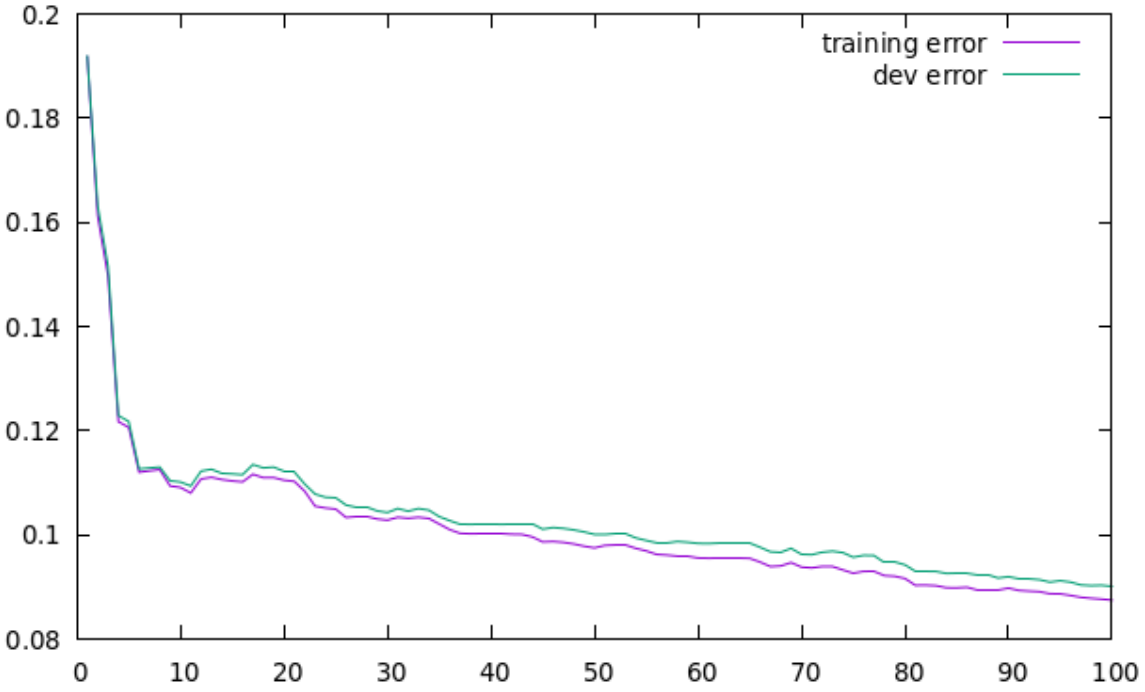
Introduction au Traitement Automatique des Langues
Evaluation détection entités nommées

- Récupération des patrons les plus fréquents :

```
m22008412@V-PJ-47-013:~/Bureau/S2/TAL/TP6$ cat corpus_en_580K.train | python3 extrait_entite_nomme.py | sort | cut -f2,3 | sort -s -t$'\t' -k1,1 | python3 entite_nomme_freq.py | sort -t$'\t' -k2,2 -r -n | python3 delete_last_col.py > patrons.txt
```

Model 1: model avec les 10 premiers patrons les plus fréquents
Les courbes d'apprentissage :



