Analyse automatique d'un *discours constituant* : enjeux méthodologiques autour de la notion de « personne »

Article inédit. Mis en ligne le 23 janvier 2009.

M. Caterina Manes Gallo et Olivier Laügt

M. Caterina Manes Gallo est professeur en sciences de l'information et de la communication à l'université de Bordeaux. Elle travaille en particulier sur la représentation des connaissances à fort ancrage langagier pour des systèmes de traitement automatique de la langue appliqués à la recherche d'information en texte plein.

Olivier Laügt est agrégé de mathématiques et maître de conférences en sciences de l'information et de la communication à l'université de Bordeaux. Il travaille sur le traitement automatique du discours, principalement appliqué aux discours de diffusion et de médiation des sciences.

Plan

PREAMBULE

Le travail que nous présentons prend son origine d'un questionnement théorique sur le fonctionnement des discours « constituants », et inscrit ce questionnement dans une approche interdisciplinaire d'analyse automatique du discours. Zones de paroles parmi d'autres et *paroles* qui se prétendent en surplomb de toute autre, les discours « constituants », comme par exemple le discours philosophique ou le discours juridique, s'autorisent d'eux-mêmes tout en se posant comme dépendants d'une source légitimante (Maingueneau & Cossutta, 1995 : 113-119). Notre objectif est de repérer des phénomènes langagiers pertinents qui contribuent à actualiser la fonction communicationnelle, à la fois *auto* et *hétéro constituante* d'un texte de loi. Notamment le projet de constitution européenne proposé au référendum en France en 2005.

S'agissant d'un gros corpus (150.860 mots, 968.579 signes) le recours à un logiciel d'analyse automatique semble indispensable pour tester le degré de généralisabilité des phénomènes langagiers qui, d'après nos hypothèses théoriques, contribuent à réaliser cette connotation à la fois *auto* et *hétéro constituante* du texte de loi à l'étude. Cependant, comme dans toute démarche empirique, l'utilisation d'une procédure de validation « instrumentée » instaure aussi des problèmes méthodologiques inédits, notamment dans l'interprétation des résultats livrés par le logiciel. Par exemple, les critères de lemmatisation peuvent contribuer à une identification et/ou une catégorisation erronées de certains termes, de la part du logiciel. La nécessité de préparer le corpus afin de pallier les limites

de l'analyseur semble donc s'imposer. D'après notre démarche, l'analyse du discours devient une pratique d'expérimentation, qui permet à terme de faire évoluer la fidélité de la reconstruction automatique du sens, à condition d'accepter un certain « libéralisme » méthodologique. L'utilisation opportuniste des ressources logicielles vise à améliorer la pertinence des résultats du traitement par rapport au « feuilleté textuel » du sens qui peut être attribué à tout discours (Bronckart, 1996).

Dans le présent travail, nous avons focalisé l'attention sur les différentes catégories de déterminants associés au terme *personne* (e.g. *une, toutes les, toute…*). Nous visons à rendre plus homogène l'ancrage langagier des univers sémantiques construits par le logiciel, en nous appuyant sur une approche énonciative. Pour cela, nous avons introduit dans le texte brut des marques qui permettent de distinguer les différentes formes d'occurrences du terme *personne*, en tant qu'objet de discours. Dans ce qui suit nous présentons les analyses du texte de la constitution européenne effectuées à l'aide du logiciel Alceste (version 4.8), en décrivant à la fois les hypothèses sur les objets de discours actualisés, par les différentes catégories de déterminants, ainsi que les résultats obtenus. D'après l'hypothèse générale sous-jacente à notre démarche, les types d'environnements co-textuels associés au même descripteur nominal contraignent la nature de l'objet de discours évoqué au niveau pragmatique (Berrendonner & Rouault, 1991 ; Rouault & Manes Gallo, 2003). Par exemple, l'association au terme « personne » d'un déterminant défini ou indéfini permet de pointer respectivement soit sur son *type* conceptuel (e.g. *le droit de LA personne*), soit sur un ressortissant quelconque de ce même *type* (e.g. *l'évasion d'UNE personne*).

Cette focalisation sur l'ancrage langagier du discours juridique fait abstraction de toute ambition de définir une ontologie du discours juridique (Bourcier, Dulong de Rosnay & Legrand, 2006), à partir des composants lexicographiques ou des primitives de sens du terme à l'étude. Nous avons en revanche fait le pari que la construction par le logiciel de différents univers sémantiques pouvait s'appuyer aussi sur des indices relatifs aux stratégies énonciatives qui accompagnent les occurrences du terme *personne*. L'objectif n'est pas d'intervenir sur les principes statistiques et sémantiques qui guident le traitement du logiciel Alceste mais d'en améliorer la pertinence des résultats par rapport au « feuilleté textuel du sens » caractéristique de tout genre de discours. Le pré traitement manuel vise à identifier la nature de l'objet de discours évoqué par les opérations énonciatives portées par les déterminants associés à un mot plein (i.e. le terme *personne*). Ce pré traitement a pour objectif d'aboutir à une analyse automatique qui permette d'identifier si il existe ou pas une corrélation entre les univers sémantiques d'appartenance de ce mot et les différents objets de discours qu'il peut évoquer (e.g. type conceptuel ou ressortissant d'un type), selon les opérations d'énonciations portées par les déterminants qui lui sont associés.

