SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801254




# Prise en charge et phénomènes de portée : retour d'expériences dans un corpus de dépêches de presse.

Damiani, Marine & Battistelli, Delphine

MoDyCo, UMR 7114, 200 avenue de la République 92000 Nanterre
marinedamiani@gmail.com, del.battistelli@gmail.com

## 1 Introduction

Le travail présenté ici s'inscrit dans le cadre du projet ANR ChronoLines[1] dont la problématique est la génération d'interfaces innovantes pour la visualisation, selon des critères temporels, d'informations contenues dans des dépêches de l'Agence France Presse (AFP). L'application visée dans ce projet s'apparente donc à des outils visuels pour des systèmes de recherche d'information de type *timelines*, très prisés des utilisateurs soucieux d'obtenir une représentation concise de l'information (voir p.ex. Alonso et al, 2009). Par rapport aux types d'outils actuellement développés, son originalité, réside dans la prise en compte du fait que, indépendamment de leur ancrage calendaire, des situations peuvent être présentées comme certaines, inaccomplies, seulement possibles ou probables, voire niées par un énonciateur qui peut être l'auteur du texte, mais qui peut aussi être un autre énonciateur (explicite ou non) dont l'auteur rapporte une partie des propos qu'il a entendus, lus, imaginés... En cela, la démarche peut être rapprochée de travaux sur l'analyse de la subjectivité ou des opinions (p.ex. Wilson et Wiebe, 2005) ou encore de travaux qui s'intéressent au degré de « factualité » des événements (p.ex. Sauri et Pustejovsky, 2012). Dans notre corpus de dépêches de presse, nous avons pu noter que près de 90% des phrases contiennent au moins un indice renvoyant à une modalité épistémique et / ou à une distanciation énonciative (de type discours rapporté), d'où l'importance que de pouvoir traiter ce type de phénomènes linguistiques qui instaurent une forme de distanciation par rapport à des contenus propositionnels. Contrairement à la plupart des autres approches, nous avons choisi de ne pas traiter ces deux types de caractéristiques séparément, puisque les deux sont impliquées dans ce qu'on appelle la « prise en charge énonciative » (terme très souvent associé à des travaux inscrits dans la théorie de l'énonciation). Comme souligné entre autres dans (Dendale et Coltier, 2011), les marqueurs de prise en charge énonciative relèvent en effet d'une étroite et complexe interaction entre les catégories de modalité et d'évidentialité.

Nous présentons dans cet article la méthodologie que nous avons mise en place en vue d'automatiser l'annotation de cette composante énonciative et modale dans ce corpus de dépêches de presse. Cette méthodologie prend en compte le fait que, en plus de la nécessité d'identifier et de classer sémantiquement les indices linguistiques entrant en jeu, il est nécessaire d'aborder la question de la portée de chacun d'entre eux. Cette question est rendue d'autant plus complexe que de nombreux indices peuvent être présents simultanément dans une phrase, (voir exemples 1 et 2.).

1. M. Arabi **a exprimé**$^{\text{indice1}}$ [**le souhait**$^{\text{indice2}}$ [d'aider la Syrie à surmonter cette phase]$_{\text{portée2}}$]$_{\text{portée1}}$
2. Paul **veut**$^{\text{indice1}}$ **sûrement**$^{\text{indice2}}$ que [Mary vienne.] $_{\text{portée}}$

Après une brève présentation de l'ancrage théorique qui est le nôtre ainsi que de nos principes méthodologiques, nous présentons notre système d'annotation automatique permettant d'identifier l'organisation du texte en segments textuels selon leurs caractéristiques énonciatives et modales. Puis dans un troisième temps, nous présenterons deux approches d'évaluation permettant un retour réflexif sur le développement du système d'annotation. Enfin, nous exposerons comment les annotations produites



Article available at http://www.shs-conferences.org or http://dx.doi.org/10.1051/shsconf/20140801254

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801254


peuevnt être utilisées au sein d'un système de recherche d'information visant à répondre aux besoins de l'AFP.

## 2 Le phénomène de prise en charge énonciative et modale

La notion de modalité, qui est étroitement liée à celle d'évidentialité[2], a été étudiée sous différentes perspectives (logique, philosophique, linguistique) - voir par exemple (Palmer, 2001; Nuyts, 2006.). En linguistique, la modalité peut être considérée avec un point de vue énonciatif - voir (Bally, 1935; Benveniste, 1966; Culioli, 1973). Selon ce point de vue, qui est celui que nous adoptons, un discours résulte de certaines opérations langagières ; parmi elles, celles de prédication et celles de prise en charge énonciative. L'une et l'autre laissent un certain nombre de traces de surface dans le discours. Ce sont les marques de prise en charge énonciative qui font l'objet du travail que nous présentons ici.

Selon une perspective théorique énonciative, tout discours (*a fortiori* une simple phrase) est nécessairement présenté du point de vue d'une source énonciative (dans notre cas, le journaliste qui a écrit la dépêche de presse) prenant en charge le(s) contenu(s) prédicatif(s) énoncé(s). Ainsi, tout discours a toujours une source par défaut qui est son auteur. Toute dépêche de presse peut donc être considérée comme un segment textuel ayant des valeurs énonciative (= 'auteur') et modale (= 'vrai') par « défaut » (selon la maxime gricéenne de qualité qui impose que le locuteur croit ce qu'il dit et qu'il a de bonnes raisons de le croire). La plupart du temps, dans un discours (et donc même au sein d'une seule phrase), des segments textuels ayant des valeurs énonciatives et modales différentes peuvent être identifiés. Par exemple, la rencontre d'un verbe de discours tel que *dire*, *répondre* ou *annoncer*, introduit la perception d'une variation de la valeur énonciative (au sens où une source secondaire peut être introduite) mais pas de variation sur le plan modal ; *a contrario*, des adverbes tels que *sûrement* ou *probablement*, introduisent une variation uniquement au niveau modal ; enfin, des verbes comme *prétendre, croire* ou *imaginer* vont eux induire une variation tant sur le plan énonciatif que modal.

