# Validations informationnelles pour l'organisation des connaissances sur le Community Manager : contexte d'étude en nano-sciences et -technologies

Sahbi Sidhom, Philippe Lambert

# Validations informationnelles pour l'organisation des connaissances sur le Community Manager : contexte d'étude en nano-sciences et -technologies

**Sahbi SIDHOM** (LORIA & Université de Lorraine, France) et **Philippe LAMBERT** (IJL & Université de Lorraine, France)

**Mails :** sahbi.sidhom@loria.fr, philippe.lambert@univ-lorraine.fr

**Résumé :**

La diversité des applications réunies de nos jours sous le terme « industrie de la langue » recouvre plusieurs pistes de réflexion. Cet article de recherche consiste à réaliser une conjoncture entre les « traitements automatiques de la langue naturelle » (NLP) et l'organisation des connaissances (KO) : apport de la ré-indexation sociale dans le domaine d'étude. En conséquent, le contexte du numérique avec le web des usages actuel montre cet intérêt et marque son importance pour la ré-indexation dans les réseaux sociaux. Il s'agit d'un nouvel axe de recherche dans un domaine actualisé pour la réflexion.

L'individu plongé dans une activité d'écriture et sans contraintes de style cognitif, comme le cas sur les réseaux sociaux actuels, préconise au moins l'existence d'une idée mentale dans le contexte. Ce type d'énoncé nous offre la réflexion première à l'étude des structures syntaxiques nouvelles comme objet d'étude dans la linguistique computationnelle. Ainsi, le texte est vu comme un ensemble cohérent d'unités plus ou moins complexes. Chaque unité s'articule avec les autres et contribue à la réalisation d'un équilibre structurel. Egalement, l'écriture sans contraintes cognitives préconise le relâchement des règles et styles sur la réalisation des contenus et intrinsèquement sur les structures syntaxiques employées. C'est la réflexion seconde à l'étude pour examiner les corpus d'analyse et procéder à la découverte dans l'ordre des nouvelles observations sur la syntaxique et la sémantique véhiculée.

Dans cette perspective, le choix porté à un modèle morpho-syntaxique particulier ne doit pas, d'une part, perdre de vue que la qualité des résultats d'un analyseur (parseur) placé dans un système ouvert à l'usage. Et d'autre part, la qualité de conception (formalisme d'implémentation) de l'analyseur doit rapprocher les concepts théoriques et pratiques en accord avec la nature de l'objet d'étude : les enquêtes d'opinion et les ressources ouvertes au web usages.

Le corpus utilisé pour l'observation et la manipulation est constitué de textes de la langue française issus d'une enquête d'opinion associant des questions fermées et ouvertes dans le domaine des nano-sciences et technologies (LNE, C'NANO, Club nanométrologie) en France. Une hypothèse faite sur la nature du texte libre véhiculé dans les questions ouvertes qu'il ne s'agit pas de contraintes de style ou de rédaction proposées ou imposées. La variété de son contenu décrit sa représentativité comme document qui valide des choix sur des questions fermées : pour nous, il s'agit d'un texte libre à préparer pour l'analyse automatique et l'extraction des connaissances.

De manière explicite, la démarche scientifique que nous suivions pour étayer les hypothèses de notre travail, consiste à corréler nos choix théoriques (modèle syntaxique) avec nos observations pratiques (texte libre des usages), par (i) l'élaboration d'hypothèses sur les structures syntaxiques (par l'étude linguistique sur le corpus) ; (ii) la transcription des observations faites sur corpus en système stable (de règles de réécriture) ; (iii) la matérialisation du système de réécriture par l'implémentation de l'analyseur et (iv) l'évaluation de l'analyseur par application directe sur corpus : sa comparaison avec les ressources des usages, aux observations retenues dans les hypothèses de l'étape (i) et la couverture grammaticale de l'étape (ii).

L'objectif de ce travail de recherche est d'apporter un modèle adaptatif d'un analyseur morpho-syntaxique aux ressources ouvertes pour la ré-indexation sociale.

Au terme de ce travail de recherche, nous prendrons appui sur l'analyse de L. Bloomfield[1] (1933) qui soulignait particulièrement que : " *Ce qui concerne le sens est le point faible des études sur le langage, et le restera jusqu'à ce que nos connaissances aient avancé bien loin de leur état actuel* ".

**Mots-clés :**

Web usages, indexation automatique, ré-indexation sociale, TALN, Analyseur morpho-syntaxique, corpus INA, nono-sciences et technologies.

[1] : Citation de Bloomfield Leonard (en 1933, *Language*, New York Press) et reproduite par Bobrow Daniel 1968, in Natural language input for a computer problem solving system, *Semantic information processing*, MIT Press, Cambridge.

## 1. Introduction

La diversité des applications réunies de nos jours sous le terme « industrie de la langue » recouvre plusieurs pistes de réflexion. Cet article de recherche consiste à réaliser une conjoncture entre les « traitements automatiques de la langue naturelle » (NLP : natural language processing) et l'organisation des connaissances (KO : knowledge organization) : apport de la ré-indexation sociale dans le domaine d'étude.En conséquent, le contexte du numérique avec le web des usages actuel montre cet intérêt et marque son importance pour la ré-indexation dans les réseaux sociaux. Il s'agit d'unnouvel axe de recherche dans un domaine actualisé pour la réflexion [14], [15], [16].

L'individu plongé dans une activité d'écriture et sans contraintes de style cognitif, comme le cas sur les réseaux sociaux actuels, préconise au moins l'existence d'une idée mentale dans le contexte. Ce type d'énoncé nous offre la réflexion première à l'étude des structures syntaxiques nouvelles comme objet d'étude dans la linguistique computationnelle. Ainsi, le texte est vu comme un ensemble cohérent d'unités plus ou moins complexes. Chaque unité s'articule avec les autres et contribue à la réalisation d'un équilibre structurel.Egalement, l'écriture sans contraintes cognitives préconise le relâchement des règles et styles sur la réalisation des contenus et intrinsèquement sur les structures syntaxiques employées. C'est la réflexion seconde à l'étude pour examiner les corpus d'analyse et procéder à la découverte dans l'ordre des nouvelles observations sur la syntaxique et la sémantique véhiculée.