LE LOGICIEL ALCESTE

Le logiciel Alceste que nous utilisons reçoit en entrée un corpus textuel brut. Après lemmatisation du texte et repérage des lemmes qui pourront contribuer à l'analyse (mots pleins suffisamment fréquents, 6 occurrences minimum dans les analyses que nous allons évoquer), il découpe le texte en UCE (unités de contexte élémentaires), qui correspondent à peu près à des phrases, sous réserve de contenir un nombre raisonnable de formes analysables en vue du traitement statistique (en moyenne 17 environ ici).

L'opération essentielle est alors une classification descendante hiérarchique (CDH) qui répartit les UCE en classes. Il s'agit d'une procédure statistique aveugle qui optimise la répartition des UCE en fonction des cooccurrences des formes qu'elles contiennent (Reinert, 2004). De ce fait, chaque classe est caractérisée par un contexte lexical, liste de

formes qui sont surreprésentées dans les UCE de la classe. Cette surreprésentation est mesurée à l'aide d'un khi2. Les formes d'un même contexte dessinent donc un univers sémantique, « *la présence simultanée de plusieurs mots pleins* [constituant par hypothèse] *une trace archaïque de l'acte d'énonciation* » (Reinert, 2004, p.84). Cette étape laisse apparaître des UCE non classées, pour la raison que leur vocabulaire est trop dispersé entre les différents contextes.

Le travail interprétatif peut alors se faire sur le compte-rendu d'analyse, qui fournit la répartition des UCE sous forme d'une arborescence (les classes sont désignées par des numéros arbitraires), les contextes lexicaux associés, ainsi que la classification ascendante hiérarchique (CAH) de chaque contexte, permettant de visualiser les proximités lexicales effectives dans le texte initial.

LES MARQUES INTRODUITES DANS LE TEXTE BRUT : QUELS OBJECTIFS ?

La fréquence d'occurrence du terme « personne » dans le texte de la Constitution européenne et l'importance que ce terme possède dans le langage juridique[1], nous ont induits à focaliser l'attention sur les objets discursifs, associés à ce terme. D'après une perspective de sémantique énonciative (Culioli, 1999 : 97-100), nous considérons que l'identité de ces objets est déterminée par les « opérations de repérage » portées par le co-texte qui entoure les différentes occurrences du terme « personne » (e.g. déterminants, construction de relatives, anaphorisation, relation de deixis,…).

Nous avons donc établi une catégorisation des déterminants (e.g. singulier, pluriel, indéfini totalisant, …) qui introduisent le terme « personne ». Cette catégorisation répond à un double objectif. D'un côté, il s'agit de mettre en évidence comment l'occurrence de ces déterminants contraint le repérage d'objets de discours distincts. Dans ce cadre, la catégorisation fonctionne comme un dispositif théorique qui nous donne la possibilité d'introduire dans le traitement d'Alceste des indices qui permettent une analyse plus fine des co-occurrences du terme « personne ». D'un autre côté, il s'agit aussi de tester quel est l'effet de l'introduction dans le texte brut de marques distinctives pour chaque catégorie de déterminants sur l'analyse du contenu effectuée par le logiciel.

Notre hypothèse

Selon la perspective interdisciplinaire que nous avons adoptée, le panachage entre analyse automatique et analyse linguistique vise à infléchir la conception du sens imposée par les présupposés sémantiques sous-jacents à la conception du logiciel. Ce qui permet d'intégrer à l'analyse du discours une pratique d'expérimentation. Dans ce cadre, la formulation

. . . . . .

[1] Dans le droit romain, le terme « persona » indique un individu vivant et est utilisé en opposition à « res », désignant un bien matériel, une chose. Cette opposition capitale dans la culture juridique occidentale est reprise dans le Code civil, dont le premier volume traite Des personnes (jouissance des droits civils, état civil, filiation, majorité), tandis que le second règle tout ce qui concerne Des biens et des différentes modifications de la propriété (Dictionnaire Encyclopédique Larousse, 2003). Néanmoins, la définition juridique de ce qu'est une personne demeure épineuse à cause de la difficulté de déterminer à partir de quand on considère qu'un individu est vivant ; in utero ou à la naissance ? Cet aspect, qui concerne surtout la législation sur l'avortement ou sur les droits de l'enfant dans le cas de la fécondation in vitro, n'est pas pris en compte dans la Constitution européenne. Le terme de « personne » renvoie soit à un individu vivant adulte (personne physique, personne à charge, droits d'être de la personne) qui a la capacité d'être titulaire de droits et de se soumettre à des devoirs, soit à un groupement d'individus auxquels est reconnue une personnalité distincte de celle de ces membres et qui est sujet de droit ou qui jouit des mêmes prérogatives (personne civile, personne morale publique).

d'hypothèses, susceptibles d'être confrontées aux résultats fournis par les différents traitements de la part du logiciel, apparaît essentielle. Notre hypothèse sera :

> *L'introduction dans le texte brut de marques distinctives pour les déterminants associés au terme « personne » a un effet sur le découpage des univers sémantiques (mobilisés dans le texte à l'étude) construits par le logiciel Alceste.*

Ce qui renvoie à l'idée que la mise en scène d'objets de discours différents va de pair avec la mobilisation d'univers sémantiques différents. Notamment les occurrences des formes nominales *toute personne* et *des personnes* sont massivement associés à des lexiques différents (verbe, substantifs, etc…).