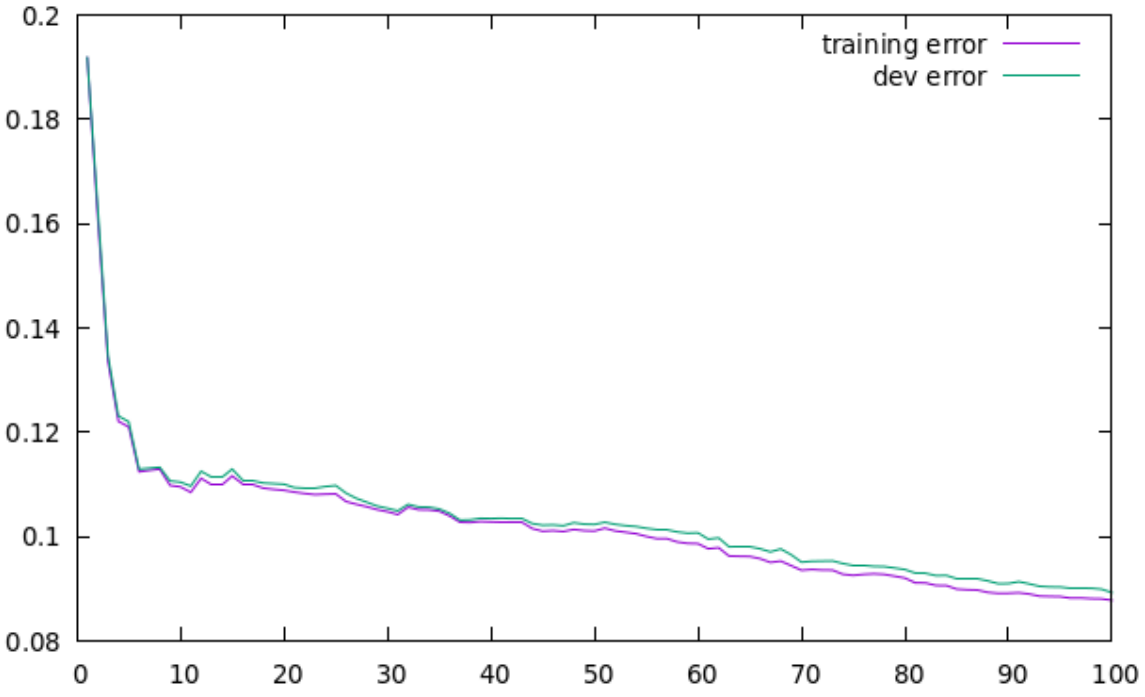

```
m22008412@V-PJ-47-013:~/Bureau/S2/TAL/TP6$ cat patrons.txt | head -n 10 > patron10.txt
m22008412@V-PJ-47-013:~/Bureau/S2/TAL/TP6$ icsiboost-master/src/icsiboost -S class_en_v1 -n 100 | tee class_en_v1.iter

m22008412@V-PJ-47-013:~/Bureau/S2/TAL/TP6$ cat class_en_v1.dev | icsiboost-master/src/icsiboost -S class_en_v1 -C | ./tagg_corpus_icsiboost -names class_en_v1.names -corpus class_en_v1.dev | ./recopie_label_entite_nommes -corpus corpus_dev.txt | ./evalue_entite_nomme
Evaluation stricte (labels et frontieres doivent etre corrects) :
 - macro-F1 : Precision: 49.45 - Rappel: 48.98 - F1: 49.22 - nbref=4896 nbhyp=4768 nbok=3164
 - micro-F1 : Precision: 66.36 - Rappel: 64.62 - F1: 65.48 - nbref=4896 nbhyp=4768 nbok=3164
 Details par label :
   -      geoloc : Precision: 64.13 - Rappel: 66.97 - F1: 65.52 - nbref=1765 nbhyp=1843 nbok=1182
```

```
   -         org : Precision: 56.05 - Rappel: 49.73 - F1: 52.71 - nbref=1508
nbhyp=1338 nbok=750
   -      person : Precision: 77.63 - Rappel: 79.23 - F1: 78.42 - nbref=1555
nbhyp=1587 nbok=1232
   -     product : Precision: 00.00 - Rappel: 00.00 - F1: 00.00 - nbref=68
nbhyp=0 nbok=0

Evaluation detection (uniquement le token de debut et le type de l'entite
doivent etre corrects) :
 - macro-F1 : Precision: 51.30 - Rappel: 50.70 - F1: 51.00 - nbref=4896
nbhyp=4768 nbok=3276
 - micro-F1 : Precision: 68.71 - Rappel: 66.91 - F1: 67.80 - nbref=4896
nbhyp=4768 nbok=3276
 Details par label :
   -      geoloc : Precision: 67.77 - Rappel: 70.76 - F1: 69.24 - nbref=1765
nbhyp=1843 nbok=1249
   -         org : Precision: 61.96 - Rappel: 54.97 - F1: 58.26 - nbref=1508
nbhyp=1338 nbok=829
   -      person : Precision: 75.49 - Rappel: 77.04 - F1: 76.26 - nbref=1555
nbhyp=1587 nbok=1198
   -     product : Precision: 00.00 - Rappel: 00.00 - F1: 00.00 - nbref=68
nbhyp=0 nbok=0
```

Model 2: model avec les 15 premiers patrons les plus fréquents
Les courbes d'apprentissage :



