1.

1.9. NEGATION

Sentence negation/ VP negation

(23) NOTA: La configuration „appareil entreposé non stocké“ n‘affecte pas la disponibilité de l‘appareil.

NP negation (ne..aucun, aucun...ne)

(45) S‘assurer qu‘il **n‘**y a **aucune** fuite sur les circuits.
(59) **Aucune** opération de stockage **n‘**est à effectuer.

AP negation

(61) les surfaces **non peintes**
(AP2) ensembles stockés en containers étanches **non pressurisés**

1.10. OTHER PHENOMENA OCCURRING INSIDE PHRASES

NP specification

Examples of NPs with determiners, demonstratives, possessives, quantifiers, numerals and partitive constructions can be extracted.

- determiners
  (99) la corde
  (34) un container ouvert
  (89) certains éléments
- demonstratives
  (27) Cette configuration
- possessives
  (93) leurs méthodes de stockage.
- quantifiers
  (82) toutes les surfaces métalliques non protégées
- numerals : cardinals
  (1) trois bâches
- numerals : ordinals
  (39) la tâche 71.00.00.051
- partitive construction
  (43) du papier anti-graisse.

Part 1 shows that the TSNLP grid is a relatively adequate mean to describe the corpus. But this positive judgement must be mitigated because in part 1, the corpus was somehow „forced“ into the grid.

Therefore, in the next part, we add some nuances and give attention to technical writing specificities.

2. Corpus description

As TSNLP focuses on syntax, the following corpus description is above all a syntactical one[1]. But we also indulge make:
• some hypothesis on the motivations for the syntactic resources selection observed,
• a short examination of some extra-sentential phenomena.

„Linguistic“ resources used in the corpus are examined at three levels:

(1) at text level.
The text structuration means and the elements used to ensure text coherence are examined.

(2) at sentence level.
In a „checking against corpora“ task, the sentence, as one of the linguistics‘ favorite unit of study, is obviously an interesting level of analysis, though there are relatively few verbal sentences in the corpus (15% of the sequences only).

(3) at nominal phrase level.
In quantity, nominal phrases are predominant as almost 85% of the sequences are nominal sequences (note that this figure does not take into account all the nominal phrases included in verbal phrases).
For syntax, the nominal phrase level is an interesting one: as will be seen, while the syntactic resources available at this level are very limited (in French), technical writing manages to create a deviant syntax by not using these resources.
At a semantic level, terminology gives a special status to some of these nominal phrases. Because we are above all aiming at a syntactic description, we have chosen to consider them as „normal“ NPs (this may appear as an extremist attitude but we want to avoid the sterile solution which consists in simply classifying all complex nominal phrases as terms).

2.1.TEXT STRUCTURATION

We will not go into the complex considerations that guided the structuration of the text examined but only want to mention that:

• the ATA100 standard imposes a logical structure throughout the text, with a hierarchical plan in chapters, sections, sub-sections, etc. Therefore, titles of the corpus can be considered both as meta-linguistic elements and as concise, precise and rational expressions of the content that follows[2]. Complex examples of ellipses can be observed that cannot be dealt with if a title and its

[1] Just a remark on morphology: as many languages for special purposes (especially those for scientific purposes) the „language of aerospace“ creates words or compounds (anémobarométrique, anti-corrosion, anti-graisse, autosoudable, désembuage, hydrotest, physico-chimique, etc.). But its creativity is difficult to define because it does not manipulate Greek and Latin roots (as the language of medicine does, for example). Rather, it imports (and possibly slightly modifies) English words or uses words of the general language but with a specific meaning.

[2] Some titles are not summarizing the text to come but are part of the text itself (ex. adjunct).

corresponding paragraph are not considered as a whole.

• the text is structured by inter-sentential linking elements.
But, because:
- the different part of texts are kept short and separate (a topic is rarely prolonged so as to make the use of anaphora desirable),
- cohesion is largely guaranteed by the logical structure,

there is little place left for the use of binding adverbs typical for „techno-scientific rhetorics“ (ainsi, alors, cependant, ci-après, donc...) and anaphora is rare.

Whenever anaphora occur, they above all use:

(1) a brachygraphical[1] expression (ie. a non linguistic expression, possibly an*ad hoc* one, that refers to a complex linguistical unit. Abbreviations, acronyms and other symbols are brachygraphical expressions.).

