
Blogoscopie

Manuel utilisateur

Auteurs :

- Helena Blancafort, Syllabs
- Estelle Dubreil, Lina
- Marguerite Leenhardt, Syllabs
- Laura Monceaux, Lina

Coordination :

- Estelle Dubreil, Lina

Objet du manuel utilisateur :

Le manuel utilisateur revêt un caractère descriptif ; il a pour objectif d'une part, de présenter le corpus Blogoscopie, sa nature et sa composition, et d'autre part, de définir et d'exemplifier chacune des balises présentes dans la DTD. En cela, il se différencie du manuel d'annotation du corpus Blogoscopie, centré sur la justification du choix des blogs, billets et commentaires à annoter et sur les règles de sélections qui sont intervenues dans le processus d'étiquetage.

Sommaire :

- Le corpus Blogoscopie
- Les balises, les attributs et les valeurs
- Annexes

Tables des matières

1.	Le corpus Blogoscopie	3
1.1	Nature du corpus	3
1.2	Composition du corpus	4
2.	Les balises, les attributs et les valeurs	5
2.1	Informations générales	5
2.2	Les concepts	7
2.3	Les instances.....	9
2.4	Les évaluations.....	9
2.5	Synthèse chiffrée	15
3.	Annexes.....	15
3.1	La DTD.....	15
3.2	Exemple de fichier annoté : Harry Potter	18

1. Le corpus Blogoscopie

1.1 Nature du corpus

Le corpus Blogoscopie est un corpus de blogs, dont les billets et commentaires ont été annotés manuellement via l'outil Oxygen (logiciel d'annotation xml). L'annotation porte sur deux catégories d'éléments : les concepts contenus dans les billets et les commentaires, et les évaluations émises par l'auteur du billet ou du commentaire à propos des concepts exprimés.

Un effort de fixation terminologique est nécessaire :

Blog : « un type de site web composé essentiellement de billets (ou d'actualité) publiés au fil de l'eau et apparaissant selon un ordre antéchronologique (les plus récents en haut de page), le plus souvent enrichis de liens hypertextes externes » [Fievet et Turrettini 2004, p. 4]¹.

Billet : texte plus ou moins court, photo, liens, etc. publiés sur un blog.

Commentaire : réaction laissée par un internaute ou l'auteur lui-même à la suite de la publication d'un billet.

Concept : nom ou groupe nominal caractéristique d'un sujet ou d'un thème.

Sujet : mot-clé ayant servi à l'extraction des billets et correspondant à la cible applicative.

Thème : catégorie proposée pour la répartition des blogs sur la plateforme Overblog (Ex : actualité, artiste, économie, etc.).

Évaluation : ensemble des marques linguistiques indiquant le sentiment de l'auteur par rapport à un concept (Ex : opinion, appréciation, jugement, etc.).

Préalablement à la collecte des données, une analyse comparative de trois typologies des blogs nous a permis de centrer nos analyses sur les blogs thématiques personnels, c'est-à-dire sur « des blogs personnels francophones, tenus par un seul et même individu, régulièrement alimenté et à caractère public » [Cardon et Delaunay-Teterel 2006, p.3]². En effet, contrairement aux blogs « journaux intimes » et aux « weblogs purs », ces blogs présentent la particularité de traiter d'un concept en particulier, les billets afférents contiennent simultanément une partie informative relative au concept discuté, et une partie évaluative relative au point de vue de l'auteur dans les billets et à celui de ses interlocuteurs dans chaque commentaire respectif.

¹[Fievet et Turrettini 2004] C. Fievet et E. Turrettini, *Blog Story*, Eds. Eyrolles, 305 p.

²[Cardon et Delaunay-Teterel 2006] D. Cardon et H. Delaunay-Teterel, La production de soi comme technique relationnelle : un essai de typologie des blogs par leurs publics, *Réseaux* N°138, pp. 15-71.

Les blogs ont été collectés à partir de la plateforme Over-Blog³ en juin 2007. Une extraction antérieure des billets et des commentaires aurait introduit un biais dans la représentation et la variété des concepts, tant l'intérêt des blogueurs était centré sur les élections présidentielles françaises. Dans un souci de représentativité du corpus par rapport aux centres d'intérêt des blogueurs, nous avons sélectionné les 10 blogs les plus visités par thème, puis les 10 billets les plus récents publiés et les 10 premiers commentaires (maximum) associés à chaque billets.

1.2 Composition du corpus

Le corpus comporte 200 billets annotés et 612 commentaires associés, ce qui représente un peu plus de 83 500 mots répartis au sein de 200 fichiers au format .xml. Au total, 5666 formes de concepts ont été annotées et 4943 formes d'évaluations.

Le corpus Blogoscopie comporte trois sous-parties explicitées ci-dessous.

La première partie du corpus comporte 76 billets et 296 commentaires. Cette première partie se décompose elle-même en deux sous-parties : l'une comportant 12 billets et 27 commentaires associés, l'autre comprenant 64 billets et 269 commentaires associés.

Les 12 premiers billets ont été sélectionnés selon un critère thématique en corrélation avec la diversité et la difficulté des marques linguistiques de l'évaluation qu'ils renfermaient⁴. Les noms des 12 billets concernés sont : blog01, blog02, blog03, blog04, blog05, blog06, blog07, blog08, blog09, blog10, BlogPolitique, blogSpiderman.