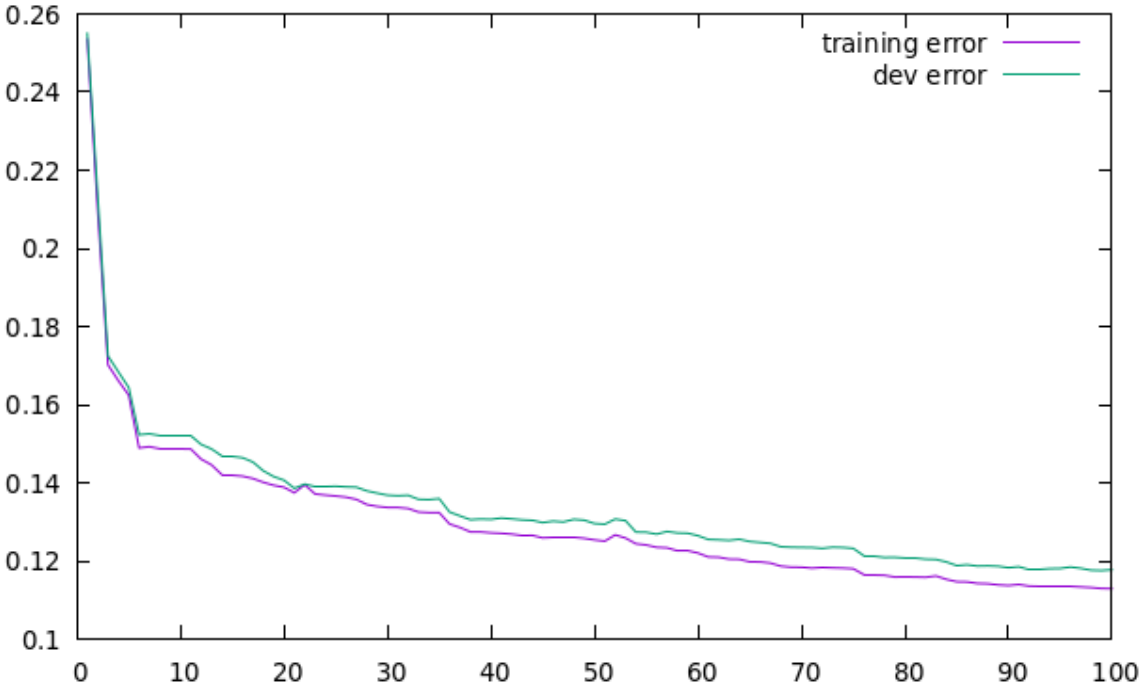

```
m22008412@V-PJ-47-013:~/Bureau/S2/TAL/TP6$ cat class_en_v2.dev |
icsiboost-master/src/icsiboost -S class_en_v2 -C | ./tagg_corpus_icsiboost
-names class_en_v2.names -corpus class_en_v2.dev |
./recopie_label_entite_nommes -corpus corpus_dev.txt | ./evalue_entite_nomme
Evaluation stricte (labels et frontieres doivent etre corrects) :
```

```
 - macro-F1 : Precision: 49.53 - Rappel: 48.82 - F1: 49.17 - nbref=4896
nbhyp=4734 nbok=3154
 - micro-F1 : Precision: 66.62 - Rappel: 64.42 - F1: 65.50 - nbref=4896
nbhyp=4734 nbok=3154
 Details par label :
   -     geoloc : Precision: 64.60 - Rappel: 66.69 - F1: 65.63 - nbref=1765
nbhyp=1822 nbok=1177
   -        org : Precision: 56.45 - Rappel: 48.47 - F1: 52.16 - nbref=1508
nbhyp=1295 nbok=731
   -     person : Precision: 77.06 - Rappel: 80.13 - F1: 78.56 - nbref=1555
nbhyp=1617 nbok=1246
   -    product : Precision: 00.00 - Rappel: 00.00 - F1: 00.00 - nbref=68
nbhyp=0 nbok=0

Evaluation detection (uniquement le token de debut et le type de l'entite
doivent etre corrects) :
 - macro-F1 : Precision: 51.66 - Rappel: 50.75 - F1: 51.20 - nbref=4896
nbhyp=4734 nbok=3279
 - micro-F1 : Precision: 69.26 - Rappel: 66.97 - F1: 68.10 - nbref=4896
nbhyp=4734 nbok=3279
 Details par label :
   -     geoloc : Precision: 68.22 - Rappel: 70.42 - F1: 69.31 - nbref=1765
nbhyp=1822 nbok=1243
   -        org : Precision: 62.86 - Rappel: 53.98 - F1: 58.08 - nbref=1508
nbhyp=1295 nbok=814
   -     person : Precision: 75.57 - Rappel: 78.59 - F1: 77.05 - nbref=1555
nbhyp=1617 nbok=1222
   -    product : Precision: 00.00 - Rappel: 00.00 - F1: 00.00 - nbref=68
nbhyp=0 nbok=0
```

Model 3: model avec les 10 premiers patrons les plus fréquents et les patrons avec les majuscule