Une analyse d'essai : analyse 0

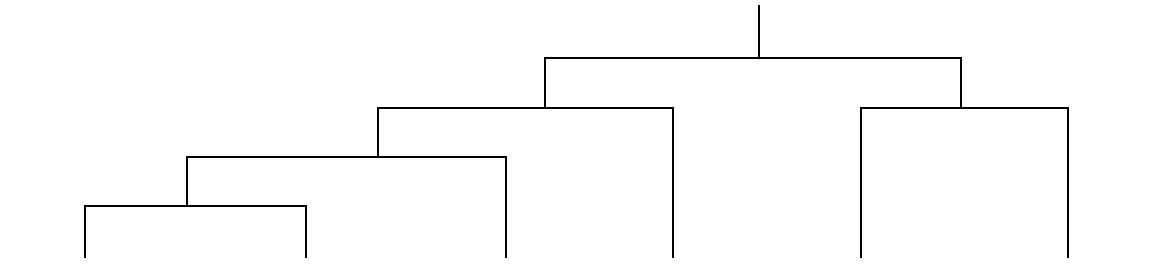
Nous présentons au logiciel le texte du projet de Constitution Européenne, sans préparation particulière. Il comporte 150 860 mots, soit 967 996 signes.

L'analyse découpe 4003 UCE, dont 2947 classées (74%) en 5 classes. La forme personne+ apparaît dans 124 UCE, dont 68 sont classées (55%). Cette forme est donc sous-classée de manière extrêmement significative (khi2 = 23,24, proba = 0,00014%). Un examen rapide du compte-rendu révèle qu'il y a eu, justement pour cette forme, un problème de lemmatisation entre nom et pronom indéfini : toutes les occurrences « personnes » du corpus ont été lemmatisés en nom personne+, alors que les occurrences « personne » l'ont été en pronom indéfini personne. Le nom est surreprésenté dans deux classes, et le pronom dans l'une seulement des deux. Cela nous permet de pointer les limites du lemmatiseur, qui entachent donc les résultats sur un point crucial pour notre interprétation.

Analyse 1

Nous préparons donc le corpus en remplaçant les 112 occurrences de « personne » et les 124 occurrences de « personnes » par la forme *personne_nc*, après avoir vérifié que le pronom « personne » n'apparaissait jamais dans le texte. Ce nouveau corpus est soumis à l'analyse, qui découpe 4002 UCE, dont 3203 classées (80%) en 6 classes.

L'arborescence se présente comme ci-dessous, compte tenu des dénominations que nous faisons en raison des contextes lexicaux (annexe 1) :



définition formelle des institutions	modalités de fonctionnement des institutions	régulation juridique	réglementation financière	droits de l'homme	harmonisation économique
490 UCE	415 UCE	343 UCE	350 UCE	493 UCE	1112 UCE

La forme personne_nc

La forme personne_nc apparaît dans 236 UCE, dont 177 sont classées (75%). La forme est donc encore sous-classée, mais maintenant de manière juste assez significative (khi2 = 3,98, proba = 4,6%). Elle est surreprésentée (et classée) en classe *droits de l'homme* :

	définition formelle des institutions	*modalités de fonctionnement des institutions*	*régulation juridique*	*réglementation financière*	*droits de l'homme*	*harmonisation économique*	total
UCE avec la forme personne_nc	1	5	18	6	104	43	177
UCE dans la classe	490	415	343	350	493	1112	3203
coefficient d'appartenance[2]	-31,4	-17,1	-0,05	-10,9	270,5	-9,0	

Nous pouvons donc considérer que nous avons ainsi pallié la faiblesse du lemmatiseur. Cette analyse nous permet de travailler convenablement sur la référence du projet de constitution à la notion de personne.

Par ailleurs, nous trouvons là validation d'une intuition, préalable à cette recherche, quant à la mise en scène de plusieurs acteurs : *personne*, *citoyen*. Cette intuition est confirmée par la proximité physique de leurs occurrences dans le corpus, mise en évidence par la CAH du contexte (annexe 2).