Ex.: T compris entre 2 et 12 mois,

when T has been previously defined: Lorsque la durée d'immobilisation (T) est supérieure à 2 mois l'appareil doit être placé dans une des configurations suivantes.

(2) a nominal group including a brachygraphical expression used as an apposition

Ex.: (61) du produit anti-corrosion C.999.

Note that it is an efficient way to avoid ambiguity.

and less frequently:

(3) a personal pronoun (but inter-sentential anaphora rarely uses such pronouns, because, in the restricted person system of the corpus, they are easily ambiguous)

Ex.: (64) Essuyer avec un chiffon propre les tiges des servocommandes principales et arrière. **Les** enduire de liquide hydraulique utilisé.

(4) a demonstrative pronoun (celui-ci, celle-ci, ...),

Ex.: (30) La configuration appareil entreposé non stocké n‘affectant pas la disponibilité de l‘appareil, s‘assurer que **celui-ci** est en permanence en état de vol et en parfait état de propreté.

(5) a nominal group with a definite article or a demonstrative article,

Ex.: (27) Cette configuration est envisagée lorsque la durée d‘immobilisation de l‘appareil est supérieure à 12 mois.

Note: cases of „exaphora“ are found. They are a sort af anaphora that can be resolved by the context only.

EX.: 3.1.1 Dans les 24 heures qui suivent la mise en container
- Vérifier la pression de gonflage du container. Si celle-ci a baissé, rechercher et supprimer la fuite avant de regonfler le container.

[1] see [Kocourek, 1991] for more details

2.2. VERBAL PHRASES, VERBAL SENTENCES

The corpus sentences are characterized:
- by a restricted person system, and consequently, by a restricted pronoun system, where the third person is the only one used. The absence of other person pronouns can partly be explained by the striving towards the impersonal.
- by a restricted tense system, where the present is mainly used (25% of the verbal forms). Some future forms are also found (less than 1% of the verbal forms).
- by a restricted mode system. The infinitive mode, used in procedures, is predominant (almost 75% of the verbal forms are infinitives). The indicative mode is used in description (about 25% of the verbal forms).

Two types of verbal sequences can be distinguished in the corpus:
- (1) definitions or descriptive statements which are (very) long and mainly complex sentences (around 2% of the sequences).
- (2) procedures (ie. simple or complex set of instructions) which are infinitive verbal sequences (around 13% of the sequences) but must be understood as imperative.

As will be seen, the communication situation and the will to be impersonal are two major factors to explain the selection of syntactic resources made by the technical writers.

2.2.1.General characteristics of the verbal phrases/sentences

These sequences are long and often complex sentences.

They include comma-bounded clause or constituents, that „interrupt" the sentence continuity.

> Ex.: (92) Certains éléments déposés**, tels les carénages,** font l'objet d'un conditionnement de transport et ne sont pas soumis à des périodicités d'inspection de stockage.

Coordination is the most obvious complexity factor. It often occurs in different constituents of the same sentences (VP, NP, PP, etc.):

> Ex.: (70) Nettoyer au white spirit **et** graisser avec G.9999 les charnières **et** les mécanismes de fermeture des capots **et** des portes d'accès, **ainsi que** les rails des sièges pilote **et** copilote.

Various adjuncts (in French: „complement circonstanciels") are included:
- time

> (7) NOTA: S'assurer que les essuie-glaces sont en position verticale **avant de poser la housse cabine.**

- intrument

> (10) **A l'aide de la perche de manutention,** engager les embouts d'amarrage de l'ensemble (Y) sur l'extrémité des pales (1) et (2).

- place

> (12) Poser les anneaux d'amarrage (S) **sur fuselage côté gauche et côté droit aux cadres X 17015 et X 5295.**

- condition

> (27) Cette configuration est envisagée **lorsque la durée d'immobilisation de l'appareil est supérieure à 12 mois.**
> (34) **Si nécessaire** déposer les pales principales, suivant sous tâche 62.10.00.031.001 de la tâche 62.10.00.031.001, les placer dans un container ouvert ou sur des supports appropriés dans un local ouvert et bien aéré.

- cause

> (29) Les appareils stockés longue durée ne sont pas susceptibles d'être utilisés immédiatement **en raison des opérations de protection et de dépose qu'ils ont subies.**

- etc.