Les courbes d’apprentissage :

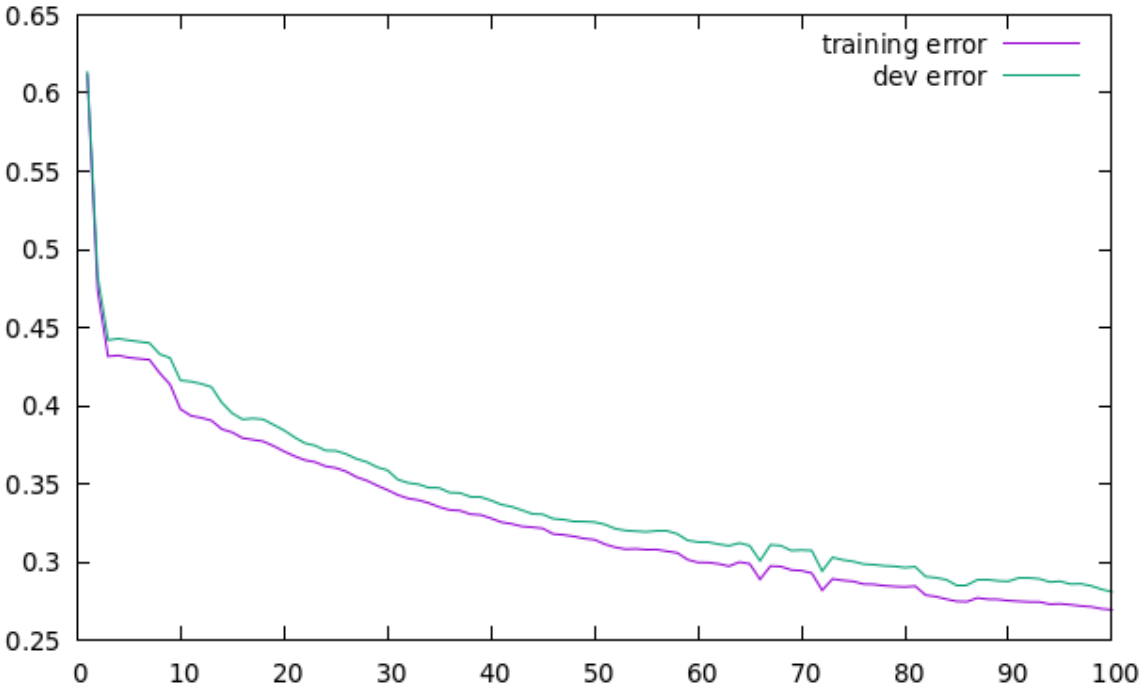



```
m22008412@V-PJ-47-013:~/Bureau/S2/TAL/TP6$ cat class_en_v3.dev |
icsiboost-master/src/icsiboost -S class_en_v3 -C | ./tagg_corpus_icsiboost
-names class_en_v3.names -corpus class_en_v3.dev |
./recopie_label_entite_nommes -corpus corpus_dev.txt | ./evalue_entite_nomme
Evaluation stricte (labels et frontieres doivent etre corrects) :
 - macro-F1 : Precision: 48.72 - Rappel: 50.14 - F1: 49.42 - nbref=4896
nbhyp=4934 nbok=3235
 - micro-F1 : Precision: 65.57 - Rappel: 66.07 - F1: 65.82 - nbref=4896
nbhyp=4934 nbok=3235
 Details par label :
   -     geoloc : Precision: 63.88 - Rappel: 66.74 - F1: 65.28 - nbref=1765
nbhyp=1844 nbok=1178
   -        org : Precision: 57.82 - Rappel: 50.99 - F1: 54.19 - nbref=1508
nbhyp=1330 nbok=769
   -     person : Precision: 73.18 - Rappel: 82.83 - F1: 77.71 - nbref=1555
nbhyp=1760 nbok=1288
   -    product : Precision: 00.00 - Rappel: 00.00 - F1: 00.00 - nbref=68
nbhyp=0 nbok=0

Evaluation detection (uniquement le token de debut et le type de l'entite
doivent etre corrects) :
 - macro-F1 : Precision: 50.18 - Rappel: 51.44 - F1: 50.80 - nbref=4896
nbhyp=4934 nbok=3320
 - micro-F1 : Precision: 67.29 - Rappel: 67.81 - F1: 67.55 - nbref=4896
nbhyp=4934 nbok=3320
 Details par label :
   -     geoloc : Precision: 66.76 - Rappel: 69.75 - F1: 68.22 - nbref=1765
nbhyp=1844 nbok=1231
   -        org : Precision: 62.48 - Rappel: 55.11 - F1: 58.56 - nbref=1508
nbhyp=1330 nbok=831
```

```
   -     person : Precision: 71.48 - Rappel: 80.90 - F1: 75.90 - nbref=1555
nbhyp=1760 nbok=1258
   -    product : Precision: 00.00 - Rappel: 00.00 - F1: 00.00 - nbref=68
nbhyp=0 nbok=0
```