Un examen manuel des 177 UCE contenant les différentes occurrences de la forme *personne* donne l'idée que le texte utilise ce mot selon différentes acceptions, avec par exemple :

L'acception P1 qui désignerait des personnes physiques ou morales résidant dans un pays membre de l'Union, ou ayant la nationalité d'un Etat membre :

u.c.e. : 254 Classe : 2 Khi2 : 24

3. tout citoyen de l' union ou toute personne_nc physique ou morale residant ou ayant son siege statutaire dans un etat membre dispose, dans les conditions prevues par la partie III, d' un droit d' acces aux documents des institutions

L'acception P2 qui désignerait la personne humaine, de manière générique, jouissant de droits fondamentaux, sans référence à une notion de citoyenneté ou de résidence sur le territoire de l'Union :

u.c.e. : 310 Classe : 2 Khi2 : 107

en-consequence, l' union reconnait les droits, les libertes et les principes enonces ci apres. dignite dignite humaine la dignite humaine est inviolable. elle doit etre respectee et protegee. droit a la vie 1. toute personne_nc a droit a la vie. 2. nul ne peut-etre condamne a la peine de mort, ni execute

L'acception P3 qui désignerait des classes de personnes humaines (âgées, handicapées, déplacées…) sans précision de résidence ou de nationalité :

u.c.e. : 983 Classe : 6 Khi2 : 9

. . . . . .

[2] Nous appelons coefficient d'appartenance le khi2 à 1 ddl, signé. Par exemple, la forme apparaît dans 177 des uce classées, dont 104 sont classées en "droits de l'homme", soit 59%. Cette classe comporte 493 uce parmi les 3203 qui sont classées, soit 15%. La forme est donc surreprésentée, avec un khi2 de 270,5. Une sous-représentation est signalée en signant négativement le khi2.

un statut uniforme de protection subsidiaire pour les ressortissants des pays tiers qui, sans obtenir l' asile europeen, ont besoin d' une protection internationale; un systeme commun visant, en-cas d' afflux massif, une protection temporaire des personne_nc deplacees; des procedures communes pour l' octroi et le retrait du statut uniforme d' asile ou de protection subsidiaire

Ce sont ces considérations qui nous ont amenés à approfondir les différentes stratégies énonciatives qui contribuent à la réalisation de ces différentes acceptions, notamment à partir de l'analyse des déterminants qui introduisent le terme *personne*. Marquer et distinguer ces déterminants pose un problème spécifique, lié au fonctionnement du logiciel. En effet, les déterminants sont d'habitude considérés comme une catégorie grammaticale supplémentaire non significative, n'intervenant pas dans l'analyse. Nous aurions pu forcer le logiciel à les prendre en compte, mais cela n'aurait pas permis de focaliser l'attention sur la forme *personne*, les déterminants étant trop nombreux : l'article *la* possède 6725 occurrences, *des* en a 4214, *les* en a 3978, *une* en a 1215, *toute* en a 244 et *toutes* en a 99. Nous avons donc opté pour un travail de préparation manuelle du corpus, à partir d'une réflexion sur les objets discursifs actualisés par les stratégies discursives qui introduisent le terme « personne ».

LES OBJETS DISCURSIFS ACTUALISES PAR LES OCCURRENCES DU TERME « PERSONNE »

D'après l'approche adoptée, le contenu d'un discours dépend en partie de l'activité re/présentatrice de son énonciation. Les productions verbales renvoient toujours à la création d'un espace dans lequel on établit un réseau de valeurs référentielles, ou un système de repérage des objets de discours par rapport à une situation d'énonciation. Cette situation est définie par différentes coordonnées : le sujet énonciateur, le temps et/ou l'aspect de l'énonciation, la détermination… (Culioli, 1999 : 43-52).

Dans ce cadre on peut considérer que la valeur référentielle du terme « personne » dépend en partie des propriétés linguistiques des marqueurs qui introduisent dans le discours *le complexe faisceau structuré de propriétés physico-culturelles,* auquel renvoie ce terme (Culioli, 1990 : 69-70). En particulier, les propriétés distinctives de chaque déterminant (i.e. le nombre et le degré de définitude) contraignent le mode référentiel par lequel le terme « personne » renvoie à sa *notion*.

Des objets intensionnels et des objets extensionnels

Les formes nominales « det + personne » actualisées dans le discours peuvent poser le signifié de ce terme, soit comme présupposé « déjà existant » dans le discours et/ou « déjà connu » par les interlocuteurs, au moment de son énonciation, soit comme « non existant » et/ou « nouveau » par rapport aux énonciations précédentes (Berrendonner & Rouault, 1991). Lorsque l'article défini n'est pas utilisé pour reprendre anaphoriquement un terme précédemment introduit, il *exhibe* simultanément à son énonciation l'existence du référent conceptuel du terme auquel il est associé. Il contribue à construire une expression sémantiquement *close* qui évoque implicitement la totalité des propriétés de la *notion* mobilisée par ce même terme (Van De Velde, 1994 : 11-15). En revanche, l'article indéfini contribue à construire une expression sémantiquement *ouverte*. Les propriétés du référent conceptuel évoqué sont sélectionnées et/ou attribuées, selon les prédications verbales et/ou les qualificatifs qui lui sont associés par le discours.