Dans cette perspective, le choix porté à un modèle morpho-syntaxique particulier ne doit pas, d'une part, perdre de vue que la qualité des résultats d'un analyseur (parseur) placé dans un système ouvert à l'usage [17], [19]. Et d'autre part, la qualité de conception (formalisme d'implémentation) de l'analyseur doit rapprocher les concepts théoriques et pratiques en accord avec la nature de l'objet d'étude : les enquêtes d'opinion et les ressources ouvertes au web usages [24].

Le corpus utilisé pour l'observation et la manipulation est constitué d'une enquête d'opinion associant des questions fermées et ouvertes dans le cadre de la constitution d'un Club thématique en nanosciences et nanotechnologies coordonné par des grands acteurs du domaine en France : le Laboratoire National d'Essai et de Métrologie (LNE), le Centre de Coméptences NanoSciences France (C'NANO) rassemblant l'ensemble des laboratoires du domaine. Une hypothèse faite sur la nature du texte libre véhiculé dans les questions ouvertes qu'il ne s'agit pas de contraintes de style ou de rédaction proposées ou imposées. La variété de son contenu décrit sa représentativité comme document qui valide des choix sur des questions fermées : pour nous, il s'agit d'un texte libre à préparer pour l'analyse automatique et l'extraction des connaissances [21].

De manière explicite, la démarche scientifique que nous suivions pour étayer les hypothèses de notre travail, consisteà corréler nos choix théoriques (modèle syntaxique) avec nos observations pratiques (texte libre des usages) [21], par :

(i) l'élaboration d'hypothèses sur les structures syntaxiques, qui se concrétise par l'étude linguistique sur le corpus de textes analysés,
(ii) la transcription des observations faites sur corpus en système stable de règles de réécriture grammaticale,
(iii) la matérialisation du système de réécriture par l'implémentation de l'analyseur morpho-syntaxique, et
(iv) l'évaluation de l'analyseur par application directe sur corpus et sa comparaison avec les ressources des usages, aux observations retenues dans les hypothèses de l'étape (i) et la couverture grammaticale de l'étape (ii).

L'objectif de ce travail de recherche est d'apporter un modèle adaptatif d'un analyseur morpho-syntaxique aux ressources ouvertes pour la ré-indexation sociale.

Au terme de ce travail de recherche, nous prendrons appui sur l'analyse de L. Bloomfield2 (1933) qui soulignait particulièrement que : " *Ce qui concerne le sens est le point faible des études sur le langage, et le restera jusqu'à ce que nos connaissances aient avancé bien loin de leur état actuel* ".

En premier lieu d'expérimentation sur corpus, nous avons intégré dans ce travail de recherche une étude sur les résumés de contenu (expérimentation sur un corpus INA), qui représentent des documents multiformes de l'audiovisuel, en adéquation avec une étude linguistique fondée sur la reconnaissance et l'extraction des syntagmes nominaux (SN). Certains de nos objectifs étaient de formuler des régularités structurelles et syntaxiques pour une grammaire d'analyse de textes écrits. Cette source de régularité, dans cette observation cognitive, était la base de la construction de la grammaire de réécriture de notre analyseur morpho-syntaxique.

En second lieu pour la phase opératoire, notre approche a consisté, d'une part, à la recherche de l'optimisation qualitative sur les résultats d'analyse automatique et, d'autre part, à l'extension des capacités du système pour la prise en compte de nouveaux phénomènes : recherche de la robustesse quantitative sur les connaissances du système.

L'aspect adaptatif était développé dans l'objectif de construire l'outil d'analyse pour l'indexation automatique et la recherche d'information vers la réindexation en analysant les traces d'usages sur de nouveaux contenus sans reconfiguration de l'outil.

Dans ces processus d'indexation et de réindexation, nous avons fait recours à la linguistique comme procédé fiable du passage des formes d'analyse textuelles aux codages recherchés.

---

[2] : Citation de Bloomfield Leonard (en 1933, *Language*, New York Press) et reproduite par Bobrow Daniel 1968, in Natural language input for a computer problem solving system, *Semantic information processing*, MIT Press, Cambridge.

Ces précepts linguistiques nous ont permis d'introduire une rigueur auto-suffisante pour catégoriser, regrouper et interpréter des mots liés à leur forme de surface vers le passage formel des représentations qui sont les connaissances.

## 2. Processus d'indexation et extraction des connaissances

Notre étude se focalise sur le discours à travers sa matérialité textuelle et s'oriente vers les unités organiques qui le compose tant sur le niveau intensionnel qu'extensionnel dans le langage de description. Nous ne prétendons pas fournir ici davantage d'orientations théoriques d'ordre linguistique, mais une esquisse de ce processus théoriquement fondé. Nous laissons au spécialiste-linguiste un grand nombre de décisions que nous ne pouvons prendre à sa place et nous retenons ses recommandations. Nos orientations conditionnent le fondement théorique de l'approche, qui est décrite dans les travaux de S. Sidhom [20], [21], sur le statut du descripteur (pour le modèle d'indexation) vs. Le statut du syntagme nominal (pour le modèle de langage). Pour les aspects pratiques, la deuxième phase de cette analyse constitue à cet égard le véritable enjeu pour faire le passage du modèle du langage (par l'extraction des structures morpho-syntaxiques) au modèle d'indexation (par la représentation ses structures sémantiques tels les SN et propriétés).

L'existence physique d'un document et sa pérennité devrait disposer des informations pour son intégrité, pour son indexation et pour faciliter la tâche de le retrouver et de le consulter [1], [2]. Dans un tel environnement, une synthèse professionnelle sur l'analyse de contenu (documentaire, professionnelle ou sociale) s'avère nécessaire. Les technologies permettant la manipulation et l'extraction automatique des connaissances seront amenées à jouer un rôle essentiel dans la société de l'information [6], [7]. Il est évident qu'un document, auquel il ne serait pas attaché des indications (traces d'usage) permettant sa lecture dans sa complétude (analyse, indexation et réindexation), ne saurait franchir le cercle de sa réutilisation ou simplement de sa consultation [10], [11].