Syntactic complexity can be related to the will to be concise: the technical author wants to express a specialised content in sentences that are semantically as heavily „loaded“ as possible, but that do not exceed certain limitation on length. Hence the use of brachygraphical signs, of interpolated or juxtaposed or coordinated or subordinate constructions. Concision leads, in particular, to the following (frequently found) constructions, which can be analysed as ellipses of relative clauses:

> Ex.: (61) S‘assurer qu‘il n‘existe pas de fuite hydraulique et protéger les surfaces **non peintes** avec du produit anti-corrosion C.999.

> Only a past participle remains: the rest of the passive verbal construction (ie. the auxilliary *être*) is omitted, as well as the relative pronoun (*qui*). *Non peintes* is a short semantic equivalent for *qui ne sont pas peintes*.

> Ex.: (100) Eliminer le produit anti-corrosif **protégeant** les surfaces non peintes avec du white spirit.

Similarly, the present participle allows the writer to say as much in one word (*protégeant*) as in two (*qui protègent*). Note that the present participle is used as a verbal form, with variations in tense and modality (affectant, étant effectuées, ayant été effectuées, etc.)

> Ex. [111]: NOTA: Toutes ces opérations de déstockage, de vérification et d‘essais étant effectuées, l‘appareil est remis en service.
> Ex.: (30) La configuration appareil entreposé non stocké n‘affectant pas la disponibilité de l‘appareil, s‘assurer que celui-ci est en permanence en état de vol et en parfait état de propreté.

As already mentioned, syntactic complexity can also be related to the communication situation and to the will to be impersonal. Constructions with an impersonal subject:

> Ex.: (26) NOTA: Sous les climats maritimes ou tropicaux, il est recommandé d‘appliquer la méthode de stockage longue durée afin d‘assurer à l‘appareil une protection maximale.

or passives forms:

> Ex.: (17) Le choix du type de stockage à réaliser est conditionné par:
> - la durée d‘immobilisation,
> - les conditions climatiques du lieu.

can be pointed out.
Passive allows the writer to say what must be done without naming the order giver and taker. Passive also has the advantadge of keeping the focus on the object.

A structure often used in the corpus is the list structure, that seems to be both a graphical convenience and a particular syntactic object.

Ex.95 shows a list that may be considered as nothing more than a graphical convention to present a four element coordination[1]. But other examples suggest that lists are more than this.

> Ex.: (95) Suivant les anneaux ou ferrures utilisés, différents modes d‘amarrage sont possibles en fonction du lieu et des conditions météorologiques. On distingue:
> - l‘amarrage normal sur terre,
> - l‘amarrage gros temps sur terre,
> - l‘amarrage normal marine,
> - l‘amarrage gros temps marine.

[1] In litterature, such coordination would be „linearly“ expressed and the writer would add the conjunction ET before the last element of the coordination: *On distingue l‘amarrage normal sur terre, l‘amarrage gros temps, l‘amarrage normal marine ET l‘amarrage gros temps marine*.

Firstly, a list can be either a coordination (Ex. 95) or a disjunction (exclusive or inclusive OU, Ex.16). Its items can be ordered (Ex. 118) or not (Ex.95).

Ex.: (16) La protection des appareils et de leurs différents équipements est normalement obtenue par l'un des trois procédés suivants:
- application d'une couche protectrice anti-corrosive,
- enveloppement de l'élément à stocker sous une barrière imperméable (matériau thermosoudable ou adhérent, housse de protection),
- maintien de l'élément dans une atmosphère suffisamment sèche (container avec agent déshydratant).

Ex.: (118) Nettoyer avec un chiffon propre les tiges coulissantes des vérins et amortisseurs comme suit :
. imbiber de white spirit,
. sécher,
. imbiber de liquide hydraulique.

Secondly, a list item can grow so as to be a coordination or a list itself (Ex. 36, Ex.124). Or a list can be used to express complex conditional procedures (Ex. 124). Or a list item can grow by including some sort of comment (that can be a whole sentence) on the item itself. In such cases, paraphrasing the list by a single level coordination of its items becomes delicate (at the minimum, some parenthesis would be needed).