Model 4: model avec les patrons qui on un majuscule
Les courbes d'apprentissage :

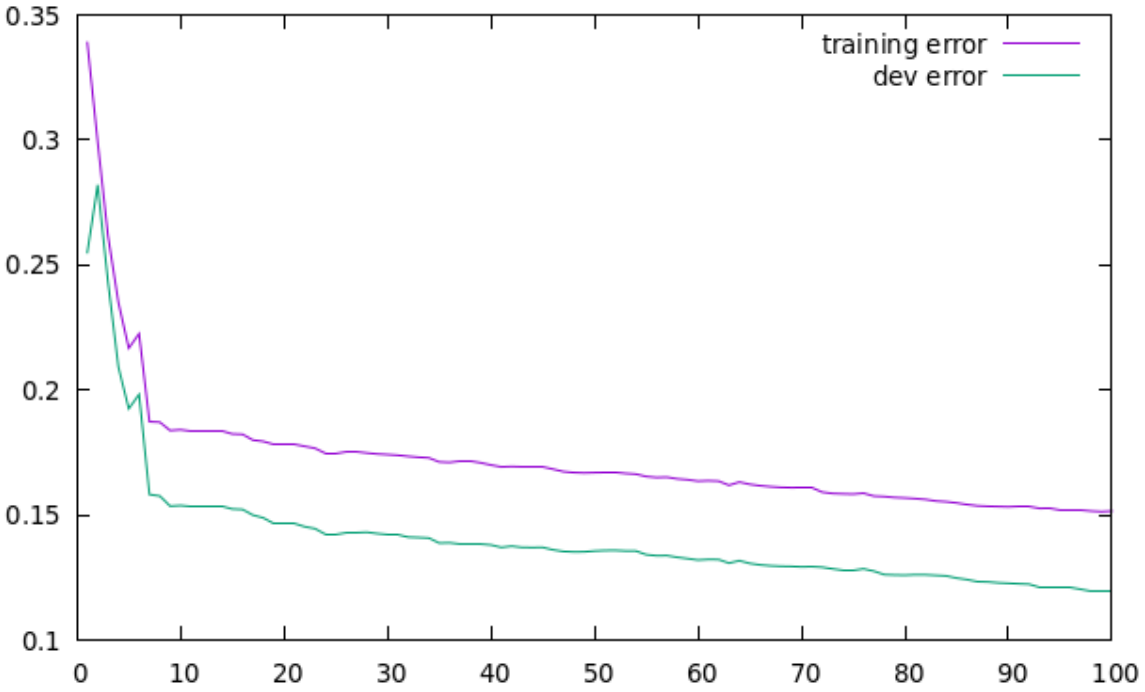



```
m22008412@V-PJ-47-013:~/Bureau/S2/TAL/TP6$ cat class_en_v4.dev |
icsiboost-master/src/icsiboost -S class_en_v4 -C | ./tagg_corpus_icsiboost
-names class_en_v4.names -corpus class_en_v4.dev |
./recopie_label_entite_nommes -corpus corpus_dev.txt | ./evalue_entite_nomme
Evaluation stricte (labels et frontieres doivent etre corrects) :
 - macro-F1 : Precision: 50.10 - Rappel: 50.83 - F1: 50.46 - nbref=4896
nbhyp=4803 nbok=3257
 - micro-F1 : Precision: 67.81 - Rappel: 66.52 - F1: 67.16 - nbref=4896
nbhyp=4803 nbok=3257
 Details par label :
   -     geoloc : Precision: 72.98 - Rappel: 66.57 - F1: 69.63 - nbref=1765
nbhyp=1610 nbok=1175
   -        org : Precision: 50.25 - Rappel: 47.41 - F1: 48.79 - nbref=1508
nbhyp=1423 nbok=715
   -     person : Precision: 77.18 - Rappel: 87.85 - F1: 82.17 - nbref=1555
nbhyp=1770 nbok=1366
   -    product : Precision: 00.00 - Rappel: 01.47 - F1: 00.00 - nbref=68
nbhyp=0 nbok=1

Evaluation detection (uniquement le token de debut et le type de l'entite
doivent etre corrects) :
 - macro-F1 : Precision: 53.47 - Rappel: 53.57 - F1: 53.52 - nbref=4896
nbhyp=4803 nbok=3445
```

```
 - micro-F1 : Precision: 71.73 - Rappel: 70.36 - F1: 71.04 - nbref=4896
nbhyp=4803 nbok=3445
 Details par label :
   -     geoloc : Precision: 73.73 - Rappel: 67.25 - F1: 70.34 - nbref=1765
nbhyp=1610 nbok=1187
   -        org : Precision: 64.09 - Rappel: 60.48 - F1: 62.23 - nbref=1508
nbhyp=1423 nbok=912
   -     person : Precision: 76.05 - Rappel: 86.56 - F1: 80.96 - nbref=1555
nbhyp=1770 nbok=1346
   -    product : Precision: 00.00 - Rappel: 00.00 - F1: 00.00 - nbref=68
nbhyp=0 nbok=0
```