Les 64 autres ont été sélectionnés selon un critère thématique parmi 33 des 43 catégories proposées sur le site Over-Blog (en éliminant les journaux intimes, les photos blogs, les blogs lesbiens, etc.), lesquels sont : actualité, artiste, blogzine, business, cinéma, collectionneur, consommation, console, croyance, décoration, détente, économie, enfants, familles, gastronomie, guide d'achat, internet, jeux, livres, loisirs, maison, musique, philosophie, politique, religions, rêver, santé, sciences, sport, techno, télévision, voyage, weblogs. Les noms des 64 billets reprennent l'intitulé de la catégorie thématique et entre parenthèses le chiffre correspondant au nombre de fichiers présents dans chacune des catégories (Ex : cinema, cinema (2), cinema (3)), mais aussi au rang d'apparition du fichier dans la liste composant le répertoire⁵.

³Over-Blog est une plateforme de blogs, c'est-à-dire un outil permettant de créer des blogs. Cette plateforme est gérée par la société JFG Network, partenaire industriel applicatif du projet Blogoscopie, chargé de l'extraction des données textuelles.

⁴Cette sélection trouve sa pertinence au regard de la procédure d'accord inter-annotateur nécessaire à la mise en place d'un schéma d'annotation (Cf. Manuel d'annotation du corpus Blogoscopie).

⁵Dans certaines catégories, l'utilisateur pourra s'étonner de ne pas trouver 3 fichiers annotés comme les noms des fichiers le laisseraient penser. Par exemple, la catégorie *actualite* comporte deux fichiers, l'un intitulé *actualite*, l'autre *actualite (3)*. Le fichier *actualite (2)* n'a pas disparu, il n'a jamais existé ! Cette absence s'explique par la nature de l'extraction qui a été fournie par Over-Blog ; elle comportait autant de répertoires que de catégories. Les annotateurs ont alors étiquetés les fichiers au fur et à mesure de leur ordre d'apparition dans la liste du répertoire. Un fichier jugé inintéressant, soit parce qu'il ne comporte pas de donnée textuelle, soit parce que les données textuelles ne comportent pas d'évaluations, n'est pas annoté. Pour autant il conserve le numéro correspondant à son rang d'apparition dans le répertoire principal et ce numéro n'est pas réattribué.

Deux à trois billets (et leurs commentaires associés) ont été annotés par catégorie. A ce stade de l'annotation, le but de l'annotation était de couvrir le maximum de thèmes possible pour une meilleure représentativité des centres d'intérêts des blogueurs. Cependant, la faible récurrence des concepts nous a poussés à appliquer un autre principe de sélection des billets, au risque sinon de ne pouvoir dégager suffisamment de régularités pour l'automatisation future du processus de détection automatique des évaluations.

La deuxième partie du corpus rassemble 34 billets et 162 commentaires, répartis en 8 thèmes; 4 billet par thème. Ces 8 thèmes sont : actualite, artiste, blogzine, business, cinema, gastronomie, guidesdachat et livres. De même que pour la première partie du corpus, le numéro figurant entre parenthèses dans le nom du fichier, correspond au rang d'apparition du fichier dans le répertoire de sa catégorie thématique. Les fichiers ont été sélectionnés selon les deux mêmes critères que précédemment, en fonction de la présence d'un contenu textuel et d'évaluations portant sur un ou des concepts. A ce stade de l'annotation, le but de l'annotation était d'annoter des évaluations sur des concepts les plus récurrents possibles. Cependant, la faible variété des évaluations recensées nous a poussés à appliquer là encore un autre principe de sélection des billets.

La troisième partie du corpus se compose de 90 billets et 156 commentaires, répartis en 10 sujets; 9 à 10 billets par sujet. Ces 10 sujets sont : beaujolais, développement durable, grève SNCF, Harry Potter, le cœur des hommes 2, loi sur la responsabilité des universités – LRU, nucléaire, Raymond Domenech, Vladimir Poutine et Wii. A ce stade de l'annotation, le but de l'étiquetage était d'annoter des évaluations variées sur des concepts les plus récurrents possibles. Cette dernière approche a été la plus fructueuse; nous avons observé une grande récurrence en corrélation avec une grande diversité des formes d'expressions de l'évaluation.

2. Les balises, les attributs et les valeurs

2.1 Informations générales

Tous les fichiers XML sont introduits par deux lignes, l'une précisant que les fichiers sont tous encodés en UTF-8 `<?xml version="1.0" encoding="UTF-8"?>`, l'autre précisant le nom du fichier correspondant à la DTD appelée `<!DOCTYPE page SYSTEM "../pagev4.dtd">`.

Un fichier XML annoté regroupe un billet, étiqueté par la balise `<billet></billet>`, et ses commentaires, étiquetés par la balise `<comment></comment>`. Le billet et les commentaires éventuels sont regroupés sous une seule et même balise appelée `<page></page>`⁶.

Figure ci-dessous le tableau présentant, définissant et exemplifiant les différents attributs associés à chacune de ces balises. Tous les exemples mentionnés sont extraits du corpus d'échantillon livré avec le manuel utilisateur. Ce corpus d'échantillon comporte 5 fichiers sélectionnés dans la troisième partie du corpus, pour leur représentativité par rapport aux balises et aux attributs présentés. Les 5 fichiers sont : beaujolais-7959750-7959966, Dev_durable-7941541-7941723, Greve_SNCF-1001625708-1019858752, Harry_Potter-7935946-7935967 et raymond_domenech-1318939-1019847978.

