Corpus LIMAH

Rémi Bois

June 27, 2016

1 Introduction

Ce rapport décrit le corpus utilisé au sein du projet Linking Media in Acceptable Hypergraphs[1] (LIMAH).

Le corpus se compose de documents issus de sources journalistiques. Son but est de regrouper des documents de modalités différentes (audio, vidéo, écrit) ainsi que des types de discours différents (articles de fond, brèves, tweets, interviews, ...). Certains documents sont également accompagnés de commentaires utilisateurs associés (commentaires d'article de presse, tweets, ...).

Tous les documents ont été récupérés entre le 20 Mai 2015 et le 8 Juin 2015, via des flux RSS. Les documents complets, qu'ils soient au format HTML, dans un format audio, dans un format vidéo, ou dans un autre format (e.g. json pour Twitter) ont été récupérés et stockés.

2 Composition du corpus

Nous décrivons ici les différents documents qui composent le corpus. Nous présentons successivement les statistiques du corpus modalité par modalité, en commençant par les documents web (articles de presse et blogs), les documents audio (podcasts radio), les documents vidéo (émissions et journaux télévisés), et enfin les données issues de réseaux sociaux. Pour chaque section, des informations statistiques sont fournies. Le tableau 1 indique le nombre de documents présents pour chaque catégorie.

[1] projet LabEx Cominlabs ANR-10-LABX-07-01 : `http://limah.irisa.fr/`

Type	Nombre de documents
Presse	4966
Radio	1556
Video	290

Table 1: Nombre de documents par type

2.1 Documents web

2.1.1 Nombre de documents et métadonnées

La première catégorie de documents récupérée comprend les pages webs. Ces pages comprennent les articles de journaux et articles de blogs. La liste des sites et leurs urls sont indiqués tableau 2. Si les flux RSS sont une porte d'entrée intéressante, il est nécessaire de récupérer les pages web entièrement afin de bénéficier des commentaires des utilisateurs, de la mise en page, et de métadonnées supplémentaires.

Les billets de blog permettent d'exposer davantage de déclarations d'opinions et proposent parfois de mettre en lumière certains aspects de l'actualité.

Source	Type de la source	Nombre de documents
Le Monde	Presse	1802
Le Point	Presse	1029
Le Figaro	Presse	812
Libération	Presse	683
Huffington Post	Presse	640
Blogs le Monde	Blog	137
Blogs le Figaro	Blog	23

Table 2: Documents web

Les métadonnées extraites pour chacun de ces documents sont les suivantes :

- Titre
- Texte (contenu principal)
- HTML du contenu principal
- Date de publication
- Url
- Source (e.g. Le Monde)
- Catégorie (blog ou presse)

- Image d'illustration
- Auteur (nom du journaliste quand présent)
- Description (texte d'introduction quand présent)

Les informations supplémentaires suivantes ont été extraites :

- Découpage en phrases et tokens
- Stemming et Part-Of-Speech tagging
- Entités nommées (personnes, lieux, ...)
- Liens hypertextes présents dans le texte de l'article
- Mots clefs

2.1.2 Taille des documents

Quelques statistiques supplémentaires ont été calculées tableau 3. La figure 2 donne une représentation visuelle de ces statistiques.

Source	Nombre de phrases (moy)	Nombre de mots (moy)
Le Monde	21.7	651
Le Point	19.6	592
Le Figaro	7.4	259
Libération	28.4	792
Huffington Post	21.0	615
Blogs le Monde	30.6	884
Blogs le Figaro	26.8	611
Moyenne pondérée	20.1	597
Médiane	15	477

Table 3: Statistiques sur les documents web

On remarque une grande diversité dans la taille des documents, avec des articles beaucoup plus courts pour le Figaro. Un exemple d'article du Figaro est montré figure 1. Le format de l'article est recréé grâce aux différentes métadonnées récupérées, ainsi que grâce au HTML du contenu principal. Les articles de blogs sont en moyenne plus longs que les articles classiques du journal qui les publie.

FIFA: Vladimir Poutine félicite Sepp Blatter

Le président russe Vladimir Poutine a adressé un télégramme de félicitations au président de la Fédération internationale de football (Fifa), Sepp Blatter, pour le féliciter à l'occasion de sa réélection.