Le degré de définitude du déterminant contribue donc à fixer la valeur référentielle du complémentaire discursif de la *notion* évoquée par un terme. Cette valeur référentielle correspond à la nature de l'objet de discours actualisé (Rouault & Manes Gallo, 2003 :

chap. 3). Ainsi le terme « personne » pourra correspondre à un des deux objets de discours suivants :

un *objet intensionnel* ou un *type*, qui selon le nombre du déterminant sera un objet individuel ou une classe. Les déterminants privilégiés sont LE/LA/LES (e.g. […] *elle place la personne au cœur de son action,* […] *ou d'instituer un niveau de protection plus élevé pour les personnes*). LA et LES introduisent directement le *type personne* dans le discours sans en sélectionner un aspect particulier (e.g. *personne physique, personne morale, personne en formation, personne sous-tutelle*) ;

un *objet extensionnel* (i.e. un objet individuel ou une classe), correspondant au ressortissant d'un *type*, auquel le discours attribue des propriétés inédites et/ou contingentes. Les déterminants privilégiés sont UN/UNE/DES (e.g. *favoriser la mobilité des formateurs et des personnes en formation*). DES introduit une classe d'exemplaires indéfinis du *type personne* à laquelle le discours attribue la propriété *d'être en formation.*.

Cette catégorisation des objets de discours a fait l'objet d'une formalisation orientée objet finalisée à la conception d'un système de représentation, pour l'extraction et l'exploitation des connaissances dans le contexte du traitement automatique de la langue (Fredj, 1993). A notre connaissance elle n'a été ni appliquée ni intégrée à un logiciel d'analyse du contenu d'un discours, car elle nécessiterait de compléter le lemmatiseur par un système expert. Nous optons donc pour une intervention manuelle sur le texte brut puisque le logiciel Alceste dans son état actuel ne peut pas prendre en compte les variations des opérations d'énonciation induites par les déterminants qui introduisent des mots pleins.

Le problème posé par « toute personne »

Dans le texte de la Constitution, on relève 80 occurrences du syntagme *toute* + *personne*, majoritairement en position de sujet grammatical et associé à un verbe d'état (i.e. non transformable à la forme progressive). Cette haute fréquence, déjà pointée par l'acception P2 décrite plus haut, oblige à s'interroger sur la nature de l'objet de discours qu'il contribue à construire.

L'indéfinition du support de prédication introduit par le déterminant *tout(e)*, par exemple dans *toute personne a droit à la liberté d'expression* (Art.II-71/1), renvoie à une totalité virtuelle non quantifiable, qui de ce fait peut aussi être vide (Kleiber & Martin, 1977 : 27). L'indéfini *tout(e)*, par un mécanisme distributif différenciateur, permet de généraliser une prédication à tous les membres d'une totalité, mais sans mettre en premier plan leur similitude (Kleibler & Martin, 1977 : 33-35).

Le syntagme indéfini « toute personne » est majoritairement marqué à gauche du verbe (il y a seulement 16 occurrences en position de complément) et ne semble renvoyer ni à un *objet extensionnel* ni à *un objet intensionnel*. En effet, d'un point de vue formel, il n'est pas possible d'opérer les mêmes transformations sur les syntagmes *la personne/une personne* d'un côté et *toute personne* de l'autre. D'où la difficulté d'identifier la catégorie d'objet de discours auquel renvoie *toute personne*.

Lorsque le substantif se trouve en position de sujet, la reprise anaphorique par un pronom qui présuppose le référent connu (e. g. pronom personnel ou démonstratif) est inapplicable. Par exemple, **Toute personne, ça/elle a le droit de travailler* vs. *Une personne, ça/elle a le droit de travailler* (Leeman, 2004 : 103-104). La possibilité d'une reprise anaphorique est

au contraire applicable dans le cas de l'indéfini *des* : e. g. *des personnes résidant dans un pays étranger ça a/elles ont le droit de voter auprès du consulat de leur pays d'origine*. Par conséquent *tout(e)* ne permet le repérage ni d'une classe, ni d'un individu *extensionnels* (Cf. supra).

De plus si l'on admet que l'indéfini *toute* permet de repérer une totalité, la plus générale possible, on est confronté à l'impossibilité d'en extraire les membres constituants, en termes de *types* et/ou de *sous-types*. En d'autres termes, il n'est pas possible d'extraire un *objet intensionnel*. Les questions en *lequel/laquelle* sont exclues (Martin 2006 : 20). Par exemple, *toute bonne nouvelle réjouit le cœur de l'être humain* → *laquelle ? *tout retard sera sanctionné* → *lequel ?, *tout candidat au doctorat doit avoir un diplôme de Master* → *lequel ? En revanche, la même question est possible avec *le* (*il faut que le livre cesse d'être une marchandise comme une autre* → lequel ? → le livre d'art, le livre de collection, le livre numérique), ou avec *les* (*les bonnes nouvelles réjouissent le cœur* → lesquelles ?→ la nouvelle d'une promotion, la nouvelle d'un mariage, la nouvelle d'une naissance), ou encore avec *tous les* (*tous les candidats au doctorat doivent avoir un diplôme de Master* → lesquels ? → les candidats français, les candidats européens, les candidats non européens). L'extraction par *lequel/laquelle* est possible aussi avec des *objets extensionnels*. Par exemple, *passe moi des verres/des crayons* → lequels ? → des verres rincés/des crayons de couleur.

En récapitulant, la question *lequel* est compatible à la fois avec un déterminant défini qui introduit un *objet intensionnel* et avec un déterminant indéfini qui pose dans le discours *un objet extensionnel*, mais est inapplicable aux syntagmes nominaux « tout(e) + substantif ». D'où la nécessité d'établir quelle est la nature de l'objet de discours construit et/ou à quel niveau (intensionnel ou extensionnel) il appartient.