Pour l'indexation, les principales étapes du processus consistent en un recueil de notices documentaires incorporant des résumés de contenus chez des fournisseurs spécialisés dans ce domaine.Notre collaboration scientifique avec des spécialistes de l'INA (Institut National de l'Audiovisuel en France), a révélé une expérience professionnelle et des acquis qui datent des années cinquante.

Actuellement, la formulation des résumés sur des contenus documentaires hétérogènes est construite, dans certains organismes spécialisés comme l'INA, selon des critères et des méthodes formelles acquises par l'expérience [3], [4], [5]. Cela permet d'assurer une régularité et une constance des traitements réalisés par les documentalistes. Cette richesse documentaire, une fois construite et mise à l'exploitation selon des traits attachés au contenu (capitalisation des sources d'information et de connaissances), pourra s'adapter aux diverses technologies d'exploitation et de diffusion [12], [13], [18].

### 2.1. Observations cognitives sur corpus

Nous avons cherché à étudier la stabilité des descriptifs textuels (ou résumés de contenu), tout particulièrement ceux dans les corpus INA (sources de INAthèque puis INAactualités), afin d'établir par une analyse statistique les composantes grammaticales et syntaxiques de la phrase (cf. Tabs. 1., 2.).

Lors de l'analyse, plusieurs situations se présentent où le repérage des syntagmes nominaux n'est pas toujours évident. Cela arrive parce qu'il y a des éléments anaphoriques, des ellipses, des syntagmes nominaux cachés, des syntagmes nominaux avec le déterminant zéro, etc.

Ainsi, il a fallu adopter quelques règles afin d'extraire les syntagmes nominaux de façon homogène pour obtenir des résultats statistiques cohérents dans un objectif précis : établir une grammaire de réécriture fondée sur les corpus. Tout en sachant que les corpus sont développés par des professionnels en texte libre et sans contraintes rédactionnelles.

Une manière de résoudre ces problèmes était de s'occuper seulement de l'extraction des syntagmes de surfaces « complets » sans traitement des cas anaphoriques, élliptiques, ou cachés. Seuls les SN avec déterminant zéro sont pris en compte, car nous supposons la facilité de remédier à ce type de problème lors de l'implémentation de l'analyseur morpho-syntaxique [21].

Les structures syntaxiques qui ont subi cette étude sont les syntagmes nominaux complexes (les SN maximaux, en abrégé SN_max), les syntagmes nominaux simples ou inclus dans les SN_max (les SN inclus, en abrégé SN_inc), les syntagmes prépositionnels (SP), les expansions prépositionnelles (EP) et les phrases relatives (REL) complétives dans un SN.

### 2.2. Identification d'une grammaire cognitive

Le corpus de départ est constitué d'environ 100 notices bibliographiques sur des documents audiovisuels INA (sur des émissions radio et télévision) et qui a été étendu graduellement jusqu'à 300 notices. Chaque notice de l'INA contient au moins deux champs résumés (chapeau pour résumé synthétique et résumé pour les descriptions détaillées) produits par les professionnels [21].

Une synthèse sur les premières données statistiques obtenues montre une forme d'homogénéité dans les structures syntaxiques composites dans la phrase, en analysant les phrases dans les parties : résumé court (chapeau) et résumé détaillé (résumé) :

| OBJET ETUDE | SN_max | SN_inc | SP | EP | REL |
|---|---|---|---|---|---|
| **chapeau** | 2.56 | 4.30 | 4.01 | 0.38 | 0.37 |
| **résumé** | 2.05 | 4.37 | 3.35 | 0.66 | 0.46 |
| **phrase** | **2.30** | **4.33** | **3.68** | **0.52** | **0.41** |

Tab.1 : Analyses structurelles de la phrase résumé INA.

A la suite de cette première analyse purement statistique, nous avons étendu l'étude pour proposer trois modèles possibles de la construction de phrase INA : la phrase à minima de structures syntaxiques (phrase min.), la phrase à maxima de structures et la phrase modèle (phrase). L'étude permetd'envisager les structurations suivantes :

| MODELE | SN_max | SN_inc | SP | EP | REL |
|---|---|---|---|---|---|
| **phrase min.** | 2 | 4 | 3 | 0 | 0 |
| **phrase max.** | 3 | 5 | 4 | 1 | 1 |
| **phrase** | **2** | **4** | **4** | **1** | **0** |

Tab.2 : Modèle de la phrase résumé INA.

En synthèse, l'analyse statistique sur le corpus de notices a révélé une stabilité grammaticale dans les descriptifs textuels (résumés). Cette révélation grammaticale cache en réalité une stabilité de rédaction des textes par les professionnels. Nous avons observé [20], que les documentalistes de l'INA lors de la rédaction des résumés n'ont pas de contraintes rédactionnelles, structurelles ou syntaxiques à respectes, si ce n'est d'appliquer la grille d'analyse de contenu adaptée à un type de document audiovisuel.

### 2.3. Organisation des structures orientée modèle de la phrase

Dans ce qui suit, nous présentons les différentes règles morpho-syntaxiques qui ont servi à l'implémentation de l'analyseur fondé sur la grammaire cognitive du corpus [21].

