

A. ANDREEWSKY - C. FLUHR

EXPERIENCE DE CONSTITUTION
D'UN PROGRAMME D'APPRENTISSAGE
POUR LE TRAITEMENT AUTOMATIQUE DU LANGAGE

Dans ce travail, on présente une méthode d'apprentissage permettant à partir d'un texte donné de constituer une syntaxe capable de lever les ambiguïtés grammaticales du langage.

Pour cela, on prend un texte *T* que l'on code entièrement en associant à chaque mot une catégorie grammaticale, celle qu'il a dans le texte.

Par exemple:

LA	→	ARTICLE
MONTRE	→	SUBSTANTIF
EN	→	PREPOSITION
OR	→	SUBSTANTIF
DE	→	PREPOSITION
MON	→	ARTICLE POSSESSIF
ONCLE	→	SUBSTANTIF
EST	→	VERBE CONJUGUE
ANCIENNE	→	ADJECTIF

Par ailleurs, lorsque une phrase est fondamentalement ambiguë on doit reproduire la même phrase avec les interprétations possibles.

Par exemple:

Le pouvoir me serait agréable.

On aura:

LE → ARTICLE		LE → PRONOM PREVERBAL
	et	
POUVOIR → SUBSTANTIF		POUVOIR → INFINITIF
etc. ...		

On constitue ensuite un dictionnaire, qui accumule les divers valeurs grammaticales possibles des mots du texte T . Par exemple on pourra avoir ainsi:

LA	(ARTICLE ou SUBSTANTIF ou PRONOM PREVERBAL)
MONTRE	(SUBSTANTIF ou VERBE CONJUGUE)
EN	(PREPOSITION ou PRONOM PREVERBAL)
OR	(SUBSTANTIF ou CONJONCTION)
EST	(VERBE CONJUGUE ou SUBSTANTIF)
etc. ...	

De cette façon on obtient un dictionnaire des mots de T avec leurs diverses valeurs grammaticales possibles dans T , on peut alors rendre le texte T ambigu en remplaçant les valeurs grammaticales de chaque mot dans le texte par les valeurs accumulées. On obtient ainsi un texte T_A . Il est clair que T représente la « vraie résolution » du texte T_A . On fait alors un relevé des configurations binaires et ternaires résolventes. On procède de la façon suivante:

— Soit par exemple $(V_1, \bar{V}_2, V_3)_i^k$ les trois valeurs d'un mot de morphologie M^k et de position i , dans T_A .

— Soit V_i^k sa valeur dans T .

On note $(V_1, \bar{V}_2, V_3)_i^k$ le passage de T_A à T .

Supposons que le mot de position $i+1$ ait la morphologie M^h et les valeurs $(W_1, W_2)_{i+1}^h$ avec la transition $(\bar{W}_1, W_2)_{i+1}^h$ lors du passage de T_A à T .

Alors il est possible de faire la corrélation binaire:

$$(V_1, \bar{V}_2, V_3)_i^k \text{ si } (\bar{W}_1, W_2)_{i+1}^h$$

ce que l'on note:

$$(V_1, \bar{V}_2, V_3)_i^k \wedge (\bar{W}_1, W_2)_{i+1}^h$$

Cela étant, on peut faire un relevé de toutes les transitions des corrélation binaires, quels que soient i, k, h , c'est-à-dire de faire un répertoire de toutes les résolutions associées deux à deux.

L'apprentissage de ces résolutions binaires est relativement rapide puisque des mots de morphologie différente peuvent avoir la même liste de valeurs.

L'association binaire écrite peut à son tour devenir ambiguë, c'est à dire pour $i \neq j$ on aura:

$$(V_1, \bar{V}_2, V_3)_i \wedge (\bar{W}_1, W_2)_{i+1} \quad \text{et} \quad (\bar{V}_1, V_2, V_3)_j \wedge (W_1, \bar{W}_2)_{j+1}$$

Par exemple soit les deux phrases:

F_1 : *Le programme entre dans la machine.*

F_2 : *Il le programme entre midi et deux heures.*

Alors on voit que l'association

$$(\text{ARTICLE, PRONOM PREVERBAL}) \wedge (\text{SUBSTANTIF, VERBE CONJUGUE})$$

a une première résolution dans F_1 et une seconde dans F_2 .

Pour essayer de lever l'ambiguïté de l'association binaire, on peut utiliser les associations ternaires, c'est à dire des expressions de la forme:

$$(\bar{V}_1, V_2, V_3)_i \wedge (W_1, \bar{W}_2)_{i+1} \wedge (Y_1, Y_2, \bar{Y}_3)_{i+2}$$

On recommence ensuite pour les associations ternaires la même procédure que pour les associations binaires.