Un tel panorama permet d'entrevoir l'imbrication des différentes sources énonciatives et des modalités (en particulier épistémiques) impactant tel ou tel contenu propositionnel. Ce type d'approche de la modalité permet en outre de mettre en avant l'hétérogénéité discursive et d'aborder la notion de modalité d'une façon originale en s'intéressant à l'influence entre la modalité et les autres catégories linguistiques proches (l'évidentialité, mais aussi le temps et l'aspect – cf. (Bybee et al, 1994)). Le développement d'un système automatique peut ainsi, selon nous, au-delà de l'intérêt applicatif, permettre d'apporter un éclairage sur le fonctionnement et l'interaction complexe de ces catégories linguistiques et ainsi d'arriver à mieux les comprendre pour les décrire.

De plus, bien que les marqueurs de modalité en français - dans leur relation étroite avec les marqueurs d'évidentialité - aient été décrits systématiquement (voir par exemple Gosselin, 2010; Le Querler, 2004), il n'existe pas encore de corpus de référence proposant l'annotation des caractéristiques énonciatives et modales comme une tâche de délimitation discursive et c'est l'objectif que nous cherchons à atteindre. Ce problème de l'identification automatique d'indices modaux et de leur portée sur la base d'une analyse syntaxique a été principalement et seulement étudié dans les textes biomédicaux en anglais (Vincze et al., 2008).

### 2.1 Comment l'annoter

L'observation des différentes formes de prise en charge énonciative et modale dans notre corpus a fait apparaître que l'on pouvait distinguer aux moins deux grands types d'indices: les *indices d'ouverture* qui correspondent à l'ouverture d'un nouveau segment textuel (cette catégorie d'indices a la propriété syntaxique de gouverner un autre segment textuel, p.ex. indice1 dans l'exemple 2 et indices 1 et 2 dans l'exemple 1.) et *les indices de type modifieur* qui viennent modifier le contexte de validation d'un segment textuel préalablement repéré (ces indices sont principalement des adjectifs et des adverbes p.ex. indice2 dans l'exemple 2). L'identification des indices d'ouverture (et des dépendants syntaxiques sous la portée de ceux-ci) amène à découper le texte en différents segments textuels tandis que le repérage des

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801254



indices de type modifieur sert à indiquer une variation du degré de prise en charge énonciative et modale au sein d'un segment textuel déjà identifié. Ainsi, l'indice 1 de l'exemple 1 *exprimer* est un indice d'ouverture qui va modifier le contexte énonciatif du segment textuel qui est sous sa portée (ce segment est noté portée 1). A l'intérieur de ce segment, on idientifie un second indice *le souhait* qui va à son tour déclencher l'ouverture d'un nouveau segment textuel (noté portée 2) et venir modifier son contexte énonciatif et modal. Sur le plan sémantique, nous faisons une différence entre les indices d'ouverture modifiant uniquement le contexte énonciatif d'un segment (dans cette catégorie on trouve les verbes de parole ainsi que les constructions prépositionnelles en *Selon*), ceux qui modifient à la fois le contexte énonciatif et modal (par exemple les verbes d'attitude propositionnelle) et ceux qui ne modifient que le contexte modal (par exemple les tournures au conditionnel ou les subordonnées de cause). La description de nos ressources et leur intégration dans le système sont décrites de façon détaillée dans Batistelli et Damiani (2013).

La tâche d'annotation que nous présentons ici concerne uniquement l'annotation de ce que nous appelons des *indices d'ouverture* (ceux-ci conduisent à la modification du niveau de prise en charge énonciatif et/ou modal d'un segment textuel) et de leurs portées. La portée d'un indice d'ouverture correspond au segment textuel impacté par la variation du niveau de prise en charge énonciative et modale induite par cet indice. Le tableau 1 présente les quatre classes d'indices d'ouverture que notre système prend actuellement en compte et donne pour chacun d'entre eux quelques exemples des dépendants syntaxiques apparaissant sous la portée de l'indice.

| **Indices d'ouverture** | **Portée** |
|---|---|
| **Verbes** | **Objet direct et/ou indirect** |
| *Verbes de parole, verbes d'attitude propositionnelle* | *Paul* ***promet***$^{\text{indice}}$ *qu'[il viendra]*$_{\text{portée}}$<br>*Paul* ***veut***$^{\text{indice}}$ *[venir]*$_{\text{portée}}$ |
| **Noms** | **Complément du nom, proposition relative** |
| *Noms prédicatifs* | *C'est* ***son souhait***$^{\text{indice}}$ *[d'être impliqué]*$_{\text{portée}}$ |
| **Morphologique** | **Tous les complements du verbe** |
| *Conditionnel* | Jean **aurait**$^{\text{indice1}}$ **annoncé**$^{\text{indice2}}$ [le départ de Paul.]$_{\text{portée}}$ |
| **Construction syntaxique** | **Proposition principale** |
| *Subordonnée de condition* | *[Marie refuse de donner son approbation]*$_{\text{portée}}$ ***à moins que Paul accepte***$^{\text{indice}}$ |
| *Construction prépositionnelle* | ***Selon***$^{\text{indice}}$ *Paul, [Marie va venir]*$_{\text{portée}}$<br>***A première vue***$^{\text{indice}}$*, [Marie a raison]*$_{\text{portée}}$ |

Tableau 1. Indices d'ouverture et portée associée

Ces quatre classes d'indices ont été choisies au terme d'une étude approfondie de notre corpus et représentent les indices quantitativement les plus présents pour ce corpus donné. Rappelons que, à l'instar de travaux ancrés dans les théories énonciatives du type de celle développée par (Culioli, 1973), un indice (ou marqueur) peut tout aussi bien être lexical, grammatical ou relever d'une construction syntaxique. Notre système reconnait ainsi une centaine de verbes et une trentaine de noms pouvant jouer le rôle d'indices d'ouverture. D'autre part, il identifie les marqueurs du conditionnel et du futur, les subordonnées de condition introduites par *si* et par *à condition que* ou encore les constructions prépositionnelles en *selon*. L'objectif n'est pas de traiter tous les verbes, ni tous les types de subordonnées ou de constructions prépositionnelles mais de travailler avec des formes représentatives de ces classes dans un corpus donné afin de valider la méthode et le mode de fonctionnement du système. La variété des indices lexicaux comme grammaticaux ou syntaxiques pris en compte pourra par la suite être étendue.

Nous allons maintenant présenter brièvement comment sur la base du repérage de ces indices notre système d'annotation fonctionne.