**a- Structures préfixées PI à la phrase : Proposition Introductive (PI)**

| PI | Exemples |
|---|---|
| 1. SP, ⊂ S | Pour les 20 ans d'AIRBUS INDUSTRIE, … |
| 2. EP, ⊂ S | En parallèle, … |
| 3. EP + SP, ⊂ S | En direct depuis l'observatoire de Meudon, … En compagnie de Marianne GRUNBERG-MANAGO, … |
| 4. PPas + SP, ⊂ S | Embarqués à bord de l'astrolabe depuis l'extrême sud de l'Australie, … |
| 5. PPré + SN, ⊂ S | Proposant un voyage à travers les sites industriels de France, … |
| 6. Prép(en) + SNdat, ⊂ S | En juin 1986, |

| | |
|---|---|
| 7. Prép(en) + PPré + SP, ⊂ S | En passant [par (la littérature)], … |
| 8. Conj, ⊂ S | Cependant, … |
| 9. Conj + Adv + SP, ⊂ S | Car contrairement aux Américains, … |
| 10. « en » + PPrés, ⊂ S | En vaccinant, |

**b- Structures du SN dans la phrase : Syntagme Nominal (SN)**

| SN | Exemples |
|---|---|
| 11. SN (détails sur SN dans [21]) | Le lac … ⊃SN<br>le lac dans le nouveau Québec (…)⊃SN |
| 12. EP ⊂ SN | ∪ne équipe ⊃de tournage …<br>Un avion Hercule ⊃de transport stratégique |
| 13. SP ⊂ SN | La présence ⊃d'un lac … |
| 14. { SN, SP, EP} ⊂ SN | L'utilisation ⊃d'images de synthèse … |
| 15. REL (relative explicative) ⊂ SN | La présence d'un lac ⊃qui se serait formé suite à la chute d'une météorite … |
| Exceptions : | |
| 16. SN$^{\nabla}$= SN sans déterminant | Psychologues et physiciens (se penchent sur leurs multiples facettes.) |

**c- Structures du REL dans la phrase : Phrase Relative (REL)**

| REL | Exemples |
|---|---|
| 17. /REL = Prel + SN/ ⊂ SN | … ,qui + son père, … |
| 18. /REL = Prel + SV/ ⊂ SN | … qui + se serait formé suite à la chute d'une météorite … |
| 19. /REL = Prel + S/ ⊂ SN | a) … qu' + il a réalisé sur le même sujet en 1973.<br>b) … dont + le pouvoir suggestif déborde largement le cadre du bâtiment lui-même. |

**d- Structures du SV dans la phrase : Syntagme Verbal (SV)**

| SV | Exemples |
|---|---|
| 20. V + (Prép + V-inf)+ SP | …**est** (de récupérer) de la matière cosmique |
| 21. V + (Prép + V-inf)+ SN | …**sont montrées** (pour comprendre) les difficultés techniques et économiques |
| 22. V + (V-inf) + SN | …**a pu** (rencontrer) AIRBUS INDUSTRIE |
| 23. V + (V-inf) + (Prép + V-inf)+ SN | …**devait** (permettre) (d'identifier) le sexe |
| 24. V + (PPrés) + SN | …**a suivi** (durant) trois semaines les activités d'une équipe |
| 25. V + SN | …**sont** le reflet de notre société |
| 26. V + SP | … **est réservée** aux avions Hercule |
| 27. V + {SN, SP, EP, PV} | … **essaie** d'expliquer le mystère de l'étoile de |

| | |
|---|---|
| | Bethléem |
| 28. V + (Adv) + SN | …**explique** (comment) les pays européens exportent des armes |
| 29. V + (Adv) + V | …**sont** intimement **liées**<br>…**est** ainsi **développé** |
| 30. V + (Adv) + (Prép + V-inf)+ SN | …**s'attache** (plus) (à expliquer) la course du côté soviétique |
| 31. V + /EP/ + SN | …**démontre** /en particulier/ la polititique de la France à ce sujet |
| 32. V + /Conj/ + SN | …**poursuit** /donc/ cette balade à la fois historique, sociologique et architecturale |
| 33. V | Ce chien **mord** |
| 34. V + (Adj) + SP | …**furent** (découvertes) en 1988 |
| 35. V + {Adv, Adj} | (il) **résout** scientifiquement… |

La construction de la phrase (S) selon notre étude s'articule autour de trois structures fondamentales, à savoir : – une structure qui précède la phrase (une proposition introductive à S), – le syntagme nominal sujet de S (sous forme d'un SN complexe ou SN_max), – le syntagme verbal de S, et – la phrase relative (REL) en option. Chacune de ces structures est identifiée en ses éléments avec son organisation morpho-syntaxique composite :

$$S \rightarrow [PI] + SN + [REL_{SN}] + SV + [REL_{SV}]$$

$$[x]: \textit{ élément facultatif}$$

Nous considérons que ce modèle de grammaire syntagmatique comme modèle « cognitif » de la phrase pourra nous servir à la fois comme outil d'indexation ou d'aide à la rédaction de textes et intrinsèquement à l'orientation de son usage pour la réindexation.

### 3. Processus de ré-indexation : application aux nanosciences et nanotechnologies

En application aux domaines des nanosciences et nanotechnologies, un nouveau corpus utilisé pour dépasser l'observation, la manipulation a été constituée à partir d'une enquête d'opinion associant des questions fermées et ouvertes dans le domaine. L'enquête avesoins des membres d'une nouvelle structure collaborative, le Club de nanoMétrologie, mise en place par le consortium des laboratoires en nanosciences et nanotechnologies France (C'Nano) et le Laboratoire National d'Essai (LNE). Une hypothèse faite sur la nature du texte : il est (i) libre, (ii) véhiculé dans les questions ouvertes et (iii) qu'il ne s'agit pas de contraintes de style ou de rédaction proposées ou imposées. La variété de son contenu décrit sa représentativité comme document pour valider nos choix sur la robustesse de la grammaire cognitive implémentée. Pour nous, il s'agit d'un texte libre à préparer pour l'analyse automatique afin de construire l'extraction des connaissances de type SN et propriétés.

Nous avons démontré la réutilisabilité du modèle de grammaire cognitive en changeant le contexte d'étude sur corpus (cf. modèles S et S' / S' $\subset$ S) tout en préservant les propriétés de cette grammaire et intrinsèquement le modèle de langage implémenté.

Lors de l'analyse des corpus d'enquête, les structures identifiées en ses éléments avec son organisation morpho-syntaxique composite se traduit par la sous-grammaire :

$$S' \rightarrow [V\inf] + SN + [REL]$$
$$[x]: \textit{ élément facultatif}$$

L'extension apporté au modèle de langage implémenté à l'origine (S) requière de rajouter simplement une nouvelle règle dans PI qui s'exprime par : Vinf $\subset$ PI.