« Toute personne » : un objet intensionnel virtuel

D'après (Jayez & Tovena, 2007 : 5-10), l'expression « tout(e) + substantif » est maladroite lorsque : a) l'énoncé dans lequel il apparaît peut avoir une *lecture rigide*[3], c'est à dire lorsque la propriété attribuée au substantif peut être appliquée aux éléments d'un ensemble identifiable, b) cette application peut faire l'objet d'un jugement de véridicité. Par exemple, **toute chaise qui était dans le jardin a été mouillée par la pluie*, est maladroit parce que l'attribut *être dans le jardin* délimite un ensemble identifiable de chaises qui impose une lecture *rigide* à l'indéfini *toute*, et donc l'attribut peut être considéré comme intrinsèque à la totalité des chaises existantes ou imaginables. « Tout(e) + substantif » est possible surtout dans le cas des énoncés normatifs parce qu'il renvoie à un ensemble mal délimité. Dans *habituellement tout étudiant reçoit un dossier d'inscription en Mai* la véridicité de l'énoncé ne dépend pas de la possibilité de constater que effectivement tous les différents étudiants reçoivent le dossier en Mai. L'énoncé permet d'asserter un fait indépendamment de sa satisfaction. La non nécessité d'identifier ou de délimiter la totalité ciblée par *tout(e)* partage un air de famille avec l'objet virtuel dont parle Robert Martin (2006 : 23).

Comment retraduire en termes d'objets intensionnels et/ou extensionnels ce mode référentiel *non rigide* de « Tout(e) + substantif » ? Une solution serait de considérer que cette forme en position de sujet grammatical permet d'introduire dans le discours un *type* (objet intensionnel) dont l'extension coïncide, non pas avec une classe comme dans le cas de *tous les* , mais avec le *type* lui même et sans possibilité d'en extraire un ressortissant ou un exemplaire particulier (objet extensionnel). Qu'il s'agisse d'un *type* et pas d'un *objet*

. . . . . .

[3] Nous ne reprenons pas le complexe dispositif formel qui vise à circonscrire cette notion. Nous nous limitons à nous inspirer de sa conceptualisation intuitive qui nous semble essentielle pour rendre compte des spécificités de l'indéfini *tout(e)*.

extensionnel dépend de l'impossibilité d'attribuer des propriétés contingentes ou inédites à l'expression nominale introduite par *tout(e)*. D'où l'impossibilité de **toute chaise est en train de se mouiller sous la pluie.*

La lecture non rigide de « tout(e) + substantif » peut être assimilée au fait que le passage de l'intension à l'extension est bloqué. D'où l'impossibilité d'un questionnement par *laquelle : toute personne a le droit de déposer une plainte pour outrage à la dignité* → *laquelle ? Cette solution semble cohérente avec l'occurrence de l'indéfini « nul », symétrique de « toute », et qui sert à exprimer l'interdiction absolue de certains droits, là où *toute personne* permet une généralisation maximale d'autres droits.

Le syntagme *toute personne* en position de sujet grammatical pose au sein du discours l'existence d'une totalité d'éléments non identifiable au niveau extensionnel, mais qui contribue à la dimension « auto constituante » du discours. Il permet de généraliser le prédicat qui lui est attribué à tous les membres virtuels de la totalité, malgré et/ou indépendamment des différences et des similitudes qu'ils entretiennent entre eux. Dans le projet de constitution européenne ce mécanisme discursif contribue à fonder une « harmonisation » de la vie publique au sein de l'Union européenne.

LES CINQ MODALITES D'ACTUALISATION DISCURSIVE DU TERME « PERSONNE »

Les marques introduites dans le texte brut sont relatives aux déterminants associés au terme « personne ». Leur catégorisation, qui s'appuie sur les distinctions précédentes, s'articule sur les trois critères suivants :

le nombre du déterminant qui peut renvoyer soit à un individu (*une/la personne*), soit à une classe (*des/les/aux*). *Aux* et *Des* sont considérés, au niveau morphologique, comme l'amalgame respectivement de : *à + les* et *de + les* ;

le mode référentiel défini et/ou indéfini, sous-jacent au déterminant qui introduit le substantif « personne », contribue à en déterminer l'existence au sein du discours, respectivement comme *objet intensionnel* ou comme *objet extensionnel* ;

la connotation totalisante exhaustive du déterminant *tout(e)* pose l'existence d'un *type virtuel* qui bloque la possibilité d'en extraire un individu ou une classe extensionnels auxquels la progression du discours attribue des propriétés contingentes.

Dans le tableau suivant nous reportons les fréquences brutes des différents déterminants associés à « personne », ainsi que leur catégorie d'appartenance et l'objet de discours qu'ils contribuent à actualiser.