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801254


## 2.2 Architecture du système d'annotation automatique

Notre système d'annotation se base dans un premier temps sur le repérage dans le corpus *des indices d'ouverture* que nous venons de décrire. Une fois ces indices repérés, le système cherche à borner (à gauche et à droite) la portée de l'indice. Cette tâche de calcul de la portée syntaxique des indices sémantiques repérés s'effectue en utilisant un analyseur syntaxique robuste, FRMG (FRench MetaGrammar), développé pour le français (De La Clergerie et al., 2009). Notons que le recours à une analyse syntaxique basée sur des structures de dépendance fait apparaitre plus directement les structures prédicats-arguments qu'une analyse en constituants immédiats[3]. L'identification de la relation entre un prédicat (qui pour nous joue le rôle d'un indice) et ses arguments (p.ex. suj, obj et a_obj dans les énoncés 9 et 10) va donc permettre d'extraire les arguments étant sous la portée de l'indice. De plus, comme le souligne (Kahane, 2010), l'analyse en dépendance se rapproche d'avantage d'une représentation sémantique de la phrase et permet de dissocier l'ordre des mots de la structure syntaxique. Le fait que les unités lexicales soient mises au centre de la structure syntaxique permet en outre d'exprimer simplement les relations lexicales comme la valence. Cette tâche de découpage s'apparente à la recherche de la portée (ou scope en anglais (p.ex. Farkas, R. et al, 2010 ; Kilicoglu et Bergler, 2010)) de l'indice dans la phrase. Notons que certains indices sémantiques peuvent également avoir une portée extra-phrastique (Charolles, 1997), mais ces cas ne seront pas abordés ici.

Ce lien entre la représentation syntaxique en dépendances et l'utilisation d'indices sémantiques (lexicaux, morphologiques ou syntaxiques) est au cœur de notre méthodologie. En effet, la spécificité de notre système d'annotation se base sur la description fine d'indices que nous repérons en lien avec les contextes d'apparition de ces indices. Le croisement de ces informations est géré sous la forme de règles symboliques au sein du système d'annotation. Chaque indice repéré par le système va ainsi être lié au segment textuel qui est sous sa portée et qui se compose des dépendants syntaxiques de l'indice (comme nous l'avons illustré dans la section précédente, le type de dépendants syntaiques considérés varie en fonction du type de l'indice).

Afin d'évaluer les performances de notre système, nous proposons deux modes d'évaluation, que nous présentons dans la section suivante. Le premier met l'accent sur le calcul de l'accord inter-annotateur entre deux annotateurs experts et les premiers résultats de notre système automatique pour la même tâche d'annotation (Damiani et Battistelli, 2013). Pour le second, nous avons développé un outillage informatique mis à la disposition du linguiste expert pour évaluer de façon rapide et simple, *via* une interface, la qualité des annotations proposées par le système.

# 3 Evaluation du système d'annotation automatique

## 3.1 Un premier mode fondé sur le calcul de l'accord inter-annotateurs

### 3.1.1 Méthodologie proposée

L'évaluation d'un système d'annotation nécessite de disposer d'un corpus annoté pouvant servir de référence. Or dans notre cas, étant donné que nous avons défini nous-même le type de segments textuels que nous souhaitions annoter, il n'existe aucun corpus de référence. Pour évaluer la qualité de notre système nous avons donc dû produire notre propre corpus de référence correspondant à notre tâche d'annotation. Ce corpus de référence a été construit à partir de la confrontation des annotations produites par deux annotateurs différents sur un même jeu de données.

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801254



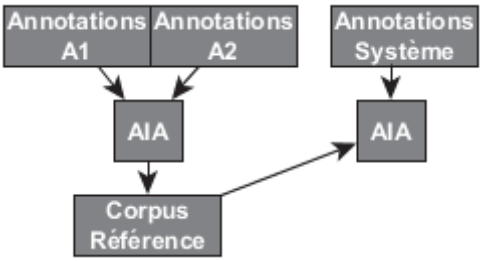


Figure 1.Construction du corpus de référence et évaluation

La figure 1 illustre les étapes qui ont été suivies pour construire ce corpus de référence et pour évaluer le système d'annotation automatique. Dans un premier temps, deux annotateurs (désormais A1 et A2), tous deux experts en linguistique, ont annoté séparément un échantillon du corpus. Cette tâche d'annotation manuelle a été réalisée à l'aide de l'outil d'annotation Glozz (Widlöcher et Mathet, 2012) sur un échantillon de 20 textes dans lesquels 256 indices d'ouverture ont été identifiés (voir tableau 2).

| # Phrase | Total | Verbes | Noms | Morpho | Syntaxique |
|---|---|---|---|---|---|
| 199 | 256 | 210 | 4 | 11 | 31 |

Tableau 2. Répartition des indices d'ouverture par classe

Nous avons ensuite confronté ces deux jeux d'annotations en réalisant un calcul d'accord-inter-annotateurs (AIA) basé sur des mesures de rappel, de précision et de F-Mesure (F1) ainsi qu'une comparaison manuelle des divergences d'annotations.
A partir de ce travail nous avons produit une version révisée des annotations de ce sous-corpus tenant compte des divergences mises en avant par l'évaluation, ce nouveau jeu d'annotations constituant alors le corpus de référence de notre tâche d'annotation. Dans un second temps, ce corpus de référence a servi à évaluer le système d'annotation automatique.

### 3.1.2 Analyse des résultats

Les résultats du calcul de l'accord inter-annotateurs experts (voir tableau 3) montrent que l'accord entre les deux annotateurs est élevé pour les indices, mais pas très bon pour le repérage de la portée. En comparant les deux ensembles d'annotations dans le détail, nous avons observé dans notre corpus que certains segments textuels peuvent être inclus ou exclus de la portée de l'indice en fonction de l'interprétation que l'annotateur privilégie. L'exemple 9 montre l'annotation de la portée proposée par l'annotateur A1. Comme nous pouvons le voir, le segment textuel *qui a débuté lundi* est inclus dans la portée de l'indice *a indiqué*, mais il est exclu dans l'annotation proposée par l'annotateur A2. Dans ce cas particulier, nous considérons que les deux interprétations sont acceptables puisque nous ne pouvons pas dire avec certitude si ce segment est présenté du point de vue du journaliste ou du point de vue de la source *Ses avocats*. Le même phénomène est souvent observé avec des adverbiaux temporels qui ne peuvent pas être interprétés sans ambiguïté comme faisant partie ou non de la portée d'un indice. Dans ces deux types de cas, l'annotateur doit utiliser le contexte ainsi que son bagage linguistique pour décider. Cette observation soulève la question - déjà mentionnée dans Farkas et al. (2010) - de savoir s'il faut ou non définir une limite stricte à la portée d'un indice.