### 3.1. Implémentations et validations

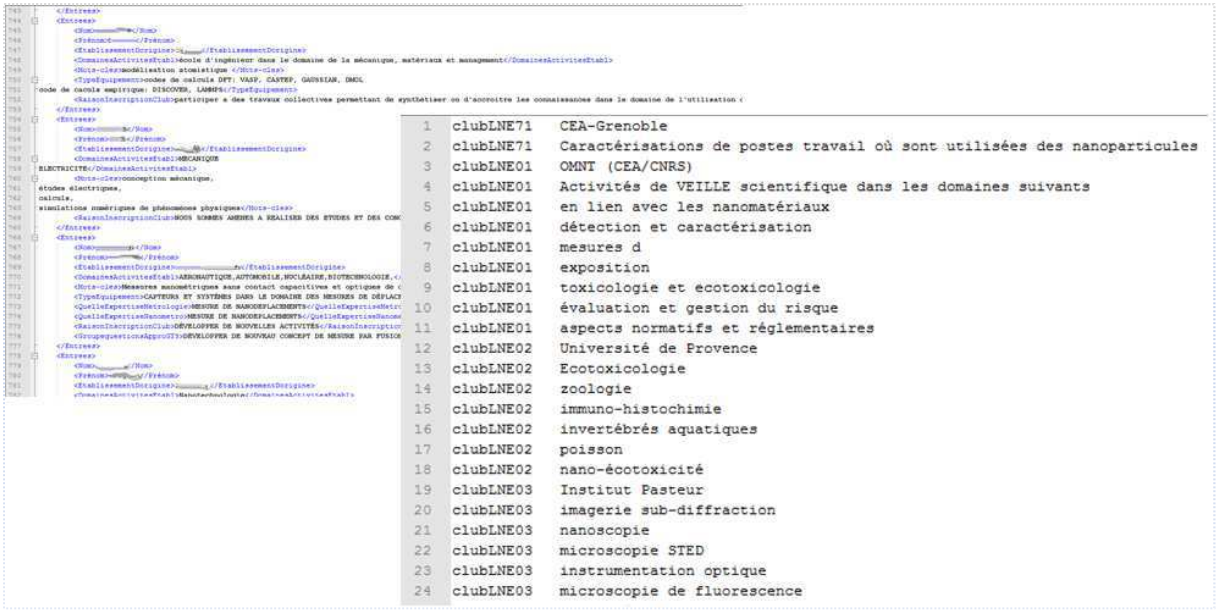
NooJ est un environnement linguistique développé par Max Silberztein (2005) de l'université de Franche-Comté (France) [25], [26]. NooJ est fondé sur la technologie .NET et reconnaît un grand nombre de formats de documents. Outre cet avantage, l'utilisation de l'outil est facile avec une prise en main relativement rapide. Dans l'optique d'un repérage terminologique pour notre corpus bibliographique, NooJ offre la possibilité de créer des grammaires locales (ie. automates finis) complètement paramétrables pour l'extraction d'informations. Les ressources de NooJ sont principalement constituées de dictionnaires (de la langue) et de graphes syntaxiques de type transducteurs à état fini, permettant le repérage d'expressions complexes, l'extraction de lemmes et l'annotation automatique de ressources textuelles [23].

Le questionnaire sur lequel s'appuie ce travail de recherche visait à identifier les raisons de l'adhésion des membres au Club de NanoMétrologie, une nouvelle structure collaborative pilotée par le LNE et le C'NANO, de mieux connaître leur besoin en information et ce qu'ils attendaient d'un tel projet. Une trentaine de questions constituaient le questionnaire. Son élaboration a été faite en étroite collaboration avec les différents comités du Club.

La méthode retenue pour toucher le plus d'adhérents et s'assurer d'avoir un maximum de réponses consistait à implémenter sous le logiciel open souce LimeSurvey (*www.**limesurvey**.org*) un questionnaire en ligne sécurisé pour recueillir les réponses des adhérents. Nous nous concentrerons pour cette étude de cas sur deux principales questions ouvertes qui ont fait l'objet d'un traitement automatique, à savoir : (i) « Quelles sont les raisons pour lesquels le répondant a adhéré au Club ? » et (ii) « Qu'est-ce qu'il attend spécifiquement d'une telle structure collaborative ? ». Une centaine de répondants ont contribué aux réponses de l'enquête et spécifiquement aux deux questions ouvertes.

La méthodologie retenue pour le traitement des réponses comprend cinq phases : (i) la sélection des données, (ii) Le nettoyage des données, (iii) l'élaboration des ressources linguistiques ad-hoc, (iv) le traitement des données, (v) l'analyse des résultats.

Le système LimeSurvey permet le déchargement des réponses au format csv. Pour la facilitation du traitement de la centaine de réponses obtenues, nous avons opté pour un reformatage du fichier de réponse au format XML. Ce choix correspond à deux critères spécifiques de traitement : (i) obtenir une meilleur structuration du document source permettant de retrouver plus aisément les réponses d'origine et (ii) donner la possibilité au système d'itérer sur ce fichier avec des traitements spécifiques selon les nœuds xml que nous voulons étudier dans leur particularité. Le fichier source a également été découpé en autant de textes que de réponse, constituant ainsi un corpus d'une centaine de fichiers xml. Cela a permis de retrouver aisément la source documentaire de l'extraction des granules ou lexèmes durant le processus de TALN (cf. Fig.1).



```
clubLNE71   CEA-Grenoble
clubLNE71   Caractérisations de postes travail où sont utilisées des nanoparticules
clubLNE01   OMNT (CEA/CNRS)
clubLNE01   Activités de VEILLE scientifique dans les domaines suivants
clubLNE01   en lien avec les nanomatériaux
clubLNE01   détection et caractérisation
clubLNE01   mesures d
clubLNE01   exposition
clubLNE01   toxicologie et écotoxicologie
clubLNE01   évaluation et gestion du risque
clubLNE01   aspects normatifs et réglementaires
clubLNE02   Université de Provence
clubLNE02   Ecotoxicologie
clubLNE02   zoologie
clubLNE02   immuno-histochimie
clubLNE02   invertébrés aquatiques
clubLNE02   poisson
clubLNE02   nano-écotoxicité
clubLNE03   Institut Pasteur
clubLNE03   imagerie sub-diffraction
clubLNE03   nanoscopie
clubLNE03   microscopie STED
clubLNE03   instrumentation optique
clubLNE03   microscopie de fluorescence
```

Fig. 1 : exemple du fichier XML des réponses du questionnaire.

