1 Présentation des données

1.0.1 pud

Phrases qui proviennent de sites informatifs, id est de journaux ou de Wikipedia

test

Nombre de phrases : 1000

Mot	Apparitions
,	1178
de	1175
.	985
la	675
et	442
le	414
les	399
à	370
l'	328
des	327

FIGURE 1 – Mots les plus utilisés

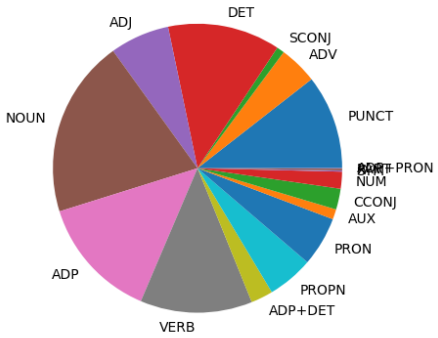

FIGURE 2 – distribution des labels

train

Nombre de phrases : 803

Mot	Apparitions
de	1114
,	1016
.	750
la	683
et	544
l'	486
des	432
à	404
d'	397
le	385

FIGURE 3 – Mots les plus utilisés

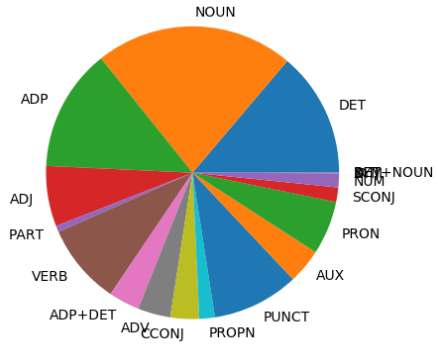

FIGURE 4 – distribution des labels

1.0.2 ftb

Phrases provenant du journal «Le Monde»

test

Nombre de phrases : 2541

Mot	Apparitions
,	4528
de	3593
.	2258
__DIGIT__	2200
la	2016
l'	1350
à	1333
le	1293
des	1255
les	1222

FIGURE 5 – Mots les plus utilisés

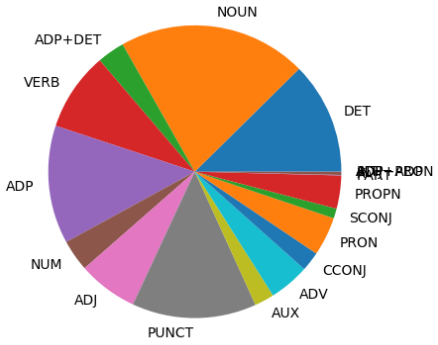

FIGURE 6 – distribution des labels

train

Nombre de phrases : 14759

Mot	Apparitions
,	27318
de	22518
.	13735
__DIGIT__	11104
la	10780
l'	8129
à	7984
le	7255
les	7086
des	7071

FIGURE 7 – Mots les plus utilisés

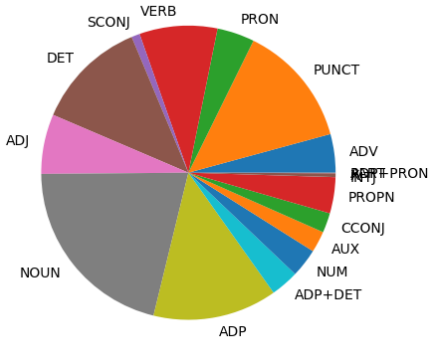

FIGURE 8 – distribution des labels

1.0.3 spoken

Données sur le français oral

test

Nombre de phrases : 726

Mot	Apparitions
de	380
est	262
euh	222
la	219
et	193
que	188
le	181
l'	177
à	169
je	162