Nous proposons d'aborder cette question en évaluant les annotations sur la portée à la fois de manière stricte et également avec un calcul plus souple. Dans l'interprétation souple, on distingue les segments dont les frontières sont détectées avec une correspondance exacte de ceux qui sont détectés avec des frontières différentes mais dont la délimitation reste correcte d'un point de vue interprétatif (comme illustré dans l'exemple 2).

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801254


3. [Le procès devant un tribunal militaire d'un blogueur égyptien arrêté pour avoir critiqué l'armée, **qui a débuté lundi**, a été ajourné à dimanche] $_{\text{portée}}$, **a indiqué**$^{\text{indice}}$ mardi un de ses avocats.

Pour mesurer la distinction entre l'utilisation de frontières strictes ou souples pour le repérage de la portée d'un indice, nous proposons d'utiliser en plus d'un calcul classique de rappel et de précision des mesures de rappel et de précision pondérées. Pour mettre en place ces mesures, nous distinguons la portée stricte (PS) de la portée flexible (PF) à laquelle nous attribuons un facteur 0.5. Nous calculons ensuite la précision et le rappel pondérés de la manière suivante :

$$Précision\ pondérée = \frac{PS + 0.5 \times PF}{Ref}$$

$$Rappel\ pondéré = \frac{PS + 0.5 \times PF}{Rel}$$

- PS (portée stricte): le nombre d'entités avec une limite de portée stricte
- PF (portée flexible): le nombre d'entités avec une limite de domaine d'application flexible
- Ref: le nombre d'entités de référence (cad idéalement identifiées)
- Rel: le nombre d'entités pertinentes (cad correctement identifiées)

La distinction entre l'évaluation tenant compte de la portée pondérée a révélé que, dans un nombre important de cas (dans cette expérimentation environ 10%), les deux annotateurs sont en désaccord dans leur annotation, alors que pourtant les deux interprétations sont correctes. Cette observation nous a permis de repenser nos objectifs d'annotation et d'intégrer cette finesse interprétative dans notre corpus de référence.

| **A1 /A2** | | **précision** | **rappel** | **F1** |
|---|---|---|---|---|
| Indices | | 0.85 | 0.86 | 0.86 |
| Portée | | 0.79 | 0.72 | 0.76 |
| Portée pondérée | | 0.84 | 0.77 | 0.80 |
| **PS** | **PF** | **Rel** | **Ref** | |
| 185 | 22 | 256 | 234 | |

Tableau 3. Evaluation des annotations de l'annotateur A1 au regard de celles de l'annotateur A2

| **Corpus Référence /Système** | | **précision** | **rappel** | **F1** |
|---|---|---|---|---|
| Indices | | 0.83 | 0.85 | 0.84 |
| Portée | | 0.52 | 0.59 | 0.55 |
| Portée pondérée | | 0.67 | 0.76 | 0.71 |
| **PS** | **PF** | **Rel** | **Ref** | |
| 59 | 33 | 100 | 113 | |

Tableau 4. Evaluation des annotations produites par le système au regard des annotations du corpus de référence

Dans un second temps, nous avons évalué la première version des annotations produites par notre système automatique avec les annotations du corpus de référence construit à l'étape précédente (voir tableau 4). Cette évaluation a été effectuée sur un sous-ensemble du corpus contenant 100 indices avec leurs portées. Les résultats de cette évaluation montrent que la détection des indices est bonne (0.84 de F1), tandis que la délimitation de la portée est plutôt faible (0.55 de F1). Cela s'explique en partie par le fait que l'analyseur syntaxique produit des erreurs d'analyse (erreurs d'étiquetage morpho-syntaxique ou de découpage syntaxique, mauvais attachement des conjonctions de coordination, *etc*.). D'autre part, on peut observer une différence significative entre le calcul de la F1 prenant en compte la portée pondérée par rapport à la F1 ne la prenant pas en compte (on passe de 0.55 à 0.71). La F1 incluant la portée pondérée se rapproche finalement de la valeur obtenue avec l'accord inter-annotateurs entre les deux experts. Cette observation tendrait à montrer l'intéret de distinguer dans ce type de tâche une portée stricte et une portée pondérée. En effet, le système d'annotation automatique n'étant pas capable comme l'humain de faire intervenir son interprétation il est dommage de pénaliser fortement l'évaluation en comptabilisant directement comme faux un cas limite. Ce phénomène des frontières difficilement décidables représente

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801254



10 % des désaccords entre les annotateurs experts (soit 22 cas) et 30 % dans l'évaluation du système (soit 33 cas), et doit être pris en compte pour améliorer la qualité des annotations.

Cette phase d'évaluation a permis de montrer que le système était capable de répondre à la tâche d'annotation proposée (malgré la présence inhérente d'erreurs) mais a également mis en avant le fait que l'évaluation de ce type d'annotations était très couteuse en temps. Afin de produire un corpus annoté de qualité tout en réduisant au maximum le temps consacré à l'évaluation puis à la révision des annotations, nous avons mis en place une interface permettant de faciliter la tâche d'évaluation.

## 3.2 Un second mode d'évaluation, ou comment faciliter la tâche d'évaluation à un linguiste

L'interface que nous présentons maintenant permet d'envisager une évaluation à la fois quantitative des résultats d'une campagne d'annotation mais permet également d'obtenir - grâce à l'expertise des évaluateurs linguistes - un retour sur l'origine des erreurs. Ce type d'évaluation permet non seulement d'évaluer l'efficience du système mais également d'envisager un retour reflexif pour envisager des améliorations dans le système et ce de manière beaucoup plus rapide qu'avec la méthode présentée dans la section précédente. En effet, outre son côté ergonomique pour l'expert, l'utilisation d'une interface pour évaluer la correction des annotations produites permet de réduire considérablement le temps nécessaire à l'évaluation. A titre d'exemple, l'évaluation de 600 indices et de leur portée prend à l'expert environ 8h sans interface contre seulement 3h *via* l'interface. Ce gain de temps permet donc d'envisager une évaluation sur un corpus plus large.