"Le chef de l'Etat russe a dit son espoir que l'expérience, le professionnalisme et la haute autorité dont il jouit aideront (Joseph) Blatter à l'avenir à encourager le développement du football à travers le monde", a déclaré le Kremlin dans un communiqué. La Russie souhaite coopérer avec la Fifa de manière générale, et tout particulièrement en vue de préparer la phase finale de la Coupe du monde 2018, qu'elle organisera.

Blatter a été réélu vendredi président de la Fifa après le retrait de son rival, le prince jordanien Ali ben al Hussein, du second tour de scrutin auquel le patron du football mondial avait été contraint. Ignorant les appels à la démission qui s'étaient multipliés ces derniers jours après les scandales de corruption à l'échelle planétaire affectant l'organisation, Joseph "Sepp" Blatter, 79 ans, a tenu tête et obtenu le droit d'assurer un cinquième mandat de quatre ans au terme duquel il a promis de laisser "une Fifa plus forte" à son successeur.

LIRE AUSSI :
» Blatter : «Je pardonne à tout le monde mais je n'oublie pas»
» Blatter réélu à la tête de la Fifa

Figure 1: Un exemple d'article du Figaro

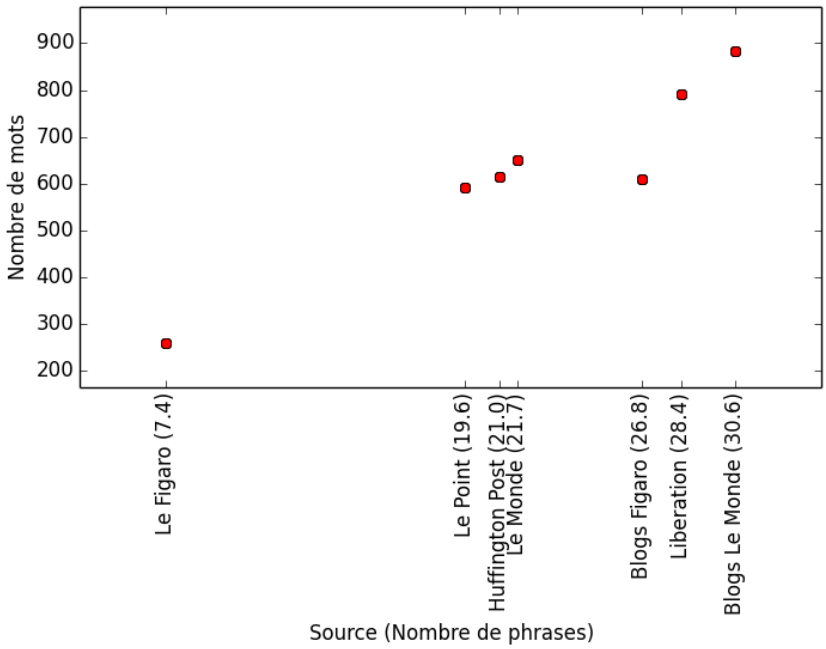

Figure 2: Rapport entre le nombre de mots et le nombre de phrases

2.1.3 Informations les plus fréquemment traitées dans la presse

L'extraction automatique de mots clefs est un bon indicateur pour connaître les principales thématiques abordées dans un corpus. On utilise ici le Term Frequency-Inverse Document Frequency (TF-IDF) afin d'ordonner les mots les plus importants dans chaque document. L'IDF est calculé sur l'ensemble du corpus. Le tableau 4 donne les vingt mots clefs les plus fréquents, ainsi

que le nombre de documents dans lesquels ils apparaissent.

Mot Clef	Fréquence
parti	141
fifa	130
sarkozy	115
roland	90
ministre	90
garros	86
groupe	82
grèce	81
milliards	81
police	80
loi	79
nicolas	79
migrants	78
pays	75
président	75
blatter	75
football	73
dollars	70
film	70
républicains	69

Table 4: Mots clefs les plus fréquents dans les documents de presse

Comme le montre le tableau 4, les thématiques les plus abordées sont le scandale de la FIFA (fifa, blatter, président, football), le renommage du parti de l'UMP en Les Républicains (parti, sarkozy, nicolas, républicains), le tournoi de Roland Garros (roland, garros), les flux migratoires (migrants), et les difficultés économiques de la Grèce (ministre, grèce, milliards, dollars).