Le nettoyage des données a consisté essentiellement à corriger les fautes d'orthographes et de nettoyer chaque nœud de tout caractère susceptible de créer du bruit dans les résultats d'analyse du système.

L'élaboration des structures linguistiques a été faite en deux étapes : le premier temps a consisté à créer des dictionnaires spécifiques thématiques sur les nanosciences et nanotechnologies. NooJ offre en cela des fonctionnalités utiles en créant automatiquement un dictionnaire constitué des entrées étiquetées comme inconnues (<UNKNOWN>) donnant ainsi l'opportunité de les intégrer au dictionnaire existant. Ainsi, un dictionnaire de plusieurs centaines d'entrées a pu être rapidement créé, rassemblant les thématiques du club métrologie, les techniques et les types de mesure spécifiques pour les nanosciences. Le deuxième temps a concerné la création d'automates à état fini permettant d'extraire les données spécifiques et leur reformatage pour le traitement ultérieur (cf. Fig.2).

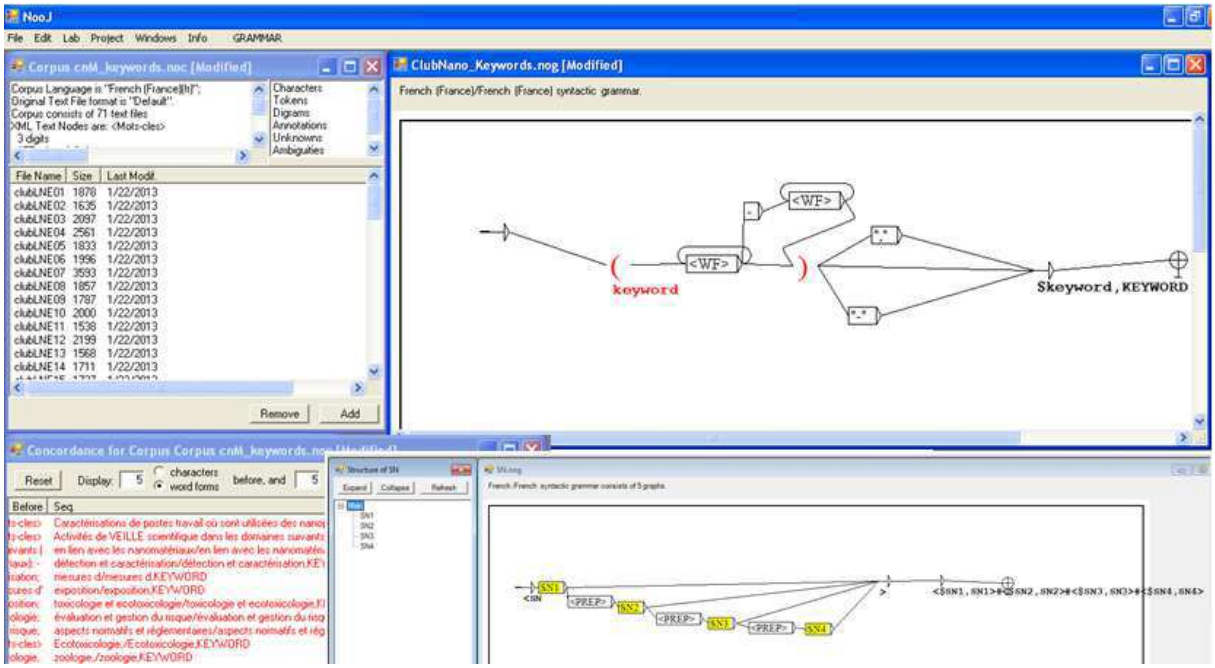



Fig. 2 : exemple d'un automate à état fini pour l'extraction des SN en cascade.

Sous NooJ, l'utilisation des automates se fait en cascade. L'itération permet l'étiquetage des structures retenues sur plusieurs niveaux ce qui permet une extraction fine.

La phase de traitement des données a porté sur l'étiquetage des nœuds « Quelles sont les raisons de votre adhésion ? » et « Qu'attendez-vous du club ? » puis sur l'extraction des structures de type S'::= <Verbe> + <SN emboîtés>. On a ainsi obtenu une vingtaine de résultats significatifs montrant que les raisons de la participation au Club sont circonscrites dans un même champ sémantique (i.e. le réseautage) avec la majorité des réponses s'inscrivant dans une logique « *de création d'un réseau* », « *d'intégration d'une communauté* », « d'indentification de Community Manager », etc. Le fait que NooJ dispose des fonctionnalités de traitement statistique, lui conférant le statut de système hybride, a permis la hiérarchisation des résultats en se basant sur l'attribution pour chaque extraction d'un indice TF-IDF. Le résultat est une série de besoins exprimés par les participants, par ordre d'importance, correspondant au poids sémantique détecté par le système.

### 3.2. Principaux résultats : le « Community Manager »

Les principaux résultats de cette démarche tiennent de trois ordres :

Le premier consistait à traiter les réponses ouvertes d'un questionnaire dédié à l'identification des besoins en information des membres et des raisons qui les avaient poussées à devenir membre du Club de nanoMétrologie. La hierarchisation des réponses à ces questions par l'intermédiaire de processus hybride de TALN et statistique en NooJ, ramène des résultats plutôt bons, avec peu de bruits. L'identification a ainsi pu être faite relativement facilement.