La figure 2 montre une vue de l'interface lors d'une des étapes de l'évaluation. Au cours de celle-ci, l'expert doit cocher si l'annotation de la portée de l'indice a été correctement effectuée ou non. De plus, une vue de l'arbre syntaxique en dépendance est proposée comme support à l'expert afin de voir si les erreurs éventuelles peuvent être imputables à une analyse syntaxique erronée.

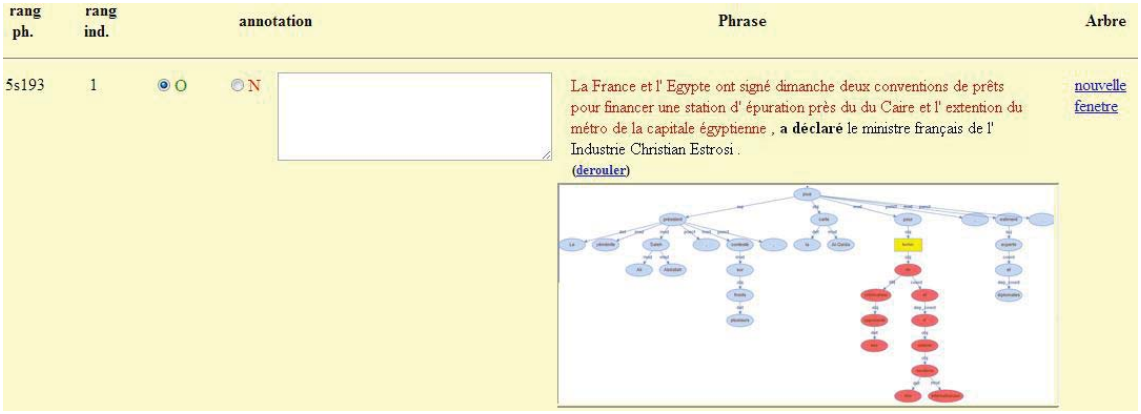


Figure 2 - Interface php pour l'évaluation par un linguiste

Comme l'illustre la figure 3 (qui décrit les étapes du processus d'amélioration de notre système), le module d'annotation automatique prend en entrée le corpus de dépêches à annoter. En sortie du module, le corpus annoté peut être récupéré, soit sous la forme d'un corpus annoté en xml (étape 7), soit sous la forme d'une base de données (étape 2) regroupant toutes les informations générées par le système d'annotation (type de l'indice sémantique, portée de l'indice, règle utilisée pour effectuer le découpage, …). Cette base de données est directement liée (étape 3) à l'application qui permet aux experts d'évaluer les annotations (étape 4). Au fil de l'évaluation, les données sont stockées dans la base de données (étape 5). Puis, en récupérant les données de l'évaluation, une phase d'amélioration du système débute. Une fois,

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801254


que le système a été modifié, il est possible soit de répéter à nouveau les étapes 2 à 6 pour évaluer à nouveau et éventuellement modifier encore le système d'annotation ou alors - si la qualité des annotations produites est satisfaisante - exporter le corpus annoté au format XML (étape 7). Ce corpus peut alors servir de corpus référence et/ou être directement exploité par exemple par un système de recherche d'informationrecherche d'information comme nous allons le montrer dans la partie suivante. Ce processus permet également de mettre de côté des annotations qui sont erronées à cause d'une erreur dans l'analyse syntaxique de départ et qui ne sont donc pas corrigeables directement par le système d'annotation. Une fois repérées, ces annotations erronées peuvent être soit corrigées manuellement, soit retirées du corpus.

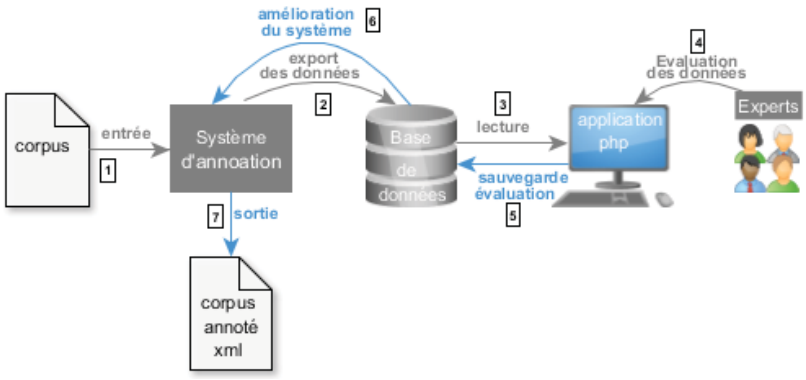


Figure 3. Processus d'amélioration du système d'annotation automatique

# 4 Utilisation du système d'annotation dans un système de recherche d'information pour des journalistes

## 4.1 Description du système de recherche d'information

Dans ce que nous venons de décrire, l'accent a été mis sur la constitution d'un système permettant d'annoter automatiquement dans des textes les segments textuels en fonction des variations de prises en charge énonciative et modale identifiées, ceci sur la base du repérage de certains indices de surface et du calcul de leurs dépendants syntaxiques. Dans la présente section, nous montrons comment les sorties de ce système peuvent être utilisées dans un système de recherche d'information. Ce dernier a pour utilisateurs des journalistes qui cherchent à construire des chronologies évenementielles en tenant compte de critères énonciatifs (par qui est énoncée une information ?), modaux (l'information est-elle présentée comme certaine ?) et temporels (l'information est-elle présentée comme à venir ou passée ?). Il prend place dans le cadre du projet ChronoLines.

Un des objectifs de ce projet dans lequel s'inscrit notre travail est de fournir aux journalistes de l'AFP un outil permettant de construire de façon semi-automatique des objets actuellement réalisées à la main au sein de l'agence etappelés « Chronologies Evènementielles » à partir des milliers de dépêches que cette agence détient. Une chronologie évènementielle représente la succession temporelle des évènements importants (ou saillants) associés à une thématique qu'un journaliste choisit d'explorer (par ex., l'affaire du Mediator, le « printemps arabe » entre 2010 et 2011 pour l'Egypte, la chute de Moubarak, *etc.*) (Battistelli et al., 2012 ; Kessler et al. 2012). Dans cette optique, un système de recherche d'information permettant de construire à partir d'une rêquete des chronologies visuelles faisant apparaître différents niveaux d'informations a été mis en place. Dans cette partie nous présentons comment notre corpus annoté est utilisé dans un tel système.