Model 5: model en l'entraînant sur le dictionnaire en plus du corpus.
Les courbes d'apprentissage :



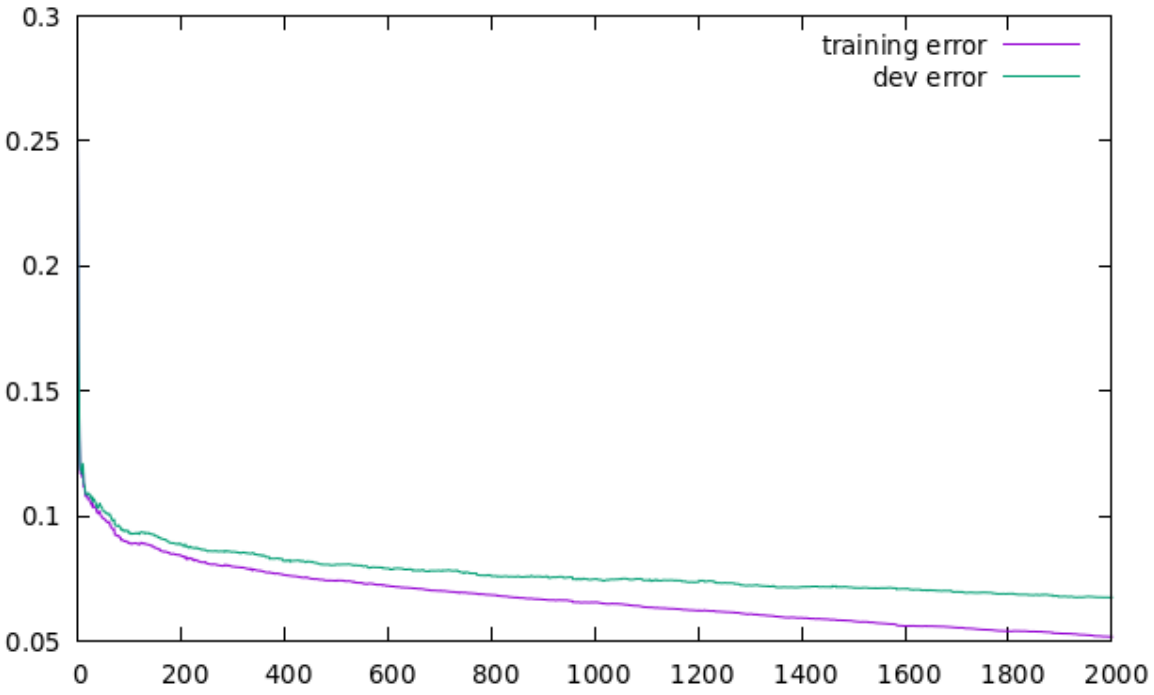

```
m22008412@V-PJ-47-008:~/Bureau/S2/TAL/TP6$ cat class_en_v5.dev |
icsiboost-master/src/icsiboost -S class_en_v5 -C | ./tagg_corpus_icsiboost
-names class_en_v5.names -corpus class_en_v5.dev |
./recopie_label_entite_nommes -corpus corpus_dev.txt | ./evalue_entite_nomme
Evaluation stricte (labels et frontieres doivent etre corrects) :
 - macro-F1 : Precision: 49.30 - Rappel: 49.07 - F1: 49.18 - nbref=4896
nbhyp=4805 nbok=3177
 - micro-F1 : Precision: 66.12 - Rappel: 64.89 - F1: 65.50 - nbref=4896
nbhyp=4805 nbok=3177
 Details par label :
   -     geoloc : Precision: 61.07 - Rappel: 69.07 - F1: 64.82 - nbref=1765
nbhyp=1996 nbok=1219
   -        org : Precision: 61.80 - Rappel: 42.37 - F1: 50.28 - nbref=1508
nbhyp=1034 nbok=639
   -     person : Precision: 74.31 - Rappel: 84.82 - F1: 79.22 - nbref=1555
nbhyp=1775 nbok=1319
```

```
   -     product : Precision: 00.00 - Rappel: 00.00 - F1: 00.00 - nbref=68
nbhyp=0 nbok=0

Evaluation detection (uniquement le token de debut et le type de l'entite
doivent etre corrects) :
 - macro-F1 : Precision: 50.74 - Rappel: 50.15 - F1: 50.44 - nbref=4896
nbhyp=4805 nbok=3249
 - micro-F1 : Precision: 67.62 - Rappel: 66.36 - F1: 66.98 - nbref=4896
nbhyp=4805 nbok=3249
 Details par label :
   -      geoloc : Precision: 63.68 - Rappel: 72.01 - F1: 67.59 - nbref=1765
nbhyp=1996 nbok=1271
   -         org : Precision: 66.73 - Rappel: 45.76 - F1: 54.29 - nbref=1508
nbhyp=1034 nbok=690
   -      person : Precision: 72.56 - Rappel: 82.83 - F1: 77.36 - nbref=1555
nbhyp=1775 nbok=1288
   -     product : Precision: 00.00 - Rappel: 00.00 - F1: 00.00 - nbref=68
nbhyp=0 nbok=0
```