Le second objectif concerne davantage l'aspect diagnostique de notre approche, en nous permettant d'obtenir une image à un temps T de la structuration du Club de nanométrologie. Pour cela, une cartographie en réseau a été réalisée (cf. Fig.4). Cette projection permet dans un premier temps d'identifier le positionnement des acteurs par rapport à la thématique générale du Club (ou notion de centralité) puis dans un deuxième temps, d'identifier les signaux faibles, c'est-à-dire les thèmes complétement excentrés mais qui peuvent se révéler déterminants pour l'évolution du Club. Enfin, ce graphe est également conçu dans une logique de diagnostic [22] : il permettra de comparer les clubs sur une échelle temporelle à T+24 mois et de diagnostiquer quelles thématiques sont les plus fortes (nombre de création de liens entre les acteurs, les thématiques, etc. [8]) et celles sur lesquelles un effort spécifique doit porter pour améliorer leur représentativité et tendre ainsi vers l'exhaustivité (cf. Fig.3).

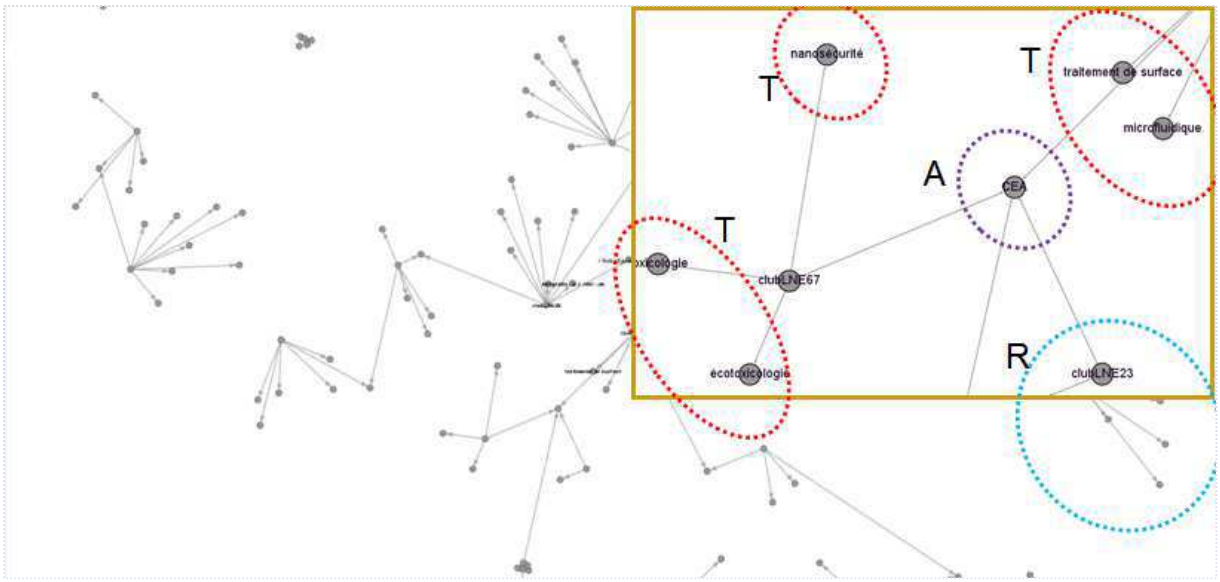



Fig. 3 : cartographie des acteurs associés aux thématiques du Club de NanoMétrologie : trilogie (A : acteurs, T : thématiques, R : ressources)

## 4. Conclusion

Dans la synthèse des objectifs construits dans ce travail de recherche, nous avons réussi à apporter un modèle adaptatif d'un analyseur morpho-syntaxique aux ressources ouvertes pour la ré-indexation sociale. Le formalisme d'implémentation en NooJ et les analyses automatiques nous ont beaucoup rapproché aux concepts théoriques et leur évolution en pratique en accord avec la nature de l'objet d'étude : des enquêtes d'opinion et des ressources ouvertes au web usages dans le domaine des nanosciences et technologies : le club de nanométrologie impliquant plusieurs organismes de recherche en France.

Dans les résultats d'analyse, nous démontrons une conjoncture entre les « traitements automatiques de la langue naturelle » (TALN ou NLP pour natutal laguage processing) et l'organisation des connaissances (KO : pour Knowledge Organization). Egalement, des

valorisations observées par la ré-indexation sociale au travers de nouveaux concepts. Deux questions ouvertes ont fait l'objet d'un traitement automatique spécifique, à savoir : (i) « Quelles sont les raisons pour lesquels le répondant a adhéré au Club ? » et (ii) « Qu'est-ce qu'il attend spécifiquement d'une telle structure collaborative ? ». A l'issu des traitements et analyses du questionnaire qui a concerné une centaine de répondants. Des recommandations en matière d'aide à la décision ont pu être proposées pour le rapprochement des activités, des projets et des acteurs associant des compétences. Ces résultats soulignent la nécessité d'une activité de Community Management. L'intérêt de cette pratique renforce la proactivité des acteurs [9],[22] ainsi que leur cohésion pour l'émergence de nouveaux projets d'appels d'offre en nano. par le rapprochement en activités et des compétences [8].

La méthodologie développée permet également le diagnostic de la structure interne du réseau. La détection de la nature hétérogène du réseau peut ainsi être mise à profit pour effectuer un ré-équilibrage autour du centre de gravité de la structure (ou le noyau de cohésion du réseau).

Sur une échelle temporelle à T+24 mois, il est essentiel de diagnostiquer comment évoluent les thématiques les plus fortes du réseau, mais également de considérer la place des réseaux secondaires (ou les réseaux associés au noyau) pour activer la création de liens entre les acteurs, les thématiques et les projets dans une démarche d'innovation.

En perspective, d'autres enquêtes complèteront ce travail sur l'identification de Community Manager en réponse à un besoin direct ou indirect de type projet. Il sera question de mettre en œuvre les valorisations attendues aux nouvelles thématiques émergeantes (en signaux faibles) et les compétences existantes par les acteurs et ressources.

## Bibliographie


[1] Bachimont B. (1999). La documentation au coeur du processus de production. in Dossier de l'Audiovisuel, Janvier-Février 1999, n°83, INA-Publications, p.38-39.