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801254



L'utilisation du système de recherche d'information qui est proposé aux journalistes s'effectue par le biais d'une interface utilisateur permettant de créer une chronologie à partir d'une requête. Dans un premier temps, l'utilisateur effectue une requête (par exemple : « révolution Egypte ») et choisit une fenêtre temporelle pour sa requête (par exemple : du 01/01/2010 au 31/12/2011). Le système de recherche d'information retourne à l'utilisateur une proposition de chronologie correspondant à sa requête. A partir de cette proposition de chronologie, l'utilisateur peut intervenir pour créer sa chronologie finale. Cette phase de construction/validation – qui s'effectue de manière visuelle avec une interface utilisateur (voir Figures 5 à 7) - permet de composer une chronologie en intégrant des filtres tenant compte de la prise en charge énonciative et modale. Nous allons ici uniquement nous focaliser sur cette phase de validation qui intègre l'utilisation des annotations produites par notre système.

## 4.2 Utilisation des annotations au sein du système

A l'heure actuelle, 20000 dépêches de l'AFP - soit plus de 250000 phrases – produites entre le 1er janvier 2010 et le 31décembre 2011 ont été annotées par notre système d'annotation automatique et intégrées au système de recherche d'information du projet. La figure 4 illustre l'annotation d'une dépêche. Les légendes sont les suivantes :

- **Source du discours en gras**
- *Indice d'ouverture en italique, de même couleur que le segment correspondant à sa portée*
- La portée de l'indice est surlignée de la même couleur que son indice
- Expression calendaire en pourpre, soulignée
- Entité nommée en bleu, soulignée

Dans l'exemple présenté sur la figure 4, deux indices d'ouverture ont été repérés par notre système (*regagnera* et *a annoncé*). Le premier indice qui est un verbe conjugué au futur indique que le segment textuel sous sa portée renvoie à une action encore non-réalisée au moment de la prodution de l'énoncé. Le second indice est un verbe de parole, il indique que le segment textuel sous sa portée est un discours attribuable à une source autre que l'auteur de la dépêche. Notre système repère en plus des indices et des segments sous la portée de ceux-ci la source associée aux indices. Dans le cas de l'indice *a annoncé*, la source du discours est *l'agence officielle Mena*. Le corpus annoté intègre également le repérage des expressions calendaires (utilisées par ailleur pour la construction de la chronologie lors de la requête, cf. (Battistelli et al., 2012 ; Kessler et al. 2012)) et des entités nommées (de type Personne, Lieu et Organisation) [4]. L'identification de ces dernières permet comme nous allons le voir de proposer une liste de sources possibles lors de la validation des chronologies (voir Figure 5).


Figure 4. Exemple de dépêche annotée

Les annotations produites par notre système sont utilisées au cours d'une phase de validation manuelle de la chronologie par le journaliste. Au cours de cette étape le journaliste va pouvoir trier dans une chronologie qui a été générée automatiquement les informations qu'il souhaite conserver dans sa chronologie finale. Afin d'effectuer ce filtrage de l'information, une interface visuelle – illustrée par sur

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801254


la Figure 5 – lui permet de décider des types de filtres qu'il souhaite utiliser. Cette étape de filtrage permet au journaliste de ne sélectionner qu'un sous-ensemble des évènements (des contenus propositionnels, dans la terminologie que nous utilisions jusqu'ici) d'une chronologie en fonction de caractéristiques énonciatives, modales ou temporelles sont associées aux évènements et repérées par notre système. Il est par exemple possible de ne sélectionner que les évènements apparaissant dans un contexte de négation ou encore ceux apparaissant dans un contexte du non-réalisé. Cette sélection s'effectue en utilisant les cases à cocher proposées qui correspondent à différents types d'indices de surface repérés par notre système. Les libellés des cases (ou filtres) proposés sont actuellement les suivants :

- Négation : repérage des morphèmes de négation « ne » portant sur un verbe
- Enonciation Secondaire : repérage de l'introduction d'une nouvelle source (par exemple par le biais d'un discours rapporté ou d'une construction prépositionelle en *selon*)
- Modaux : repérage d'indices modaux lexicaux épistémiques (verbes modaux, adverbes ou adjectifs)
- Conditionnel : repérage des morphèmes flexionnels du conditionnel
- Futur : repérage des morphèmes flexionnels du futur
- Condition : repérage des subordonnées de condition

Ces libellés, fortement liés aux différents types de marqueurs linguistiques pris en compte dans notre système, ne sont pas ceux qui seront proposés aux journalistes dans la version finale de l'application de recherche d'information. Ils préfigurent ainsi seulement pour l'instant les grands types de niveaux d'information que nous souhaitons pouvoir proposer à un journaliste lors de la construction de sa chronologie évènementielle. En l'état actuel du développement du système de recherche et de visualisation de l'information destiné *in fine* aux journalistes, il est permis en tous les cas aux linguistes impliqués dans le projet de tester l'intercation entre les différents filtres envisagés pour ensuite proposer un jeu d'étiquettes de filtres moins directement liées aux catégories d'indices repérés..

Il devient donc possible de sélectionner par exemple des contenus propositionnels en fonction de la source les prenant en charge. Pour cela, il faut tout d'abord choisir le filtre *Source Sec.*, afin de sélectionner tous les segments ayant été introduits par une source secondaire, c'est-à-dire une source autre que le journaliste rédacteur de la dépêche. Puis, le système répertorie toutes les entités nommées ayant été trouvées dans un segment annoté comme étant une source et en propose une liste à l'utilisateur qui peut sélectionner les sources qu'il souhaite retenir dans sa chronologie. Ainsi, il est possible de comparer les événements d'une même chronologie en fonction de la source qui les a mentionnés dans son discours.

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801254





Figure 5. Exemples de sélection de filtres

La figure 5 présente la phase de construction d'une chronologie se composant de trois lignes temporelles concernant la Tunisie entre 2010 et 2011. Dans cette chronologie, l'utilisateur souhaite comparer les événements cités par les premiers ministres de différents états, les événements cités par la presse et les événements cités par des partis politiques. Pour cela, il crée trois lignes temporelles et sélectionne pour chacune d'elles la liste des sources qu'il souhaite voir apparaitre. Une fois la sélection validée, il peut visualiser les événements sur trois lignes temporelles distinctes (voit Figure 6).