⁶Le fichier correspondant à dernière DTD utilisée figure en annexe de ce manuel.

Informations	Balise XML	Attributs
Blog	<page>	<i>mes_blog_rank</i> (OPT ⁷) : mesure du blog rank du blog Ex : <i>mes_blog_rank="10"</i> <i>mes_mediametrie</i> (OPT) : mesure médiamétrie du blog <i>tags_blog</i> (OPT) : thématiques entrées par l'auteur pour son blog Ex : <i>tags_blog="humour,littérature,écriture,publicité,pub,antidote"</i> <i>thematique</i> (OBL ⁸) : thématique dans laquelle est répertorié le blog Ex : <i>thematique="Harry Potter"</i> <i>url</i> (OBL) : url du blog à partir de laquelle est extrait le billet Ex : <i>url="http://sarahleslie.over-blog.com/</i>
Billet	<billet >	<i>age, profession</i> (OPT) : âge et profession de l'auteur du billet <i>auteur</i> (OBL) : pseudo de l'auteur du billet Ex : <i>auteur=" DOMINIQUE"</i> <i>id_b</i> (OBL) : identifiant du billet Ex : <i>id_b=" B7959966"</i> <i>url</i> (OBL) : url du billet Ex : <i>url=" http://fresitaroja.over-blog.com/article-4531728.html"</i> orthographe ⁹ (OBL) : orthographe standard ou erronée du billet Ex : <i>orthographe=" erronée "</i> syntaxe (OBL) : syntaxe correcte ou erronée du billet Ex : <i>syntaxe=" erronée "</i>
Commentaire	<comment >	<i>auteur</i> (OBL) : pseudo de l'auteur du commentaire Ex : <i>auteur=" Dominik Vallet"</i> <i>id_co</i> (OBL) : identifiant du commentaire Ex : <i>id_co="CO1"</i> orthographe (OBL) : orthographe standard ou erronée Ex : <i>orthographe="standard"</i> syntaxe (OBL) : syntaxe correcte ou erronée du commentaire Ex : <i>syntaxe="correcte"</i>

Tableau 1 : Définition et exemple des attributs associés aux 3 balises principales

Les attributs en italique sont automatiquement pré remplis lors de la mise en forme des pages extraites par Overblog. En revanche, l'annotateur évalue l'orthographe et la syntaxe comme suit :

- Syntaxe : style et structure de phrase employés ;
- Orthographe : présence faute d'orthographe et d'accord.

Les balises <billet> et <comment> sont composée de plusieurs balises :

- <date> : correspond à la date de publication du billet ou du commentaire
 Ex : <date>2006-05-29 11:48:00</date>
- <titre> : correspond à l'intitulé donné par l'auteur du billet ou du commentaire
 Ex : <titre>373) Sémantique.</titre>

⁷OPT : attribut optionnel

⁸OBL : attribut obligatoire

⁹Lorsqu'un attribut figure en gras, c'est que son assignation reste à la charge de l'annotateur.

- <texte> ou <textco> (pour le texte des commentaires) : correspond aux données textuelles composant le billet ou le commentaire

Ex : <textco><partie organisation="énonciatif">Le gouvernement connaît très bien les raisons de la grève, mais il fait tout pour faire passer les <CA cc="C1, train">cheminots</CA> pour des privilégiés, mais qui s'est augmenté de 172 % ? </partie></textco>

- <partie> : le texte d'un billet ou d'un commentaire est découpé en partie (au moins une), chacune des parties correspondant à mode d'organisation du discours (Ex : <partie organisation="énonciatif">). A chaque changement de type de discours, on annote une nouvelle partie. Il existe quatre modes d'organisation du discours présentés dans le tableau ci-dessous¹⁰.

Valeurs possibles de l'attribut organisation	Descriptif
narratif	Succession des actions d'une histoire
énonciatif	Opinion avec marque de locuteurs : rapport d'influence, point de vue, témoignage
argumentatif	Opinion sans marque de locuteurs (assertion de départ / assertion d'arrivée / une ou plusieurs assertions de passage)
descriptif	Identifier une succession (Nommer / Localiser / Qualifier)

Tableau 2 : Modes d'organisation du discours

Chacun de ces modes d'organisation du discours est représenté par un attribut `organisation=""` dont les quatre valeurs possibles reprennent les noms ci-dessus mentionnés.

2.2 Les concepts

D'une façon générale, un concept est un **groupe nominal, occurrence du texte** du billet ou d'un commentaire. Nous avons définis trois types de concepts pour annoter le corpus.

¹⁰Pour une présentation exhaustive des modes d'organisation du discours, se reporter à P. Charaudeau 1992, *Grammaire du sens et de l'expression*, Eds Hachette Education, 928 p.