FIGURE 9 – Mots les plus utilisés

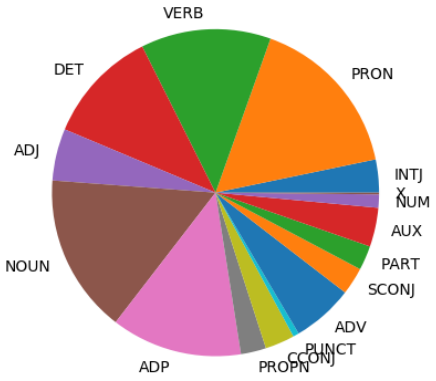

FIGURE 10 – distribution des labels

train

Nombre de phrases : 1153

Mot	Apparitions
de	529
euh	489
est	362
la	339
et	328
le	294
à	282
c'	264
je	223
qui	218

FIGURE 11 – Mots les plus utilisés

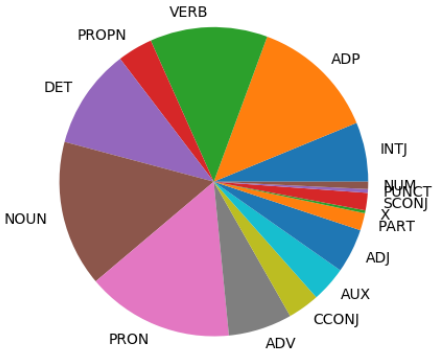

FIGURE 12 – distribution des labels

1.0.4 sequoia

Ensemble de texte, à la base disponible avec des annotations «profondes», ie plus complètes. Néanmoins, ces annotations ont été normalisées par rapport aux autres données

test

Nombre de phrases : 456

Mot	Apparitions
de	463
,	431
.	344
la	216
__DIGIT__	204
l'	186
à	169
et	165
des	152
d'	139

FIGURE 13 – Mots les plus utilisés

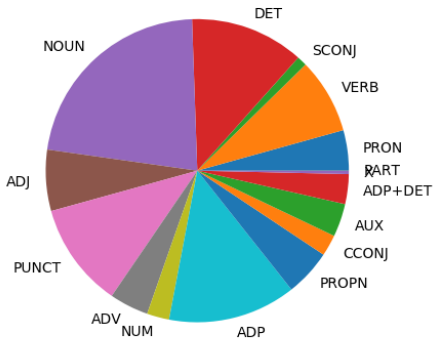

FIGURE 14 – distribution des labels

train

Nombre de phrases : 2231

Mot	Apparitions
de	2435
,	2354
.	1683
la	1176
__DIGIT__	1039
l'	916
à	857
et	845
des	759
le	743

FIGURE 15 – Mots les plus utilisés

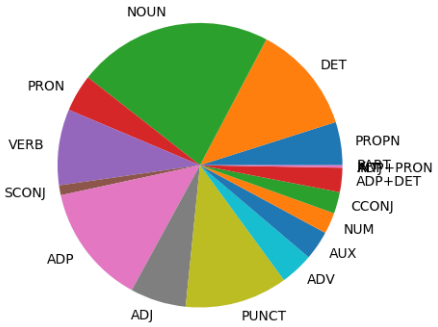

FIGURE 16 – distribution des labels

1.0.5 partut

Ensemble varié de textes, incluant de extraits de conférences comme des extraits de Wikipedia ou des textes légaux

test

Nombre de phrases : 110

Mot	Apparitions
de	120
.	110
,	79
la	69
à	58
l'	56
les	46
et	42
des	39
le	39

FIGURE 17 – Mots les plus utilisés

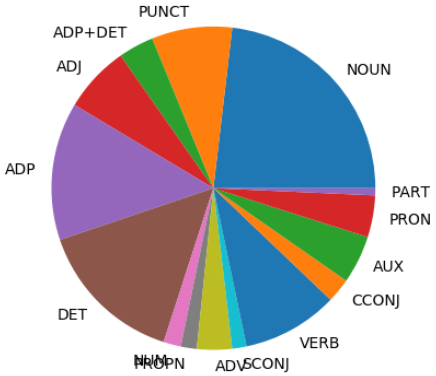

FIGURE 18 – distribution des labels

train

Nombre de phrases : 803

Mot	Apparitions
de	1114
,	1016
.	750
la	683
et	544
l'	486
des	432
à	404
d'	397
le	385