Model 6:

Les courbes d'apprentissage :



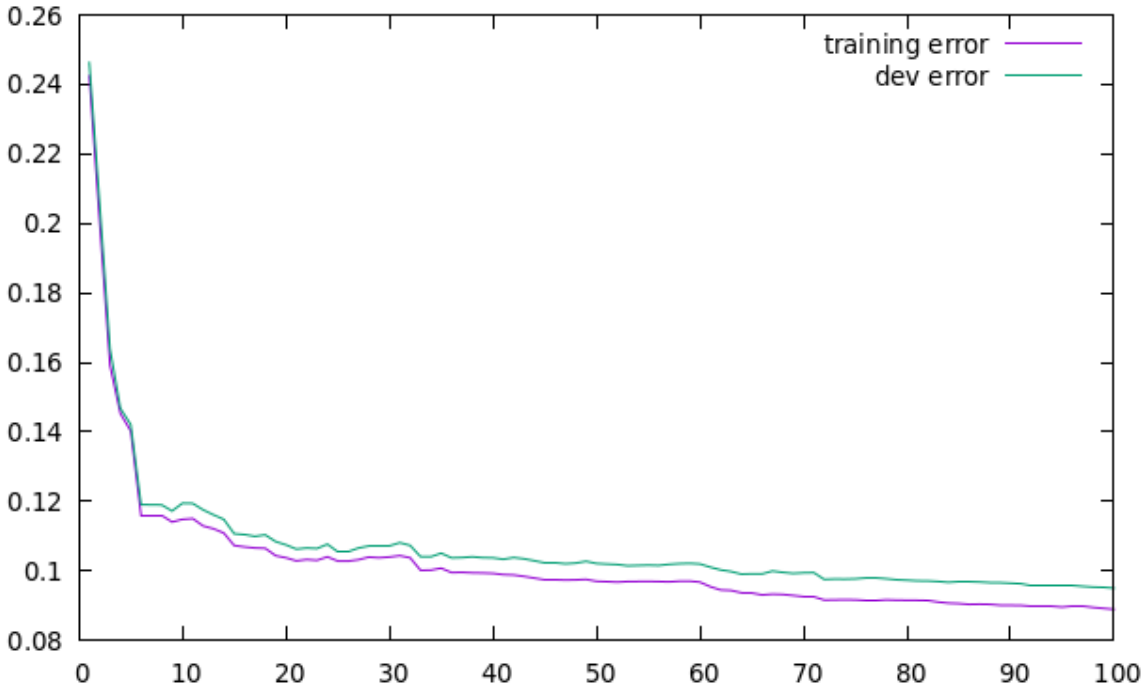

```
m22008412@V-PJ-47-049:~/Bureau/S2/TAL/TP6$ cat class_en_v6.dev |
icsiboost-master/src/icsiboost -S class_en_v6 -C | ./tagg_corpus_icsiboost
-names class_en_v6.names -corpus class_en_v6.dev |
./recopie_label_entite_nommes -corpus corpus_dev.txt | ./evalue_entite_nomme
Evaluation stricte (labels et frontieres doivent etre corrects) :
 - macro-F1 : Precision: 67.69 - Rappel: 61.76 - F1: 64.59 - nbref=4896
nbhyp=4637 nbok=3893
 - micro-F1 : Precision: 83.96 - Rappel: 79.51 - F1: 81.67 - nbref=4896
nbhyp=4637 nbok=3893
```

```
 Details par label :
   -      geoloc : Precision: 83.21 - Rappel: 81.13 - F1: 82.16 - nbref=1765
nbhyp=1721 nbok=1432
   -         org : Precision: 79.18 - Rappel: 66.58 - F1: 72.33 - nbref=1508
nbhyp=1268 nbok=1004
   -      person : Precision: 89.31 - Rappel: 93.44 - F1: 91.33 - nbref=1555
nbhyp=1627 nbok=1453
   -     product : Precision: 19.05 - Rappel: 05.88 - F1: 08.99 - nbref=68
nbhyp=21 nbok=4

Evaluation detection (uniquement le token de debut et le type de l'entite
doivent etre corrects) :
 - macro-F1 : Precision: 86.26 - Rappel: 68.56 - F1: 76.40 - nbref=4896
nbhyp=4637 nbok=4010
 - micro-F1 : Precision: 86.48 - Rappel: 81.90 - F1: 84.13 - nbref=4896
nbhyp=4637 nbok=4010
 Details par label :
   -      geoloc : Precision: 84.43 - Rappel: 82.32 - F1: 83.36 - nbref=1765
nbhyp=1721 nbok=1453
   -         org : Precision: 85.33 - Rappel: 71.75 - F1: 77.95 - nbref=1508
nbhyp=1268 nbok=1082
   -      person : Precision: 89.55 - Rappel: 93.70 - F1: 91.58 - nbref=1555
nbhyp=1627 nbok=1457
   -     product : Precision: 85.71 - Rappel: 26.47 - F1: 40.45 - nbref=68
nbhyp=21 nbok=18
```