Type de concept	Définition	Balise XML	Attributs
concerné	Il répond à la question « De quoi parle le billet ? » et correspond en ce sens à la nature du référent présent uniquement dans le billet.	CC	<p>id_c (OBL) : identifiant du concept, attribué manuellement par l'annotateur au fur et à mesure des occurrences textuelles rencontrées. Ex : <CC id_c="C1">Grèves</CC></p> <p>hyponyme (OPT) : relation sémantique dominante du concept rencontré par rapport à un autre concept déjà annoté, C3 correspondant dans l'exemple ci-dessous à Beaujolais nouveau. Ex : <CC id_c="C2" hyponyme="C3">BEAUJOLAIS</CC></p> <p>niveau (OPT) : niveau de la langue du concept, précisé si différent de courant. Il existe 4 valeurs possibles : familier (Ex : bagnole), soutenu (Ex : automobile), terminologique (Ex : autolocomoteur), vulgaire (Ex : chiotte).</p>
associé	Il est associé aux champs sémantiques des CC dans le billet ou dans l'un des commentaires.	CA	<p>cc (OBL) : identifiant du CC auquel le concept associé est sémantiquement rattaché. Il peut avoir plusieurs CC de rattachement. Ex : <CA cc="C1">grévistes</CA></p> <p>niveau (OPT) : niveau de la langue, à préciser si différent de courant.</p>
non associé	Il est non associé aux champs sémantiques des CC, mais obligatoirement porteur d'une évaluation dans le billet ou l'un des commentaires. Ex : <CN>photos</CN> dans un billet sur le Beaujolais.	CN	<p>niveau (OPT) : niveau de la langue, précisé si différent de courant.</p>

Tableau 3 : Trois catégories de concepts

Sans rentrer dans les détails des règles adoptées pour l'annotation, l'utilisateur doit être averti de certaines conventions :

- Un concept est **annoté dans la page une seule fois**, même s'il apparaît plusieurs fois, et cela pour une meilleure visibilité d'annotation. On pourra retrouver automatiquement les occurrences identiques ;
- **Les variantes** d'un concept déjà annoté sont **de nouveau annotées**, une seule fois par variante, précisément parce qu'une procédure automatique ne pourrait permettre de les retrouver. Ex : <CA cc="C1">mouvements</CA> ; <CA cc="C1">mouvement</CA> ;
- Pour les concepts concernés, les **synonymes** sont regroupés sous le **même identifiant**. Ex : fichier blog01.xml <CC id_c="C1"> série </CC> ; <CC id_c="C1"> sitcom </CC>. Cela facilite l'annotation des évaluations, en particulier leur rattachement aux « bons » concepts ;
- Enfin, pour faciliter la construction de futurs lexiques basés sur les concepts trouvés en corpus, nous avons fait le choix d'annoter tous les CC et tous les CA, y compris ceux ne faisant pas l'objet d'une évaluation.

Une évaluation peut porter sur un nom propre, dans ce cas on le qualifie d'instance. Les instances sont définies ci-dessous.

2.3 Les instances

D'une façon générale, une instance est une **entité nommée, occurrence du texte** du billet ou d'un commentaire. Une Entité Nommée correspond à l'acceptation la plus large que l'on peut faire du nom propre, elle « regroupe les noms propres communément reconnus comme tels (la classe ENAMEX des conférences MUC), ainsi qu'un certain nombre d'entités qui ne sont pas toujours considérées comme noms propres : les noms collectifs (les Français, les néandertaliens, etc.), les maladies ou encore les noms de personnages mythiques ou fictifs (Hercule, Colombo, etc.). » [Fourour et Morin 2003, Revue québécoise de linguistique, <http://www.erudit.org/revue/rql/2003/v32/n1/012243ar.html>].

Nous avons défini deux types d'instances pour annoter le corpus : les instances associées et les instances non associées.

Type d'instance	Définition	Balise XML	Attribut
associée	Elle est associée aux champs sémantiques des CC et porteuse d'une évaluation.	IA	concept (OBL) : forme lexicale du concept concerné ou associé à laquelle l'instance est associée. Ex : <IA cc="film, livre, livres, roman">Harry Potter</IA>
non associée	Elle est non associée aux champs sémantiques des CC étiquetés dans le billet et/ou les commentaires, mais porteuse d'une évaluation.	IN	De par sa nature, l'instance non associée n'a pas d'attribut, au sens où elle n'est pas reliée aux concepts présents dans le billet ou les commentaires.

Tableau 4 : Deux catégories d'instances

De même que pour les concepts, sans rentrer dans les détails des règles adoptées pour l'annotation, l'utilisateur doit être averti de certaines conventions :

- Une instance est annotée une seule fois, même si le texte du billet ou des commentaires compte plusieurs occurrences ;
- **Les variantes** d'une instance déjà annoté sont **de nouveau annotées**, une seule fois par variante, précisément parce qu'une procédure automatique ne pourrait permettre de les retrouver ;
- Les instances associées sont rattachées au concept dont elles représentent le champ sémantique, par l'attribut concept `cc=""`. Ex : <IA cc="gouvernement">Ministère de l'Ecologie et du Développement Durable</IA>.

2.4 Les évaluations

On distingue cinq types d'évaluations : l'opinion, l'appréciation, l'acceptation – refus, l'accord – désaccord et le jugement. Chacune de ces évaluations comporte un certain nombre de sous-catégories, qui représentent autant de valeurs possibles pour l'attribut `type=""`. Ces

valeurs sont décrites et exemplifiées ensuite, à raison d'un tableau par catégorie d'évaluation.