2.2 Documents audio

Les émissions de radio d'information ainsi que des chroniques traitant de l'actualité sont ciblées. Certaines émissions impliquent plusieurs orateurs en même temps (eg. Les Grandes Gueules), tandis que d'autres se concentrent sur un présentateur (eg. Journal 13h) ou des interviews de personnalités ou d'auditeurs.

2.2.1 Nombre de documents et métadonnées

Les podcasts récupérés sont décrits dans le tableau 5. Chaque document a été transcrit automatiquement et segmenté automatiquement. Le nombre de documents apparaissant dans le tableau 5 représente ce nombre de segments, et non le nombre d'émissions distinctes.

Emission	Source	Genre	Nombre de documents
Bourdin Direct	BFM	Public	556
Chroniques	France Inter	Chronique	392
Les Grandes Gueules	BFM	Public	283
Divers	France Inter	Emission	144
Carrément Brunet	BFM	Public	50
Journal de 8h	France Culture	Journal	67
Journal de 13h	France Inter	Journal	46

Table 5: Documents audio

Les métadonnées extraites pour chacun de ces documents sont les suivantes :

- Nom de l'émission
- Transcription automatique
- Segmentation automatique
- Date de publication
- Url du podcast
- Source (e.g. RMC)

Les informations supplémentaires suivantes ont été extraites :

- Découpage en psoeudo-phrases et tokens
- Stemming et Part-Of-Speech tagging
- Entités nommées (personnes, lieux, ...)
- Mots clefs

2.2.2 Taille des documents

Le tableau 6 fournit quelques statistiques sur la taille des différents segments.

Source	Nombre de phrases (moy)	Nombre de mots (moy)
Bourdin Direct	55.7	835.2
Chroniques	36.8	516.6
Les Grandes Gueules	72.9	1154.6
Divers	49.5	786.5
Carrément Brunet	115.1	1446.9
Journal de 8h	20.6	329.7
Journal de 13h	31.1	407.2
Moyenne	52.9	789.7
Médiane	49.5	697.4

Table 6: Statistiques sur les documents audio

2.2.3 Informations les plus fréquemment traitées à la radio

Comme pour la presse, les mots clefs ont été extraits automatiquement des documents radio. Le tableau 7 donne les vingt mots clefs les plus fréquents ainsi que le nombre de documents dans lesquels ils apparaissent.

On peut a première vue s'étonner de la présence du terme "euh" en première position des mots-clefs. En effet, "euh" étant fréquemment attendu, son Inverse Document Frequency (IDF) devrait être très faible. En fait, cet IDF est calculé sur l'ensemble du corpus, y compris sur les documents de presse écrite, qui ne comportent logiquement aucune fois le terme "euh". De plus, ce terme est très peu fréquent sur des émissions de radio de type journal d'information, et au contraire extrêmement fréquent lors d'émissions faisant intervenir des auditeurs (*e.g.* Bourdin Direct). Ce terme apparaît donc comme fortement discriminant pour certains documents, et apparaît donc en tête de liste. Le même raisonnement peut être appliqué à d'autres termes tels que "heures" ("Il est huit heures"), "merci", "rmc", ... Il apparaît que les mots clefs extraits des documents audio sont moins exploitables dû à ce bruit. On repère tout de même des références aux thématiques abordées dans la presse (FIFA, Les Républicains).

Mot Clef	Fréquence
euh	255
heures	109
gens	91
hui	76
parti	68
merci	60
fifa	60
france	59
pays	55
jean	52
rmc	48
matin	44
travail	40
question	39
gauche	38
nicolas	38
droite	37
problème	36
discours	36
républicains	35

Table 7: Mots clefs les plus fréquents à la radio

2.3 Documents vidéos

2.3.1 Nombre de documents et métadonnées

Des émissions quotidiennes ou hebdomadaires ont été récupérées, parmi lesquelles des journaux télévisés, des débats politiques, des magazines d'information. La liste des sources utilisées et des quantités récupérées est disponible dans le tableau 8. Comme pour les radios, le nombre de documents correspond au nombre de segments après segmentation automatique.