Model 7:

Les courbes d’apprentissage :



```
m22008412@V-PJ-47-049:~/Bureau/S2/TAL/TP6$ cat class_en_v7.dev |
icsiboost-master/src/icsiboost -S class_en_v7 -C | ./tagg_corpus_icsiboost
-names class_en_v7.names -corpus class_en_v7.dev |
./recopie_label_entite_nommes -corpus corpus_dev.txt | ./evalue_entite_nomme
Evaluation stricte (labels et frontieres doivent etre corrects) :
 - macro-F1 : Precision: 58.25 - Rappel: 54.73 - F1: 56.44 - nbref=4896
nbhyp=4521 nbok=3535
 - micro-F1 : Precision: 78.19 - Rappel: 72.20 - F1: 75.08 - nbref=4896
nbhyp=4521 nbok=3535
 Details par label :
   -     geoloc : Precision: 75.91 - Rappel: 74.11 - F1: 75.00 - nbref=1765
nbhyp=1723 nbok=1308
   -        org : Precision: 70.33 - Rappel: 53.45 - F1: 60.74 - nbref=1508
nbhyp=1146 nbok=806
   -     person : Precision: 86.75 - Rappel: 91.38 - F1: 89.01 - nbref=1555
nbhyp=1638 nbok=1421
   -    product : Precision: 00.00 - Rappel: 00.00 - F1: 00.00 - nbref=68
nbhyp=14 nbok=0

Evaluation detection (uniquement le token de debut et le type de l'entite
doivent etre corrects) :
 - macro-F1 : Precision: 80.92 - Rappel: 61.32 - F1: 69.77 - nbref=4896
nbhyp=4521 nbok=3710
 - micro-F1 : Precision: 82.06 - Rappel: 75.78 - F1: 78.79 - nbref=4896
nbhyp=4521 nbok=3710
 Details par label :
   -     geoloc : Precision: 80.21 - Rappel: 78.30 - F1: 79.24 - nbref=1765
nbhyp=1723 nbok=1382
   -        org : Precision: 78.10 - Rappel: 59.35 - F1: 67.45 - nbref=1508
nbhyp=1146 nbok=895
   -     person : Precision: 86.81 - Rappel: 91.45 - F1: 89.07 - nbref=1555
nbhyp=1638 nbok=1422
   -    product : Precision: 78.57 - Rappel: 16.18 - F1: 26.83 - nbref=68
nbhyp=14 nbok=11
```