Évaluation	Balise XML	Attributs
Opinion	Opinion	type (OBL) : sous-catégorie d'opinion exprimée forme (OBL) : forme lexicale du concept ou de l'instance sur laquelle porte l'opinion ironie (OPT) : le segment textuel annoté présente un procédé ironique
Appréciation	Appreciation	type (OBL) : sous-catégorie d'appréciation exprimée forme (OBL) : forme lexicale du concept ou de l'instance sur laquelle porte l'appréciation ironie (OPT) : le segment textuel annoté présente un procédé ironique
Acceptation Refus	Acceptation_Refus	type (OBL) : sous-catégorie d'acceptation ou de refus exprimé forme (OBL) : forme lexicale du concept ou de l'instance sur laquelle porte l'acceptation ou le refus ironie (OPT) : le segment textuel annoté présente un procédé ironique
Accord Désaccord	Accord_Desaccord	type (OBL) : sous-catégorie d'accord ou de désaccord exprimé forme (OBL) : forme lexicale du concept ou de l'instance sur laquelle porte l'accord ou le désaccord ironie (OPT) : le segment textuel annoté présente un procédé ironique
Jugement	Jugement	type (OBL) : sous-catégorie de jugement exprimé forme (OBL) : forme lexicale du concept ou de l'instance sur laquelle porte le jugement ironie (OPT) : le segment textuel annoté présente un procédé ironique

Tableau 5 : Attributs obligatoires des cinq catégories d'évaluations

Là encore, l'utilisateur doit être averti de quelques conventions d'annotation :

- Lorsqu'une évaluation porte sur plusieurs concepts ou sur plusieurs instances, les formes lexicales de ces concepts et/ou instances figurent toutes en valeur de l'attribut **forme=""**, séparées par une virgule. Ex : `<Appreciation type="AIF" forme="adaptation, livre">le plus intéressant</Appreciation>` ;
- L'attribut **ironie=""** n'admet qu'une valeur possible "oui". Ex : Allez, `<Appreciation type="AID" forme="fonction publique" ironie="oui">on remet ça dès le 20 novembre avec tous nos amis</Appreciation>` de la `<CA cc="C1">fonction publique</CA>`¹¹.

¹¹Le corpus d'échantillon ne comportant pas de procédé ironique, cet exemple est extrait du fichier Greve_SNCF-419690-1019685906.xml.

L'opinion

L'opinion comporte cinq sous-catégories.

Balise <opinion> : valeurs possibles de l'attribut « type »	Exemple d'indice linguistique Exemple d'annotation en corpus
<i>Conviction</i>	Ex : je suis persuadé Ex : <Opinion type="Conviction" forme="action, Gouvernement, travail de sensibilisation">ne doit pas se résumer</Opinion>
<i>Supp¹²_Certitude_Forte</i>	Ex : je suppose Ex : <Opinion type="Supp_Certitude_Forte" forme="grève reconductible">Nous pensons</Opinion>
<i>Supp_Certitude_Moyenne</i>	Ex : je crois Ex : <Opinion type="Supp_Certitude_Moyenne" forme="explication">vient très probablement</Opinion>
<i>Supp-Certitude-Faible</i>	Ex : je doute Ex : <Opinion type="Supp_Certitude_Faible" forme="mouvement">je doute</Opinion>
<i>Supp-Pressentiment</i>	Ex : je sens Ex ¹³ : <Opinion type="Supp_Pressentiment" forme="industrie">j'ai comme l'impression</Opinion>

Tableau 6 : Valeurs possibles de l'attribut « type » associé à la balise <opinion>

L'appréciation

L'appréciation comporte six sous-catégories.

Balise <appreciation> : valeurs possibles de l'attribut « type »	Exemple d'indice linguistique Exemple d'annotation en corpus
<i>AEF¹⁴</i>	Ex : je suis satisfait

¹²Supp- signifie supposition.

¹³ Extrait du fichier *economie.xml*

¹⁴AE signifie *Appréciation Explicite*, AI signifie *Appréciation Implicite*, F signifie *Favorable* et D signifie *Défavorable*.

	Ex : <Appreciation type="AEF" forme="réalisation">J'attends avec impatience</Appreciation>
AED	Ex : je trouve décevant Ex : <Appreciation type="AED" forme="lever beaucoup plus tôt">Je serais moi-même énervé</Appreciation>
AE_forme_exclamative_F	Ex : Youpi ! Ex : <Appreciation type="AE_forme_exclamative_F" forme="humour">j'adore</Appreciation> ce style d'<CA cc="C1">humour</CA>!!
AE_forme_exclamative_D	Ex : Tant pis ! Ex : <Appreciation type="AE_forme_exclamative_D" forme="2">j'étais moyennement chaud</Appreciation> pour aller le voir !!!
AIF	Occurrence textuelle positive relevant du locuteur et exprimée autrement que par un verbe modal ou une forme exclamative. Ex : <Appreciation type="AIF" forme="semaine du développement durable">est une bonne chose</Appreciation>
AID	Occurrence textuelle négative relevant du locuteur et exprimée autrement que par un verbe modal ou une forme exclamative. Ex : <Appreciation type="AID" forme="action">nettement insuffisante</Appreciation>

Tableau 7 : Valeurs possibles de l'attribut « type » associé à la balise <appreciation>

L'accord-désaccord

L'accord – désaccord comporte quatre sous-catégories.