	INDIVIDU	CLASSE
Objet intensionnel *virtuel*	Toute personne (80) Toute personne physique ou morale (4) Toute autre personne (1)	Toutes les (6) Toutes autres (1)
Objet intensionnel *type*	La personne (16)	Aux (24) Les (autres) (15), Ces (1)
Objet extensionnel	Une (telle) personne (11)	Des (67), D'autres (3), *entre* personnes (2)…

Comme on peut le constater il y a une nette différence de distribution entre les déterminants qui introduisent respectivement une classe et/ou un individu. Notamment, la plus grande richesse des marqueurs de classe par rapport à ceux d'individu et la présence massive de *toute* au singulier. Ce phénomène, comme on le verra par la suite semble conforter notre hypothèse.

Nous avons négligé de distinguer les cas où DES contribuait à construire un syntagme complexe (44 sur 67) (e.g. *la protection des personnes, la sélection des personnes,* …) des cas d'occurrence simple (e.g. *les dispositions de l'article 196 du dit traité s'appliquent à des personnes …, vérifier si des citoyens d'états membres ou des personnes à leur charge* …).

Nous avons alors modifié manuellement le corpus, en remplaçant la forme *personne_nc* par des marques spécifiques, permettant de la distinguer en fonction du déterminant qui l'introduit (par exemple, *des personne_nc* devient *des ppp_indp*).

Table des correspondances

détermination	nouvelle forme	nombre de remplacements
des d'autres entre	ppp_indp	72
une telle	ppp_inds	11
la	ppp_defs	16
les de aux (quelques cas d'énumération) ces	ppp_defp	45
toute toute autre	ppp_tot	85
toutes les toutes autres	ppp_tot_p	7

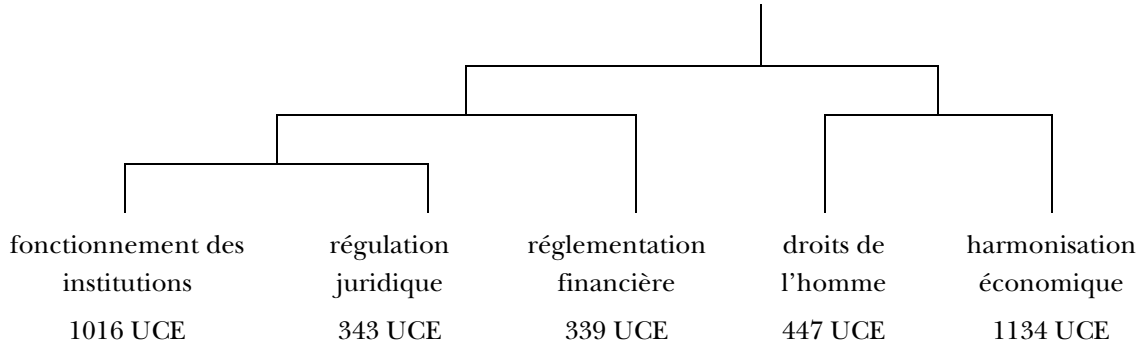
Cette modification du corpus touche une forme analysée, toutes choses égales par ailleurs. Elle va donc avoir mécaniquement une répercussion sur les résultats du traitement, répercussion liée à la prise en compte plus précise des univers sémantiques mobilisés autour des différentes acceptions de la forme *personne*. Nous nous attendons donc à obtenir une arborescence et un classement différents de celui de la première analyse.

Analyse 2

L'analyse découpe alors 4000 UCE, dont 3279 classées (82%) en 5 classes.

Il y a deux UCE de moins car le découpage a regroupé les trois UCE 380 à 382 de l'analyse1 en les UCE 380 et 381, ainsi que les trois UCE 3527 à 3529 en les UCE 3526 et 3527. Ces variations ont pour seule explication que la forme analysée *personne_nc* a été remplacée par six formes : il y a dans l'analyse2 1380 mots analysés, au lieu de 1375 dans l'analyse1.

L'arborescence présente ainsi seulement 5 classes. L'examen des contextes lexicaux montre que les classes que nous avons dénommées *définition formelle des institutions* et *modalités de fonctionnement des institutions* dans l'analyse1 se résolvent en une seule que nous nommons ici *fonctionnement des institutions*, les autres contextes restant suffisamment stables pour que nous ne modifiions pas les dénominations (annexe 3). Il en résulte que l'arborescence se présente comme ci-dessous :



Deux phénomènes se superposent : il y a davantage d'UCE classées que dans l'analyse1 (82% vs 80%), et certaines UCE ont changé de classe (13), ou ont été déclassées (97), du fait de la légère recomposition des contextes lexicaux (annexe 4). Il apparaît ainsi que parmi les 799 UCE non classées de l'analyse1, 175 ont réussi à être classées dans l'analyse2, principalement dans le contexte *fonctionnement des institutions*. Une compréhension fine de ces mutations nécessiterait une plongée dans l'algorithme de traitement statistique du logiciel. Ce travail n'entre pas dans notre objectif initial de recherche de corrélation entre les univers sémantiques du terme *personne* et le type de déterminant qui lui est associé. Il impliquerait en outre un investissement du concepteur du logiciel.