La méthode de résolution employée ensuite pour l'analyse peut être décrite à l'aide d'un exemple:

Soit la phrase: *Ce programme commence bien.*

Le répertoire des résolutions binaires analyse tout d'abord *Ce*
Ce qui donne les résolutions suivantes:¹

$$\begin{aligned} & (\overline{\text{PONCTUATION}})_0 \wedge (\overline{\text{ARTG, PROC}})_1 \\ & (\text{PONCTUATION})_0 \wedge (\text{ARTG, PROC})_1 \end{aligned}$$

Les valeurs possibles de *Ce* sont: *ARTG, PROC*

Si on analyse ensuite *Ce programme* cela donne les résolutions suivantes:

$$\begin{aligned} & (\overline{\text{ARTG, PROC}})_1 \wedge (\overline{\text{SUBS, SUSU, VT}})_2 \\ & (\text{ARTG, PROC})_1 \wedge (\text{SUBS, SUSU, VT})_2 \end{aligned}$$

¹ ARTG = article général

SUBS = substantif complément

VT = verbe conjugué

PROC = pronom complément

SUSU = substantif sujet

Les valeurs possibles de C_e sont ici: $ARTG$

La résolution pour C_e sera donc l'intersection des deux listes

$(ARTG, PROC) \cap (ARTG) = ARTG$ (ici \cap est l'intersection ensembliste).

Le processus continue, ainsi de proche en proche, d'attribuer à chaque mot la valeur grammaticale compatible avec les valeurs répertoriées des mots et des associations binaires.

L'ANALYSE PAR DÉFAUT ²

L'analyse par défaut peut être définie comme suit:

Soit un texte T dont on veut effectuer l'analyse automatique. Supposons qu'un certain nombre de mots de T soient inconnus du dictionnaire. En d'autres termes, au cours de l'apprentissage ces mots n'ont pas encore été rencontrés. On se propose alors d'utiliser les informations syntaxiques accumulées (corrélations binaires et ternaires) pour attribuer aux mots inconnus du dictionnaire *la* ou *les* valeurs compatibles avec le contexte (valeurs par défaut).

La possibilité d'une telle opération est évidente.

En effet, soit une phrase φ de la forme: $M_1 M_2 \dots M_i X_j M_K \dots M_n$ où X_j est un mot inconnu du dictionnaire, auquel on doit donc attribuer une valeur par défaut.

Supposons pour simplifier les choses, que ni M_i ni M_K ne soient ambigus. Il correspond donc à M_i une valeur V_i et à M_K une valeur V_K . Il existe un ensemble de corrélations binaires ayant comme terme initial V_i . Elles s'écriront:

$$(V_i, V_x^1) (V_i, V_x^2) \dots (V_i, V_x^S)$$

où $V_x^1, V_x^2, \dots, V_x^S$ peuvent a priori être considérées comme l'une des valeurs possibles de X_j dans le texte étudié.

De même il existe un certain nombre de corrélations binaires possibles ayant V_K comme valeur terminale. Elles s'écrivent:

$$(V_x^1, V_K) (V_x^2, V_K) \dots (V_x^t, V_K) \text{ où } t \neq S \text{ a priori.}$$

² Voir commentaires 5 et 6.

La partie commune des ensembles

$$L_1 \equiv \{V_{X_j}^1, V_{X_j}^2, \dots, V_{X_j}^S\}, L_2 \equiv \{V_{X_j}^1, V_{X_j}^2, \dots, V_{X_j}^t\}$$

constituera *la* ou *les* valeurs par défaut de X_j .

Si $L_1 \cap L_2 = 0$, il y a incompatibilité, on pourrait afficher un message signalant cette incompatibilité, mais ce cas ne se produit en fait jamais.

Si $L_1 \cap L_2 \neq 0$, la partie commune des deux listes est proposée comme analyse par défaut.

Cette stratégie n'est évidemment pas la seule possible.

Le procédé d'analyse par défaut présente un certain nombre d'aspects linguistiques importants.

On peut souligner en particulier les faits suivants:

L_1 : un tel système fait apparaître les notions de classes de mots, de grammaire et de syntaxe produites par le corpus sur lequel est effectué l'apprentissage.

L_2 : il permet de connaître la réaction de la syntaxe apprise, à un mot inconnu, avant de décider du statut de ce mot.

L_3 : il permet d'évaluer la redondance syntaxique d'un certain texte dans la mesure où l'on peut toujours effectuer l'apprentissage sur une partie de ce texte et analyser le reste par défaut.

L_4 : selon que l'analyse par défaut est effectuée à l'aide des corrélations binaires ou ternaires, on peut mesurer l'importance dans la langue de ces corrélations.

L_5 : on peut essayer d'utiliser ce système dans le but de simuler l'apprentissage de la langue par l'enfant ou l'adulte, sous son aspect grammatical et syntaxique. Cependant on doit prendre ici une précaution et souligner que sur un texte donné, l'apprentissage fait par programme est optimum, alors que pour un être humain il y a des pertes assez nombreuses de l'information enregistrée.