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801254


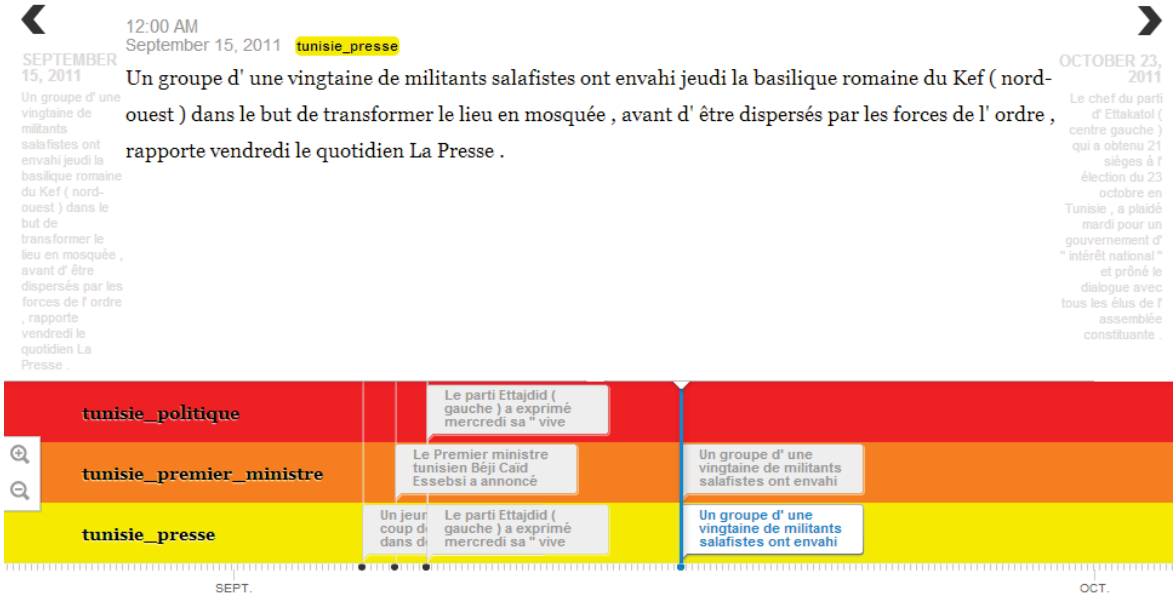


Figure 6. Chronologie avec choix des sources

En suivant le même mode opératoire, il est possible de créer des filtres sans tenir compte de la source. Par exemple, celle présentée dans la figure 7 distingue d'un côté des contenus niés et de l'autre des contenus non-réalisés. La sélection des ces derniers se fait en combinant plusieurs filtres. Pour la création de cette chronologie, ce sont les filtres *Futur*, *Conditionnel* et *Condition* qui ont été utilisés, l'objectif étant de comparer des contenus non encore réalisés- puisqu'ils sont au moment où ils sont énoncés projetés comme pouvant se réaliser de façon plus ou moins certaine dans le futur – avec des événements qui sont niés et qui donc sont donc présentés comme ne pouvant mener à une réalisation.

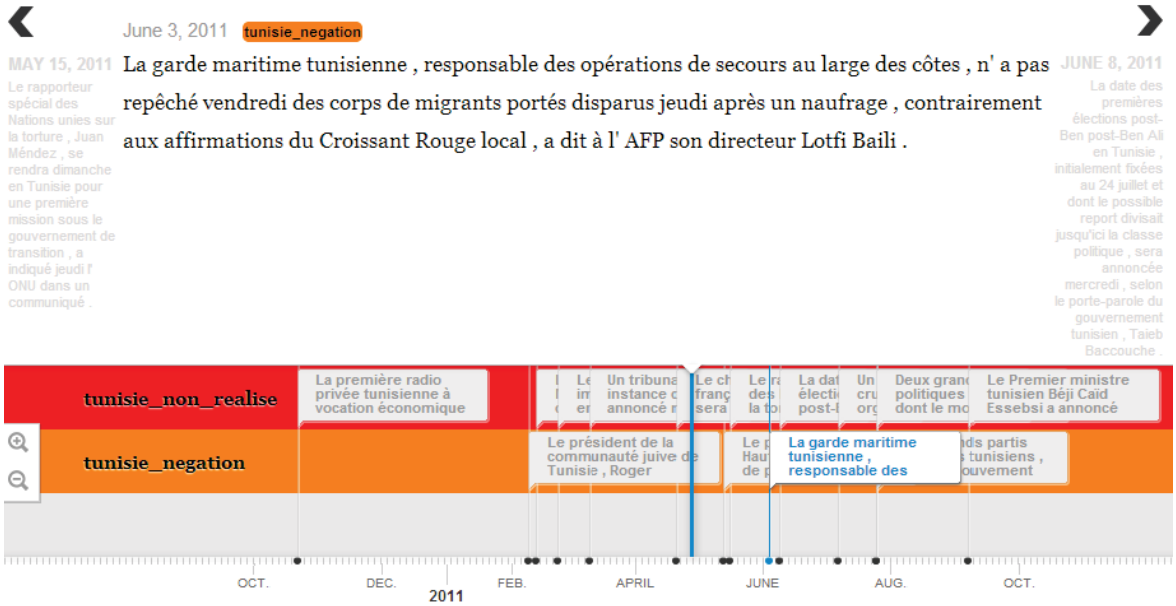


Figure 7. Chronologie avec filtres « modaux » uniquement

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801254



L'évenement du 3 Juin 2011 apparaissent au premier plan sur la chronologie présentée sur la Figure 7 fait partie de la chronologie s'intéressant aux contenus niés et on peut voir que le contenu qui a été repéré est *la garde maritime tunisienne […]* ***n'a pas repêché*** *vendredi de corsps de migrants […]*.

Ainsi, si cette interface est développée au demeurant pour répondre à un besoin émanant des journalistes de l'AFP, elle peut également être utile aux linguistes.