Balise <accord_desaccord> : valeurs possibles de l'attribut « type »	Exemple d'indice linguistique Exemple d'annotation en corpus
Acc ¹⁵ _total	Ex : oui, bien sûr Ex : <Accord_Desaccord type="Acc_total" forme="Domenech, Henry">vous avez raison</Accord_Desaccord>

¹⁵Acc signifie *Accord* et Desacc signifie *Désaccord*.

<i>Acc_approximatif</i>	<p>Ex : je suis à peu près d'accord</p> <p>Ex : <Accord_Desaccord type="Acc_approximatif" forme="Darroussin">je suis assez d'accord</Accord_Desaccord></p>
<i>Rectificatif</i>	<p>Ex : oui, mais (annonce une objection)</p> <p>Ex : C'est une <CA cc="C1">action</CA> <Appreciation type="AIF" forme="action">soutenable</Appreciation> <Accord_Desaccord type="Rectificatif" forme="action">mais</Accord_Desaccord> <Appreciation type="AID" forme="action">nettement insuffisante</Appreciation>.</p>
<i>Desacc_total</i>	<p>Ex : certainement pas</p> <p>Ex : <Accord_Desaccord type="Desacc_total" forme="majorité de la population, grèves">contre</Accord_Desaccord></p>

Tableau 8 : Valeurs possibles de l'attribut « type » associé à la balise <accord_desaccord>

L'acceptation-refus

L'acceptation – refus comporte seulement deux sous-catégories.

Balise <acceptation-refus> : valeurs possibles de l'attribut « type »	Exemple d'indice linguistique Exemple d'annotation en corpus
<i>acceptation</i>	<p>Ex : je consens à</p> <p>Ex : <Acceptation_Refus type="Acceptation" forme="réforme des universités">Oui</Acceptation_Refus> à la <CC id_c="C1">réforme des universités</CC></p>
<i>refus</i>	<p>Ex : je m'oppose à</p> <p>Ex : <Acceptation_Refus type="Refus" forme="nucléaire militaire">je refuse</Acceptation_Refus></p>

Tableau 9 : Valeurs possibles de l'attribut « type » associé à la balise <acceptation_refus>

Le jugement

Le jugement comporte six sous-catégories.

Balise <jugement> : valeurs possibles de l'attribut « type »	Exemple d'indice linguistique Exemple d'annotation en corpus
<i>JE¹⁶_positif</i>	Ex : bravo Ex : <Jugement type="JE_positif" forme="miracle">j'espère</Jugement> toujours un <CA cc="C2, Beaujolais nouveau">miracle</CA>
<i>JE_négatif</i>	Ex : je te reproche Ex : <Jugement type="JE_negatif" forme="sélectionneur, match couperet">beaucoup ne lui auraient pas pardonné d'avoir perdu</Jugement>
<i>Jl_positif_félicitation</i>	Ex : vous avez été magnifique Ex ¹⁷ : <Jugement type="Jl_positif_félicitation" forme="auteurs">Je tiens surtout à féliciter</Jugement> les <CA cc="C1">auteurs</CA>
<i>Jl_positif_pardon</i>	Ex : n'en parlons plus, c'est oublié Ex ¹⁸ : <Jugement type="Jl_positif_pardon" forme="imperfections">il faut nous pardonner</Jugement> les diverses <CA cc="C7">imperfections</CA>
<i>Jl_négatif_accusation</i>	Ex : vous avez enfoncé ma porte Ex ¹⁹ : Quand <Jugement type="Jl_négatif_accusation" forme="milieux">on a la prétention de</Jugement> vouloir diriger l'humanité vers la lumière du paradis Social
<i>Jl_négatif_reproche</i>	Ex : ton attitude n'est pas correcte Ex ²⁰ : <Jugement type="Jl_négatif_reproche" forme="look, hrithik">ils aurais pue fair 1 effort</Jugement> quand meme

Tableau 10 : Valeurs possibles de l'attribut « type » associé à la balise <jugement>

¹⁶ JE signifie Jugement Explicite, JI signifie Jugement Implicite.

¹⁷ Extrait du fichier *rever (3).xml*

¹⁸ Extrait du fichier *blogzine (16).xml*

¹⁹ Extrait du fichier *politique.xml*

²⁰ Extrait du fichier *cinema (5).xml*

2.5 Synthèse chiffrée

En guise de conclusion, nous proposons une synthèse chiffrée des annotations effectuées. Figure d'abord le tableau correspondant à la quantité de concepts et d'instances annotés par sous catégories, puis le tableau récapitulatif de la distribution des évaluations annotées selon les différentes parties du corpus.

CC (formes différentes)	CA	CN	IA	IN	TOTAL
1000	4322	344	1161	49	6876

Tableau 11 : Détail du nombre de concepts et d'instances annotées par sous-catégorie

PARTIES DU CORPUS	BILLETS	COMMENTAIRES	TOTAL
1 - 76 billets	1135	941	2076
2 - 34 billets	608	465	1073
3 - 90 billets	1389	405	1794
TOTAL	3132	1811	4943

Tableau 12 : Distribution de la quantité d'évaluations annotées selon les différentes parties du corpus

	Nb phrases « . »	Nb phrases « ! »	Nb phrases « ? »	Nb phrases total
Corpus 1.1	3495	802	184	4481
Corpus 1.2	817	264	54	1135
Corpus 1.3	2831	485	219	3535
Total	7143	1551	457	9151

Tableau 13 : Distribution de la quantité et de la qualité des mots et phrases annotées dans le corpus

Pour plus de transparence et de clarté, figurent en annexes, le fichier DTD et un fichier xml d'exemple sur le sujet Harry Potter.