Ex.: (36) Tous les deux mois:
- brancher les batteries de bord, poser éventuellement les pales suivant sous tâche 62.10.00.031.001 de la tâche 62.10.00.031.061,
- effectuer le dégommage des garnitures des vérins hydrauliques suivant paragraphe 4.5 de la tâche 29.00.00.611,
- effectuer la visite avant le premier vol de la journée suivant tâche 05.21.00.211,
- effectuer un vol d'entretien de 20 minutes,
- effectuer une visite type après dernier vol de la journée suivant tâche 05.21.00.213,
- procéder aux vérifications et remises en état éventuelles.

Ex.: (124)
3.1.1 Dans les 24 heures qui suivent la mise en container:
- Vérifier la pression de gonflage du container. Si celle-ci a baissé, rechercher et supprimer la fuite avant de regonfler le container.

3.1.2 Au cours des inspections périodiques:
- Si la pression du container est normale et l'indicateur d'humidité de couleur bleue, aucune opération n'est à effectuer.
- Si la pression du container a baissé et l'indicateur d'humidité de couleur bleue, remettre le container en pression et effectuer la vérification décrite en 3.1.1.
- Si la pression du container a baissé et l'indicateur d'humidité de couleur rose :
. dégonfler et ouvrir le container,
. examiner l'ensemble stocké,
. s'il n'y a aucune trace de corrosion, remplacer les sachets déshydratant et l'indicateur d'humidité,
. s'il y a des traces de corrosion, déstocker l'ensemble mécanique, éliminer les traces de corrosion, stocker l'ensemble mécanique,
. remplacer les sachets de déshydratant et l'indicateur d'humidité,
. fermer le container, effectuer le balayage (éventuellement) et gonfler le container.

Thirdly list items either play the role of a coordinated constituent in the sentence (Ex.95, Ex.36) or are an explanation (definition in extension) of a sentence constituent (Ex.118, Ex.16). Ex.16 clearly shows that the list (list of items coordinated by OU) is the agent complement, as is the nominal phrase *l'un des trois procédés suivants* .
But a sentence with two agent complements would be an ill-formed sentence with regard to verbal complementation...

2.2.2. Particularities of procedures

Examples:

> Ex.: (5) Caler les roues principales et avant.
>
> Ex.: (6) Poser les houses: MRP (V) entrée d'air (U), poste de pilotage (W).
>
> Ex.: (10) A l'aide de la perche de manutention, engager les embouts d'amarrage de l'ensemble (Y) sur l'extrémité des pales (1) et (2).
>
> Ex.: (34) Si nécessaire déposer les pales principales, suivant sous tâche 62.10.00.031.001 de la tâche 62.10.00.031.001, les placer dans un container ouvert ou sur des supports appropriés dans un local ouvert et bien aéré.
>
> Ex.: (118) Nettoyer avec un chiffon propre les tiges coulissantes des vérins et amortisseurs comme suit :
> . imbiber de white spirit,
> . sécher,
> . imbiber de liquide hydraulique.

The most striking feature of procedures is the use of verbs in the infinitive[1], with the meaning of verbs in the imperative.

This systematic use of infinitive can naturally be related to the will to be impersonal and to „soften" the authority that lies in these text parts (basically, a procedure is an order given to the reader).

As the examples show, not only is the subject not mentioned (in French the imperative mode also allows the writer to meet this requirement, as in the imperative the subject of the verb, who is the order executor, is not explicitly mentioned) but the verb ending is such as to erase all of its subject's features (the subject person and number can still be inferred from the imperative verb ending, they evidently cannot be inferred from an infinitive verb).

Note the high number of nominal phrases that include brachygraphical expressions (Ex.: *MRP*) or numeric references in apposition (Ex.: *tâche 62.10.00.031.001*) or references to non textual objects (Ex.: *entrée d'air (U)*, to be read as *entrée d'air called U on the drawing nearby* or *l'extrémité des pales (1) et (2)* to be read as *l'extremité des pales numered (1) and (2) on the drawing nearby*).