D'autres émissions ont été récupérées mais n'ont pas été incluses dans le corpus. En effet, celles-ci étaient très irrégulières (*e.g.* les émissions hebdomadaires) et/ou ont été récupérées de façon incomplète.

Emission	Source	Type	Nombre de Documents
JT 20h	France 2	Journal	160
JT 13H	France 2	Journal	41
C dans l'air	France 5	Actu Débats	37
C a vous	France 5	Actu Divertissement	35
JT 06H	France 2	Journal	17

Table 8: Documents vidéos

Les métadonnées extraites pour chacun de ces documents sont les suivantes :

- Nom de l'émission
- Transcription automatique
- Segmentation automatique
- Date de publication
- Url de la vidéo
- Source (e.g. France 2)

Les informations supplémentaires suivantes ont été extraites :

- Découpage en psoeudo-phrases et tokens
- Stemming et Part-Of-Speech tagging
- Entités nommées (personnes, lieux, ...)
- Mots clefs

2.3.2 Taille des documents

Le tableau 9 récapitule les différentes tailles des documents vidéos.

Emission	Nombre de phrases	Nombre de mots
JT 20h	66.3	628
JT 13H	50.9	515
C dans l'air	108.2	1230
C a vous	49	492.5
JT 06H	11.2	127
Moyenne	64	643
Médiane	51	515

Table 9: Statistiques sur les documents vidéos

Le rapport entre nombre de phrases et nombre de mots est constant. Cela peut être attribué à deux facteurs : les émissions traitées sont majoritairement très préparées et avec prompteur (*e.g.* journaux télévisés), et évitent donc les phrases très longues régulièrement présentes dans des émissions de type radio. Seule l'émission C dans l'air possède un nombre de mots par phrase plus important, probablement dû au fait que l'émission consiste à poser des questions à des experts qui ne sont pas soumis à un prompteur.

2.3.3 Informations les plus fréquemment traitées à la télévision

Une fois encore, les mots clefs ont été extraits automatiquement des documents vidéos. Le tableau 10 donne les vingt mots clefs les plus fréquents ainsi que le nombre de documents dans lesquels ils apparaissent.

Mot Clef	Fréquence
euh	45
ans	16
france	12
soir	12
ville	10
gens	10
coalition	9
élèves	8
mois	8
fifa	8
président	8
réforme	7
prix	7
exposition	7
euros	7
année	7
armée	7
produits	6
françois	6
décret	6

Table 10: Mots clefs les plus fréquents à la télévision

Une fois encore, le terme "euh" est discriminant pour certaines émissions. Il est ainsi rare dans les journaux télévisés, mais très présent dans les émissions plus longues basées sur des questions-réponses telles que C dans l'air.

Certaines thématiques réapparaissent telles que le scandale de la FIFA, et de nouvelles émergent telles une mention des “élèves” et de la “réforme” qui correspondent à la réforme de l'enseignement au collège et à la consultation des enseignants sur les nouveaux programmes qui a eu lieu en mai 2015.

2.4 Réseaux sociaux et commentaires utilisateurs

Les réseaux sociaux sont un lieu de prolongement de l'information. Les consommateurs de médias y échangent autour de l'actualité, dans des formes bien plus diversifiées que les médias que ne le sont les sources de leurs échanges.

Nous avons récupéré tous les tweets publiés sur les comptes de journaux présents dans le corpus, ainsi que tous les tweets faisant mention d'un journal sur la période visée. Ces milliers de tweets ont ensuite été filtrés pour ne récupérer que ceux qui font une mention explicite (via l'url) d'un article de presse faisant partie de notre corpus. 15940 tweets ont été récupérés de cette manière. Les mots clefs utilisés pour récupérés ces tweets sont indiqués table 11. Les tweets ayant une taille limitée (140 caractères), on obtient sans surpise uniquement 1.3 phrase par tweet en moyenne.

Mots Clefs
lemonde
lepoint
huffingtonpost
rmc
radiofrance
francetv
arte
lcp
afp

Table 11: Recherche Twitter

Les tweets ont été stockés au format json et contiennent donc la totalité des informations récupérables : auteur, date de publication, nombre de favoris et de retweets au moment de la récupération, urls citées, géolocalisation, etc... [2]

[2] pour une liste exhaustive, voir `https://dev.twitter.com/overview/api/tweets`