3. Annexes

3.1 La DTD

```
<!-- Une page est composée d'un billet et éventuellement des commentaires -->
<!-- Tous les attributs sont remplis par OVERBLOG -->
<!-- tags_blog : thématiques secondaires entrées par auteur pour son blog -->
```

```
<!-- mes_mediametrie : mesure médiamétrie du blog -->
<!-- mes_blog_rank : mesure du blog rank du blog -->
```

```
<!ELEMENT page (billet,comment*)>
<!ATTLIST page
  url CDATA #REQUIRED
  thematique CDATA #REQUIRED
  tags_blog CDATA #IMPLIED
  mes_mediametrie CDATA #IMPLIED
  mes_blog_rank CDATA #IMPLIED >
```

```
<!-- Un billet est composé d'une date, d'un titre et d'un texte -->
<!-- Attributs OVERBLOG : identifiant du billet, url du billet, auteur, agenn profession -->
<!-- Attributs ANNOTATION : type de texte, organisation du discours, orthographe du billet,
syntaxe du billet -->
```

```
<!ELEMENT billet (date,titre,texte)>
<!ATTLIST billet
  id_b ID #REQUIRED
  url CDATA #REQUIRED
  auteur CDATA #REQUIRED
  age CDATA #IMPLIED
  profession CDATA #IMPLIED
  syntaxe (correcte|erronee) #REQUIRED
  orthographe (standard|erronee) #REQUIRED>
```

```
<!-- REMARQUES ESTELLE -->
<!-- type
(Scientifique|Faits_Divers|Commentaires|Information|Editorial|Reportage|Politique|Récit|Rec
ette|Notice|Règles_du_jeu) #IMPLIED -->
<!-- organisation (énonciatif|argumentatif|descriptif|narratif) #REQUIRED -->
```

```
<!-- Dans un titre et un texte, on peut noter des concepts et des observations lexicales -
sentimentales -->
```

```
<!ELEMENT date (#PCDATA)>
<!ELEMENT titre (#PCDATA | CC | CA | CN | IA | IN | Opinion | Appreciation |
Acceptation_Refus | Accord_Desaccord | Jugement)*>
<!ELEMENT texte (partie)+>
<!ELEMENT textco (partie)+>
<!ELEMENT partie (#PCDATA | CC | CA | CN | IA | IN | Opinion | Appreciation |
Acceptation_Refus | Accord_Desaccord | Jugement)*>
```

```
<!-- pour exprimer qu'une partie appartient à un mode de discours -->
<!ATTLIST partie
  organisation (enonciatif|argumentatif|descriptif|narratif) #REQUIRED>
```

```
<!-- Un commentaire est composé d'une date, d'un titre et d'un texte comme le billet -->
<!-- Attribut : identifiant du commentaire, identifiant du billet qu'il commente, type de texte ,
organisation du discours, orthographe du billet, syntaxe du billet -->
```

```
<!ELEMENT comment (date,titre,textco)>
<!ATTLIST comment
```

```

id_co ID #REQUIRED
auteur CDATA #REQUIRED
syntaxe (correcte|erronee) #REQUIRED
orthographe (standard|erronee) #REQUIRED>

<!-- REMARQUES ESTELLE -->
<!-- organisation (énonciatif|argumentatif|descriptif|narratif) #REQUIRED -->

<!-- Les concepts : 3 types -->
<!-- cc : concept concerné / ca : concept associé / cn : concept non associé -->
<!-- attention l'attribut niveau n'est pas obligatoire, si non présent, on considère qu'il est égal à
courant -->

<!ELEMENT CC (#PCDATA)>
<!ATTLIST CC
    id_c CDATA #REQUIRED
    niveau (terminologique|soutenu|familier|vulgaire) #IMPLIED
hyponyme CDATA #IMPLIED>
<!ELEMENT CA (#PCDATA)>
<!ATTLIST CA
    cc CDATA #IMPLIED
    niveau (terminologique|soutenu|familier|vulgaire) #IMPLIED>

<!ELEMENT CN (#PCDATA)>
<!ATTLIST CN
    niveau (terminologique|soutenu|familier|vulgaire) #IMPLIED>

<!ELEMENT IN (#PCDATA)>

<!ELEMENT IA (#PCDATA)>
<!ATTLIST IA
    cc CDATA #IMPLIED >

<!-- Les différentes observations lexicales et sentimentales -->
<!-- chaque observation est spécifiée par son type et id_c : le ou les identifiants des concepts
sur lesquels l'observation est faite (si plusieurs : espace entre les identifiants des concepts -->
<!ELEMENT Opinion (#PCDATA)>
<!ATTLIST Opinion
    type
(Conviction|Supp_Certitude_Forte|Supp_Certitude_Moyenne|Supp_Certitude_Faible|Supp_Pr
essentiment) #REQUIRED
    forme CDATA #REQUIRED
    ironie (oui) #IMPLIED>

<!ELEMENT Appreciation (#PCDATA)>
<!ATTLIST Appreciation
    type (AIF|AID|AEF|AED|AE_forme_exclamative_F|AE_forme_exclamative_D)
#REQUIRED
    forme CDATA #REQUIRED
    ironie (oui) #IMPLIED>