2.2.2. Particularities of definitions and descriptions

Examples:

> (1) Le lot de l'appareil comprend différents obturateurs et bâches destinés à le protéger en extérieur; trois bâches sont prévues pour la cabine, l'entrée d'air et le MRP.
>
> (2) Des obturateurs munis de flamme protègent les prises totale/statique et les tuyères GTM des agents extérieurs.
>
> (3) Des embouts d'amarrage en bout de pale permettent la tenue des pales des amarres fixées à l'appareil par vent supérieur à 40 noeuds et amarres avec perches à partir de 60 noeuds.
>
> (4) Enfin, l'appareil est immobilisé par les cales de roues sur chaque train d'atterrissage.

---

[1] Hence, examples to illustrate subject-verb agreement cannot be found in procedures and examples of verbs in the imperative are not to be found in the corpus.

(26) NOTA: Sous les climats maritimes ou tropicaux, il est recommandé d'appliquer la méthode de stockage longue durée afin d'assurer à l'appareil une protection maximale.

(73) Les opérations consistent:
- à déposer les housses, les obturateurs, les protections par papier anti-graisse et les obturations par papier adhésif,
- à ouvrir ou déposer les portes d'accès, les capots et les carénages afin de:
    - vérifier la tenue des produits de protection anti-corrosifs et le comportement des commandes et circuits,
    - rechercher et éliminer les traces de corrosion et de moisissure.

Definitions and descriptions (around 2% of the sequences) are the most „literary" part of the corpus. They are (often very long) verbal sentences (including lists), including a subject (third person only) and a conjugated verb.

2.3. NOMINAL PHRASES

The great majority of the corpus sequences are nominal sequences (almost 85% of the sequences). Nominal phrases can be independent (titles, and (two) procedures, etc.) or appear in a sentence.

2.3.1.General characteristics of the nominal phrases

Syntactic complexity also shows in nominal phrases, many of which are nominalizations.

Ex.: remise en condition d'un appareil entreposé non stocké

Most of the phenomena considered relevant by TSNLP for Nominal Phrase description are present in the corpus: Nominal Phrases are composed of:

• a noun N (possibly a proper noun or a brachygraphical expression),

Ex. : Généralités
Ex.: le MMA

and, possibly, one or several of the following elements:

• a specifier: determiner, quantifier, demonstrative, possessive, numeral, partitive

Ex.(determiner): (99) la corde
Ex. (quantifier): (89) certains éléments
Ex. (quantifier): (82) toutes les surfaces métalliques non protégées
Ex. (demonstrative): (27) Cette configuration
Ex. (possessive): (93) leurs méthodes de stockage
Ex. (numeral): (1) trois bâches
Ex. (partitive construction) : (43) du papier anti-graisse

• an adjectival phrase (used as epithet or predicate),

Ex.(agreement): eau déminéralisée
Ex.(placement): (102) leur libre débattement.
Ex.(adjective modified by an adverb): (16) une atmosphère suffisamment sèche
Ex.(adjective with a PP complement): (88) étanches à la vapeur d'eau et aux eaux de ruissellement
Ex. (adjective with an infinitive complement): (29) susceptibles d'être utilisés immédiatement
Ex. (comparative complement): (18) inférieure ou égale à 2 mois

• a noun (or a proper noun or a brachygraphical expression, etc.) in apposition

Ex.: type COCOON
Ex.: les obturateurs tuyères
Ex.: du produit anti-corrosion C.623
Ex.: Tâche 00.00.00.111.111
Ex.: le radiateur d'huile BTP

• a prepositional phrase (many PPs complements are enough to create a monstruous NP like "palan électrique à engrenages à chaine à maillons à commande par boîte pendante à boutons poussoirs" [Kocoureck, 1991])

Ex.: Consignes d'utilisation des lubrifiants dans les ensembles mécaniques.
Ex.: Les obturateurs et bandes adhésives des mises à l'air libre des réservoirs.

• a relative clause

Ex. (extraction of the subject) (90) La cellule et les éléments **qui restent à poste sur celle-ci** font l'objet d'un stockage courte durée et sont soumis aux périodicité d'inspection du stockage suivant MMA 10.00.00 paragraphe 3.1.

Ex. (extraction of the object) (29) Les appareils stockés longue durée ne sont pas susceptibles d'être utilisés immédiatement en raison des opérations de protection et de dépose **qu'ils ont subies.**

• an infinitive complement (? No example in the corpus but examples in the manual).

Two more corpus specific NP modifiers are:

• an adverbial phrase

Such phrases can be considered as the result of an ellipsis.