FIGURE 19 – Mots les plus utilisés

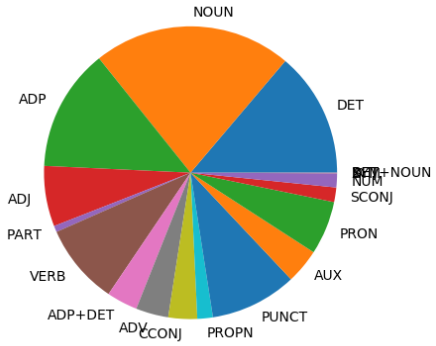

FIGURE 20 – distribution des labels

1.0.6 gsd

Phrases variées

test

Nombre de phrases : 416

Mot	Apparitions
,	489
de	426
.	353
la	222
et	182
l'	171
à	168
__DIGIT__	166
le	155
les	127

FIGURE 21 – Mots les plus utilisés

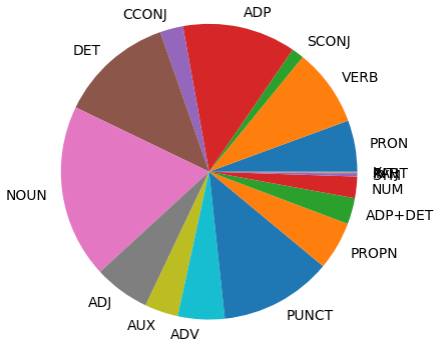

FIGURE 22 – distribution des labels

train

Nombre de phrases : 14450

Mot	Apparitions
de	16964
,	16595
.	13359
la	8676
et	7148
__DIGIT__	7095
le	6389
à	6216
l'	5864
en	4793

FIGURE 23 – Mots les plus utilisés

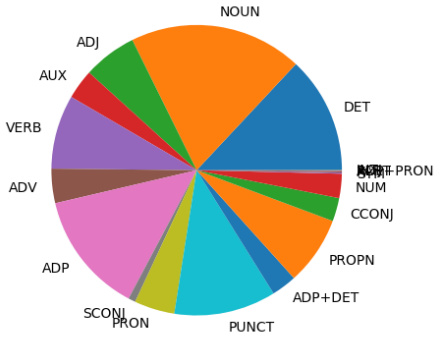

FIGURE 24 – distribution des labels

1.0.7 foot

test

Nombre de phrases : 743

Mot	Apparitions
__MENTION__	549
la	433
__SHARP__	395
Juventus	368
:	338
de	325
!	314
le	257
,	247
__URL__	226

FIGURE 25 – Mots les plus utilisés

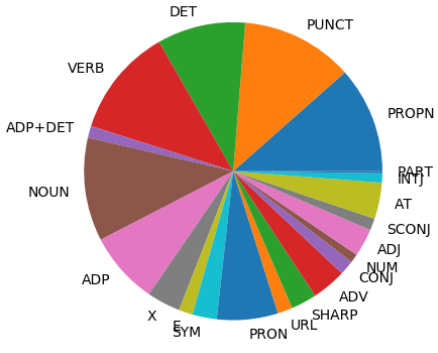

FIGURE 26 – distribution des labels

1.0.8 natdis

Ensemble de tweets

test

Nombre de phrases : 622

Mot	Apparitions
de	994
__URL__	469
__DIGIT__	411
terre	338
__SHARP__	300
tremblement	295
:	285
chaleur	244
au	220
vague	206

FIGURE 27 – Mots les plus utilisés

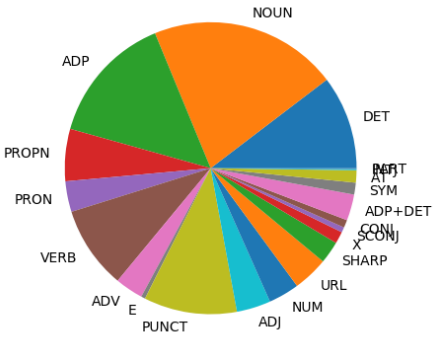

FIGURE 28 – distribution des labels