[2] Bachimont B. (2004). Signes formels et computation numérique : entre intuition et formalisme. In H. Schramm, L. Schwarte & J. Lazardzig (Eds.), Instrumente in Kunst und Wissenschaft - Zur Architektonik kultureller Grenzen im 17. Jahrhundert. Berlin: Walter de Gruyter Verlag.

[3] Browne G. (1996). Automatic indexing and abstracting. in Indexing in Electronic Age Conference, Robertson, NSW 20-21 April 1996, Australian Society of Indexers, 8p.

[4] Carbonell J.G., alii. (1997). Translingual Information Retrieval: a comparative evaluation. in Proceedings IJCAI-97, Nagoya, Japan, Morgan Kaufmann, San Mateo, CA (1997).

[5] Champenier T., Pautet D. (1996). Mise à disposition à travers le réseau Internet de la littérature grise produite à l'INSA. Projet de PFE 1996 , INSA-Lyon, 65p.

[6] Guimier-Sorbets A-M. (1993). Des textes aux images : accès aux informations multimédias par le langage naturel. Documentaliste – Sciences de l'information, 1993, vol.30, n°3, p.127-134.

[7] Régimbeau G. (1998). Accès thématiques aux d'art contemporaines dans les banques de données. in Documentaliste Sciences de l'Information, Volume 35, n°1, janvier 1998, p.15-23.

[8] Lambert P., Sidhom S. (2011). Problématique de la veille informationnelle en contexte interculturel : étude de cas d'un processus d'identification d'experts vietnamiens". in Proceedings : ISKO-Maghreb'11 – Concept and Tools for Knowledge Management (KM). ESCE-University of la Manouba Edition. Hammamet (Tunisia) May. 2011.

[9] Lambert P., Sidhom, S. (2010). Vers le Design d'information pour valoriser les résultats d'une veille sur les maladies chroniques. in Proceedings: Journée d'étude sur la "Mutualisation des ressources documentaires : Hétérogénéité des ressources et accessibilité dans un espace collaboratif." ELICO - Université Jean Moulin Lyon3, 05/11/2010 Lyon (France).

[10] Maniez J. (1993). L'évolution des langages documentaires. Documentaliste et Sciences de l'information, 1993, vol.30, n°4-5, p.254-259.

[11] Harbaoui A., Ghenima M., Sidhom S. (2009). Enrichissement des contenus par la réindexation des usagers : un état de l'art sur la problématique. in 2nd International Conference on Information Systems and Economic Intelligence - SIIE 2009 vol.1 (2009) pp. 932-942 IHE Edition Tunis.

[12] VAN SLYPE G. (1987). Les langages d'indexation : conception, construction et utilisation. in dans les systèmes documentaires Paris : Editions d'organisation, 1987. 277 p. - (Systèmes d'information et de documentation).

[13] Maret P., Pinon J-M., Martin D. (1994). Capitalisation of consultants' experience in document drafting. Conference Proceedings RIAO 1994, Printed by CID Paris France, p.113-118.

[14] Calmet J., Maret P. (2013). Toward a trust model for knowledge-based communities. WIMS 2013: 47.

[15] Vercouter L., Maret P. (2012). Introducing Web Intelligence for communities. Web Intelligence and Agent Systems 10(1): 91-92 (2012).

[16] Stan J., Do V-H., Maret P. (2011). Semantic User Interaction Profiles for Better People Recommendation. ASONAM 2011: 434-437.

[17] Mseddi R., Sidhom S., Ghenima M., Ben Ghezala H. (2011). From information to decision: information management methodology in decisional process. in in Proceedings SIIE'2011 : Information Systems and Economic Intelligence (SIIE'2011) vol.1 (2011) pp.219-226, IGA Edition. Marrakech (Morocco) Feb. 2011.

[18] Pinon J-M. (1996). Projet SEMUSDI : Serveur de documents Multimédia en Sciences de l'Ingénieur. Rapport de Présentation Technique : insa de Lyon, Juillet 1996, 15p.

[19] B. Bertin, V. Scuturici, J.M. Pinon, E. Risler. (2012). CarbonDB : a Semantic Life Cycle Inventory Database. in Conference on Information and Knowledge Management (CIKM) 2012, Maui, Hawaï.2012.

[20] Sidhom S., Hassoun M., Bouché R. (1999). Cognitive grammar for indexing and writing. ISKO-España Conference Proceedings, 22-24 april 1999 Granada, p.11-16.

[21] Sidhom S. (2002). Plateforme d'analyse morpho-syntaxique pour l'indexation automatique et la recherche d'information: de l'écrit vers la gestion des connaissances. Thèse de doctorat de l'Université Claude Bernard Lyon1. France. Mars 2002. p.247.

[22] Sidhom S., Ghenima M., Lambert P. (2010). Systèmes d'information et Intelligence économique : enjeux et perspectives. in Proceedings IEMA-4, 4ème Colloque International sur l'Intelligence Économique et le Knowledge Management - vol.1 (17/05/2010) Alger. (Sidhom S. comme conférencier invité).

[23] Sidhom S., and Lambert P. (2011). "Information Design for Weak Signal detection and processing in Economic Intelligence: case study on Health resources". in Proceedings SIIE'11: Information Systems and Economic Intelligence, IGA Edition. Marrakech (Morocco) Feb. 2011.

[24] Sidhom S. (2013). Conjoncture des processus d'indexation et de gestion des connaissances : vers la réindexation par les usages. in Didactiques et métiers de l'humain et de la relation : nouveaux espaces et dispositifs en question. (direction de Frisch M.), ID Collection L'Harmattan. pp.85-125. Paris, 2013.

[25] Donabédian A., Khaskarian V., Silberztein M. (2013). NooJ Computational Devices. In *Formalising Natural Languages with NooJ*. Eds. Cambridge Scholars Publishing: Cambridge.

[26] Silberztein M., Anaïd Donabédian. (2013). *Formalising Natural Languages with NooJ: selected papers from the NooJ 2012 International Conference*. Cambridge Scholars Publishing: Cambridge.