## 5 Conclusion

Dans cet article, nous avons présenté dans un premier temps notre méthodologie d'analyse et d'annotation automatique du phénomène de prise en charge énonciative et modale dans un corpus de dépeches de presse en français. Le schéma d'annotation que nous proposons est basé sur la détection d'indices d'ouverture et des segments textuels correspondant à leur portée. Les résultats de l'évaluation présentée montrent que la tâche la plus complexe pour le système d'annotation n'est pas de trouver les indices mais de délimiter leurs portées et de définir la façon d'évaluer les frontières des segments en particulier dans les cas où l'interprétation n'est pas univoque. Nous avons aussi montré l'intérêt d'utiliser un outil permettant à des experts linguistes d'évaluer simplement et rapidement ce type de corpus annoté. Dans un second temps, nous avons présenté comment notre corpus annoté pouvait être utilisé au sein d'un système de recherche d'information répondant à un besoin émanant des journalistes de l'AFP. Ainsi, ce système permet de créer des chronologies en tenant distinguant des contenus propostionnels avérés, incertains ou niés et également de sélectionner des contenus propositionnels en fonction de la source les ayant énoncés. A l'heure actuelle, nous travaillons à la définition d'un jeu d'étiquettes sémantiques rendant compte des différentes combinaisons possibles de filtres (actuellement, rappelons-le, essentiellement basés sur des classes d'indices de surface). Notre corpus annoté sera mis à disposition de la communauté dès la fin du projet.

## Remerciements

Nous tenons à remercier R. Sauri qui a mis à notre disposition le code php lui servant à gérer l'annotation de son corpus et dont nous nous sommes inspirés pour notre interface d'évaluation, A. Guha qui a adapté cette interface à notre problématique et A. Fanet pour le développement de l'interface permettant de générer les chronologies.

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801254


## Références bibliographiques

Alonso, O., Gertz, M. Et Baeza-Yates, R. (2009). Clustering and exploring search results using timeline constructions. *In Proceedings of the 18th ACM conference on Information and knowledge management*, 97-106.

Bally, C. (1932*). Linguistique générale et Linguistique française*. Paris : Leroux, 2éd. (1944), Berne.

Battistelli D., Damiani M. (2013). Analyzing modal and enunciative discursive heterogeneity: how to combine semantic resources and a syntactic parser analysis, *In Proceedings of s WAMM (Workshop on Annotation of Modal Meaning in Natural Language), held in conjunction with IWCS'13 (10th International Conference on Computational Semantics)*.

Battistelli D., Charnois T., Minel J.-L;, Teissèdre C. (2013). Detecting Salient Events in Large Corpora by a Combination of NLP and Data Mining Techniques, *In Actes Cicling'13 ( International Conference on Intelligent Text Processing and Computational Linguistics )*.

Benveniste, E. (1966). *Problèmes de linguistique générale*, 1, Paris : Gallimard.

Bybee, J. L., Perkins, R., Et Pagliuca, W. (1994). *The evolution of grammar: tense, aspect and modality in the language of the world*. Chicago: University of Chicago Press.

Charolles M. (1997). L'encadrement du discours – univers, champs, domaines et espace, *Cahiers de recherche linguistique*, *6*, 1-73.

Culioli, A. (1973). Sur quelques contradictions en linguistique. *Communications*, *20(1)*, 83-91.

Damiani M., Battistelli D. (2013). Enunciative and modal variations in newswiretexts in French: From guideline to automatic annotation, In *Actes LAW The 7th Linguistic Annotation Workshop & Interoperability with Discourse), held in conjunction with ACL 2013*.

De La Clergerie E., Sagot B., Nicolas L. Et Guenot M. (2009). FRMG: évolutions d'un analyseur syntaxique TAG du français. *In Journée ATALA "Quels analyseurs syntaxiques pour le français ?"*.

Dendale, P. Et Coltier, D. (Eds). (2011). *La prise en charge énonciative. Études théoriques et empiriques*. Bruxelles : De Boeck/ Duculot.

Farkas, R., Vincze, V., Mora, G., Csirik, J., Et Szarvas, G. (2010). The CoNLL-2010 shared task: learning to detect hedges and their scope in natural language text. *In Proceedings of the Fourteenth Conference on Computational Natural Language Learning---Shared Task*, 1-12.

Gosselin, L. (2010). *Les modalités en français: La validation des représentations*. Amsterdam : Rodopi.

Kahane S. (2001). Grammaires de dépendance formelles et théorie Sens-Texte, Tutoriel, *Actes TALN'2001*, 17-76.

Kessler R., Tannier X., Hagège C., Moriceau V., Bittar A. (2012). « Finding Salient Dates for Building Thematic Timelines *», in Proceedings of the 50th Annual Meeting of the Association for Compatutional Linguistics (ACL 2012)*.

Kilicoglu, H., Et Bergler, S. (2010). A high-precision approach to detecting hedges and their scopes. *In Proceedings of the Fourteenth Conference on Computational Natural Language Learning---Shared Task* , 70-77.

Kilicoglu, H. (2012*). Embedding Predications*. PhD thesis, Concordia University.

Le Querler N. (2004). Les modalités en français, *Revue belge de philologie et d'histoire*, *82 fasc. 3*, 643-656.

Nuyts, J. (2006). Modality: Overview and linguistic issues. *In W. J. Frawley (Éd.), The Expression of Modality*. Berlin: De Gruyter Mouton.

Palmer, F. (2001). *Mood and Modality*. Cambridge : Cambridge University Press.

Sauri, R. Et Pustejovsky, J. (2012). Are You Sure That This Happened? Assessing the Factuality Degree of Events in Text. *Computational Linguistics*, *38(2)*, 261-299.

Vincze, V., Szarvas, G., Farkas, R., Móra, G., & Csirik, J. (2008). The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC bioinformatics*, *9(Suppl 11)*, S9.

Wilson, T. Et Wiebe, J. (2005). Annotating Attributions and Private States. *In ACL 2005 Workshop: Frontiers in Corpus Annotation II: Pie in the Sky*, 53-60.

SHS Web of Conferences 8 (2014)
DOI 10.1051/shsconf/20140801254




[1] http://www.chronolines.fr/

[2] Cette notion est convoquée lorsqu'une nouvelle source d'information est introduite dans le discours (à travers un discours rapporté, par ouï-dire, par déduction, ...).

[3] On trouve la même remarque dans (Kilicoglu 2012, p121).

[4] L'annotation des entités nommées est fournie par Xerox