Ex.: visite avant premier vol (hypothesis: elliptical form for *visite qui doit être effectuée avant le premier vol*, or for *visite à effectuer avant le premier vo*l)

Ex.: visite après dernier vol

• a participial clause (past participle or present participle)

Ex.: (76) l'étiquette **précisant le type de stockage**

Ex.: (93) Des tableaux **renseignés lors de l'expédition et joints au livret de l'appareil** répertorient les éléments déposés et indiquent leurs méthodes de stockage.

NPs are often coordinate or juxtaposed. But simple examples of NP coordination is rare, as coordination mainly cooccurs at several levels of constituent.

Ex.: vérification et entretien périodiques des ensembles déposés et stockés

Ex.: conditionnement et stockage d'un appareil expédié par voie maritime ou par aérotransport

Ex.: la BTP, la BTA, la BTI, le radiateur d'huile BTP

NP coordination sometimes uses both AND and OR, which can create interpretation problems (cf. boolean logic).

Ex.: ensembles déposés en containers **ou** sous enveloppe étanche à l'humidité avec déshydratant **et** indicateur d'humidité

2.3.2. Particularities of titles

Ex.N°1 : Tâche 32.00.00.642 Graissage

Ex.N°2 : 1.4. Informations de références

Ex.N°3 : Remise en condition d'un appareil entreposé non stocké

Ex. N°4 : Stockage appareil complet

The omission of prepositions and determiners in a PP complement (last example) occurs in particular in titles (but also in NPs inside clauses).

For the reader, this omission is not a source of ambiguity, because, when using the manual, s/he is facing a particular aircraft and would never think of interpreting the omitted determiner as something other than a definite article (specific, not generic). This is specification by context.

Note that the omission of the preposition and determiner even occurs in nominalizations, with a logic that is hard to guess.

Ex.N°4: Dépose-Pose pompe hydraulique gauche

Ex.N°5: Dépose-Pose d'un élément de transmission arrière, d'un flector

Ex.N°6: Dépose-Pose du palier de transmission N° 1

In Ex.N°4, the preposition „de" and the definite article „la" are omitted. In Ex. N°5, the preposition „d(e)" is not omitted maybe because the PP is a coordinated one and the article „un" is not omitted either, presumably because it is an indefinite article. IN Ex. N°6 neither the preposition nor the

definite article is omitted (du = de + le).

In general language syntax, the omission of prepositions in the examples above is not correct. But the phenomena is frequent in technical language: it is an obvious case of „deviant“ syntax, showing, at the level of NP syntax, the will to be concise (omission of „grammatical“ words to keep just the most „semantically loaded“ words).

The situation is complex for the omission of the specifier because it looks just like a „zero“ determiner[1], which is correct in general language:

• for example titles in general (and this holds for the corpus titles) have no specifier (no specifier for the noun that is the head of the NP),
Ex. : 1.4. Informations de références

• in a noun complement PP, the specifier is often a „zero“ determiner
Ex.: (10) A l‘aide de la perche **de manutention**, engager les embouts d‘amarrage de l‘ensemble (Y) sur l‘extrémité des pales (1) et (2)

• in enumerations, items have a „zero“ determiner
Ex.: (16) La protection des appareils et de leurs différents équipements est normalement obtenue par l‘un des trois procédés suivants:
- **application d‘une couche protectrice anti-corrosive,**
- enveloppement de l‘élément à stocker sous une barrière imperméable (matériau thermosoudable ou adhérent, housse de protection),
- maintien de l‘élément dans une atmosphère suffisamment sèche (container avec agent déshydratant).

So, while the syntactic means available for nominal complementation are limited in general standard French (use of prepositions), technical writing further „simplifies“ NP syntax by simply not using prepositions (and specifiers). A possible explanation for this is the little use for semantic interpretation of the most common French prepositions, DE and À. This „deviance“ makes the NPs particularly ambiguous for „average readers“, who, for example, cannot resort to their knowledge of the domain to understand the NP „vidange groupe réservoirs“ as „vidange de tous les réservoirs du groupe (turbomoteur)“ rather than as „vidange du groupe de réservoirs“. This last example is particularly disturbing: not only are the prepositions missing, but the word order is not as expected.

2.3.3. Particularities of nominal procedures

These are nominalizations to be understood as imperative.