```

```
<!ELEMENT Acceptation_Refus (#PCDATA)>
<!ATTLIST Acceptation_Refus
  type (Acceptation|Refus) #REQUIRED
  forme CDATA #REQUIRED
  ironie (oui) #IMPLIED>
```

```
<!ELEMENT Accord_Desaccord (#PCDATA)>
<!ATTLIST Accord_Desaccord
  type (Acc_total|Acc_approximatif|Rectificatif|Desacc_total) #REQUIRED
  forme CDATA #REQUIRED
  ironie (oui) #IMPLIED>
```

```
<!ELEMENT Jugement (#PCDATA)>
<!ATTLIST Jugement
  type
  (JE_positif|JE_negatif|JI_positif_felicitation|JI_positif_pardon|JI_negatif_accusation|JI_negatif_reproche) #REQUIRED
  forme CDATA #REQUIRED
  ironie (oui) #IMPLIED>
```

3.2 Exemple de fichier annoté : Harry Potter

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE page SYSTEM "../pagev4.dtd">
<page mes_blog_rank="3" mes_mediametrie=""
tags_blog="humour,littérature,écriture,publicité,pub,antidote" thematique="Harry Potter"
url="http://sarahleslie.over-blog.com/">
  <billet age="" auteur="hitOmi" id_b="B7935967" profession="" url="http://sarahleslie.over-
blog.com/article-1081261.html" orthographe="standard" syntaxe="correcte">
    <date>2005-10-26 00:00:00</date>
    <titre><IA cc="film, livre, livres, roman">Harry Potter et La Coupe De Feu</IA> [<IA
cc="film, livre, livres, roman">The Gobelet Of Fire</IA>]</titre>
    <texte>
    <partie organisation="enonciatif">
      Quand ? Le 30 Novembre ! Un nouvel <CA cc="C1">évènement</CA> <Appreciation
type="AIF" forme="évènement">très attendu</Appreciation> chez les <CA cc="C1, C2, C3,
Harry Potter">fans</CA> de <IA cc="film, livre, livres, roman">Harry Potter</IA> (comme
moi, bien que <Appreciation type="AED" forme="adaptations">relativement
déçue</Appreciation> par les précédentes <CA cc="C2">adaptations</CA> par rapport aux
<CC id_c="C2" hyponyme="C3">livres</CC>) est le quatrième <CA cc="C1, C2,
C3">opus</CA> de l'<CA cc="C3">épopée</CA> de <IA cc="roman, épopée, Harry
Potter">J.K. Rowling</IA>, <IA cc="film, livre, livres, roman">Harry et la Coupe de
Feu</IA> (<Appreciation type="AIF" forme="Harry et la Coupe de Feu">le
meilleur</Appreciation> des 6 Harry Potter selon moi). D'après les premières <CA
cc="C1">images</CA> qui ont été offertes à mes petits yeux ébahis, il me semble que l'<CA
cc="C2, Harry et la Coupe de Feu">univers plutôt sombre</CA> <Appreciation type="AIF"
forme="univers plutôt sombre">a été bien restitué</Appreciation>... <Jugement
type="JE_positif" forme="univers plutôt sombre, film">J'espère</Jugement> ne pas trop
m'avancer même si je sais d'ores et déjà que le <CC id_c="C2" hyponyme="C3">livre</CC>
```

ne peut être développé dans le quart de son intégralité à cause du peu de temps que dure le <CC id_c="C1">film</CC> (2h37, en comptant le <CA cc="C1">générique</CA> !). En tout cas si cette <CA cc="C2">adapatation</CA> <Appreciation type="AIF" forme="adaptation, livre">reflète fidèlement</Appreciation> le livre il sera <Appreciation type="AIF" forme="adaptation, livre">le plus spectaculaire</Appreciation> et <Appreciation type="AIF" forme="adaptation, livre">le plus intéressant</Appreciation> de quatre <CA cc="C2, C3">volets</CA>... et aussi le début de l'assombrissement de l'<CA cc="C2, Harry et la Coupe de Feu">univers magique</CA> d'Harry Potter, dont l'âge et les épaules semblent croître aussi vite que ses problèmes avec <IA cc="C1, C2, C3">Lord Voldemort</IA>... <Appreciation type="AEF" forme="réalisation">J'attends avec impatience</Appreciation> la <CA cc="C1">réalisation</CA> de <IA cc="C1, réalisation">Mike Newell</IA> (<IA cc="film, réalisation, Mike Newell">Le Sourire de Mona Lisa</IA>, et <IA cc="film, réalisation, Mike Newell">Donie Brasco</IA>) pour retrouver <IA cc="C1, Harry Potter">Daniel Radcliffe</IA>, <IA cc="C1, Harry Potter">Emma Watson</IA> et <IA cc="C1, Harry Potter">Rupert Grint</IA> dans le <CC id_c="C3">roman</CC> <Appreciation type="AIF" forme="roman">le plus magique</Appreciation> de J.K. Rowling... "Bande Annonce 1 "Bande Annonce 2 "Bande Annonce 3

</partie>
</texte>
</billet>
</page>