L'examen du classement des formes issues de la recodification de *personne_nc* (annexe 5) fait apparaître que si elle était précédemment classée dans le contexte *droits de l'homme*, les formes *ppp_tot*, *ppp_defs*, *ppp_inds* et *ppp_tot_p*, qui correspondent aux syntagmes *toute personne*, *la personne*, *une personne* ou *toutes les personnes* restent dans le contexte *droits de l'homme* (annexe 6). En d'autres termes, dans cet univers sémantique, la notion de personne est actualisée surtout comme un objet intensionnel et comme un objet intensionnel virtuel. En revanche, les formes *ppp_defp* et *ppp_indp*, correspondant à *les personnes* ou *des personnes*, sont maintenant dans le contexte *harmonisation économique*.. Dans cet univers, la notion de personne est actualisée surtout comme classe intensionnelle et/ou extensionnelle.

La différenciation par leurs déterminants des occurrences de la forme personne a donc permis un regard plus fin, et confirme la distinction entre individu et classe. De plus, cela conforte l'intérêt de l'approfondissement de la distinction ébauchée plus haut entre les acceptions P1, P2 et P3.

En revanche, le problème laissé ouvert par ce travail est qu'il ne met pas à jour de corrélation entre *personne* comme objet intensionnel virtuel et son appartenance à un univers sémantique spécifique différent de celui construit par l'objet intensionnel *la personne*.

OUVERTURE SUR D'AUTRES ANALYSES

Les résultats des deux analyses précédentes font en outre émerger des questions plus profondes à propos de la classe *droits de l'homme*.

Tout d'abord, la surreprésentation de la forme *droit* dans ce contexte lexical est écrasante. Il apparaît de plus que cette forme est facilement cooccurrente avec la forme *personne_nc*. Cela nous a amenés à effectuer deux autres analyses pour étudier quels déterminants de personne concourent à cette cooccurrence, et élucider les circonstances (forme simple ou syntagme complexe) qui amènent la forme *droit* à cooccurrer avec *personne*, comme objet intensionnel ou comme objet intensionnel virtuel. Ces analyses, que nous n'avons pas la place d'exposer ici, constituent donc des pistes alternatives pour mettre en évidence cette distinction aussi au niveau du traitement automatique du discours.

Références bibliographiques

Berrendonner, Alain ; Rouault, Jacques (1991), *Sémantique des objets et calcul des noms*, KMET'91.

Bourcier, Danièle ; Dulong de Rosnay, Mélanie ; Legrand, Jacky (2006), « Susciter la construction interdisciplinaire d'ontologies juridiques : bilan d'une expérience », *Société des connaissances*, Nantes 2006, 1-10.

Bronckart, Jean-Paul (1996), *Activité langagière, texte et discours*, Lausanne : Delachaux-Niestlé.

Culioli, Antoine (1990 et 1999), *Pour une linguistique de l énonciation*, Tomes 1 et 2, Paris : Ophrys.

Dictionnaire Encyclopédique Larousse (2003), Paris.

Fredj, Mounia (1993), *SAPHIR : un système de représentation des connaissances pour la langue naturelle*, Thèse, Université de Sciences Sociales, Grenoble 2.

Jayez, Jacques ; Tovena, Lucia (2007), « Déterminants et irrévérence. L'exemple de *tout.* » In : Reichler-Béguelin, Marie José (éd.), *Référence nominale et temporelle*, Berne : Peter Lang,, 1-24.

Joly, Martine ; Versel, Martine ; Laügt, Olivier (1998), « Alceste et *Le Monde*, Image(s) Virtuel(les) », *in* Beau F., Dubois P. et Leblanc G. (dir.), *Cinéma et dernières technologies*, Bruxelles : De Boeck - INA, 129-152.

Kleibler, Georges ; Martin, Robert (1977), « La quantification universelle en français :*Le, un, tout, chaque, n'importe quel* », *Semantikos*, Vol.2, Fasc.1, 19-36.

Laügt, Olivier (dir.) (2004), « Du traitement du discours dans les recherches en communication », *Cahiers de Jérico-st* n°4, Tours : Presses de l'Université François Rabelais de Tours.

Leeman, Danielle (2004), *Les déterminants du nom en français : syntaxe et sémantique*, Paris : PUF - Linguistique nouvelle.

Maingueneau, Dominique ; Cossutta, Frédéric (1995), « L'analyse des discours constituants », *Langages*, n° 117, 112-125.

Manes Gallo, M. Caterina (2003), « Communication Humain/Machine en langue naturelle: Un nouvel enjeu pour la psycholinguistique », *Revue Interaction Homme/Machine*, vol. 4, N° 2, 67-87.

Martin, Robert (2006), « Définir l'indéfinition », In : Corblin, Francis ; Ferrando, Sylvie ; Kupferman, Lucien (dir.), *Indéfini et prédication*, Paris : Presses de l'Université Paris Sorbonne, 11-24.

Reinert, Max (2004), « La méthode d'analyse exploratoire des données textuelles "Alceste" et le problème de l'analyse de contenu », in (Laügt, 2004).

Rouault, Jacques ; Manes Gallo, M. Caterina (2003), *Intelligence Linguistique : Le couple sémantique -pragmatique et le calcul du sens des énoncés élémentaires*, Paris : Hermès-Science.

Van De Velde, Danièle (1994), « Le défini et l'indéfini », *Le Français Moderne*, LXII, N°1, 11-35.