Ex.: (109) Rotation des roues tous les 15 jours.
Ex.: (112) Rotation des ensembles mécaniques par point fixe si les GTM sont en place ou en rotation de 4 à 5 tours à la main.

Note: these two examples are likely to be cases of a support verb ellipsis because Effectuer la rotation des roues tous les 15 Jours. also appears in the corpus.

2.3.4. Particularities of Nominal Phrases inside sentences

Titles often re-occur as NPs in sentences, or rather, slightly modified titles re-occur as NPs in sentences[2].

---

[1] [Maingueneau, 94] The author also mentions „zero“ determiners in the following two cases:
- a predicative NP (ex.: devenir ingénieur),
- certain verbal forms (ex.: faire crédit, donner raison, ...)

[2] This is interesting if we think of building test suites by:
- extracting as many examples as possible from a corpora,
- and artificially deriving the missing item to make the test suite as systematic and complete as possible
because it means that the corpus itself naturally contains „variations“ (for example, the same NP with or without

In these Nominal Phrases the omission of the preposition and/or the specifier is less frequent than in titles. A hypothesis is that such omissions occur when the NP is an exact copy of a title in which this „deviance“ already exists.

Conclusion

This „checking against corpora“ task shows that TSNLP can be useful for an industrial NLP user as a „non esoteric“, „representative“ object of „general“ linguistic: TSNLP provides the users with a reference, in particular for corpora study.

For us, this reference has proved helpful to define:
• which linguistics we could need. For example, we believe that the distinction between nouns (ex.: helicopter) and nouns that are nominalizations (rotation) is pertinent for the corpus.
• what are the corpus specific phenomena, that could be taken into account in TSNLP, possibly after some „linguistization“. Examples are: lists, NP with preposition and/or specifier omission, brachygraphical expressions, etc.

As a reference, the TSNLP Test Suite could also facilitate communication between NLP system users and developers: each of them can express their demand or capacity, using the TSNLP Test Suite as a yardstick.

This „checking against corpora task“ also, hopefully, results in a useful industrial user contribution to TSNLP consisting in:

• an evaluation of the adequacy of TSNLP Test Suite coverage to some real (technical) text.
This report points out parts of the TSNLP „grid“ that suit the corpus well (NP specification, NP modification, NP complementation...) and suggests directions for further refinements (distinction of nominalizations...). It shows parts of the TSNLP „grid“ not pertinent for the corpus analysis (interrogative sentences, linear order/clitization, ...) and structures from the corpus that do not fit in TSNLP (lists, ...).

• suggestions of some possible phenomena combinations (combination of different level constituent coordination ...).
This task proved difficult because most of the corpus sentences or NPs illustrate too many phenomena (and various level phenomena) at the same time. Only few „intermediate complexity examples“ can be directly extracted from the corpus. Others could be simplified corpus extracts. To provide guidelines for such simplification, we would have to find out how inter-dependent the many phenomena occurring in a sentence are.

modifier, the same NP in singular and in plural, etc.).

References

• [Dauphin, 1995], E. Dauphin, Elaboration d'un support de données textuelles bilingues annotées pour l'aide à l'évaluation de la traduction automatique par un utilisateur industriel: l'apport de la pragmatique, thesis, Université Paris VII, 1995

• [Grevisse, 1986] M. Grévisse, *Le bon usage*, 12th edition, Duculot, Paris, 1986

• [Kocourek, 1991] R. Kocoureck, *La langue française de la technique et de la science - vers une linguistique savante-*, Oscar Brandstetter Verlag, Wiesbaden, second edition, 1991

• [Kittredge, 1982] coll. edited by R. Kittredge and John Lehrberger, *Sublanguage: studies of language in restricted semantic domains*, de Gruyter, Berlin, 1982

• [Lerat, 1995] P. Lerat, *Les langues spécialisées*, PUF, Paris, 1995

• [Maingueneau, 94] D. Maingueneau, *Syntaxe du français*, Hachette, Paris, 1994

• [Estival, Falkedal, Balkan et al., 1994], Analysis of Existing Test Suites, Report to LRE 62-089 (D-WP1), TSNLP, University of Essex

• [Vatnsdal, 1987] A. Vatnsdal, *Register analysis: the language of air-traffic control* in „Occasional papers on systemic linguistics", Nottingham, 1987