D'autre part, l'analyse par défaut présente aussi des avantages du point de vue de la souplesse et de la performance de l'algorithme d'apprentissage.

En effet:

Performance: dans la mesure où l'on évite le codage de mots déjà connus du dictionnaires et dans la mesure où, comme le montrent déjà les résultats, on trouve presque toujours dans la décision par défaut au moins la bonne valeur, on a la possibilité d'*accélérer l'accumulation des données de l'apprentissage.*

Souplesse:

S_1 : l'analyse par défaut, permet de ne pas rejeter une phrase dans laquelle il y a relativement peu de mots inconnus. (Dans les programmes des grammaires arborescentes, un seul mot inconnu rejette toute la phrase).³ En pratique, on parvient avec les corrélations binaires à obtenir une analyse par défaut qui n'est pas trop ambiguë en présence d'un mot inconnu entre deux connus et même en présence de deux mots consécutifs inconnus. Par contre, trois mots consécutifs inconnus ne peuvent plus être analysés. Cependant, ce dernier cas, au bout d'un certain degré d'apprentissage, se rencontre rarement. Ainsi dans le texte de 1000 mots analysé par défaut (voir commentaire 6) on ne trouve le cas de trois mots consécutifs inconnus que deux fois.

S_2 : D'autre part, la possibilité d'une analyse par défaut permet d'envisager à un certain moment l'arrêt de l'apprentissage, ce qui dans certaines applications peut être important.

S_3 : Enfin, il est prévu, pour faciliter l'introduction des données linguistiques, de réaliser tout le système en conversationnel.

En pratique, une première expérience a été réalisée, dans laquelle, après avoir effectué un apprentissage sur 4000 mots, on a pris au hasard dans le même ouvrage 1000 autres mots qui ont été analysés par défaut à l'aide de corrélations binaires.

Les premiers résultats montrent que lorsqu'on a un mot inconnu entouré de deux mots connus, le nombre de valeurs attribuées par défaut est faible.

Bien entendu, l'emploi de corrélations ternaires ne pourra qu'améliorer ces résultats. D'autre part, il arrive que dans certains cas on ne puisse pas analyser un mot, lorsque le mot qui le précède, ou bien est inconnu du dictionnaire, ou bien contient encore un trop grand nombre d'ambiguïtés.

Enfin, soulignons que des mots enregistrés dans le dictionnaire avec une analyse fautive ne peuvent être relevés à l'aide de l'analyse par défaut, qui ne peut contester les valeurs attribuées aux mots par l'apprentissage. Pour connaître la réaction par défaut du système à ces mots, on doit les supposer inconnus, c'est-à-dire les effacer du dictionnaire.

³ Dans une arborescence une décision par défaut entraîne l'exploration d'un grand nombre de branches qui sera variable selon la position du mot dans l'arborescence.

CONCLUSION

L'expérience effectuée ne représente pour l'instant qu'un essai de constitution d'un algorithme d'apprentissage d'analyse syntaxique, effectué sur 4000 mots. Elle a permis de tester la méthode et le programme et de confirmer les hypothèses de travail.

A - Elle montre pour la partie manuelle:

— que le temps de constitution d'un tel programme peut être de l'ordre de deux ou trois mois.

— que le temps de constitution de l'échantillon linguistique est de l'ordre d'une dizaine de jours pour 4000 mots.

— par contre le temps nécessaire pour choisir les catégories grammaticales peut être plus long et dépend du système grammatical adopté.

B - Par ailleurs le traitement automatique a permis la constitution d'un dictionnaire avec les diverses catégories grammaticales relatives à chaque mot, et d'une syntaxe permettant de lever les ambiguïtés grammaticales, le répertoire syntaxique et le dictionnaire subissant un apprentissage au fur et à mesure de l'extension du corpus.

Le niveau d'analyse qui est prévu pour l'instant, est celui d'une analyse grammaticale comprenant environ 80 catégories grammaticales et la recherche du sujet et de certains groupements (les adjectifs relatifs aux substantifs et les adverbes regroupés autour des verbes). Les temps d'analyse sont très courts par rapport aux temps nécessaires dans les analyses en arbres.

Le traitement effectué pourra servir de prétraitement pour l'analyse des niveaux supérieurs, l'inégalité suivante étant vérifiée:

temps de traitement pour la levée des ambiguïtés gram- maticales	+	temps de traitement pour obtenir les ni- veaux linguistiques plus élevés	≪	Temps de traite- ment par une stra- tégie globale arborescente
---	---	---	---	---

C - Il est important de remarquer que la connaissance du niveau grammatical peut *suffire* pour des applications telles que la documentation automatique. Un algorithme d'analyse documentaire utilisant les résultats d'analyse obtenus par le système décrit est en voie d'élaboration et sera exposé ultérieurement.

D - L'analyse par défaut constitue enfin un élément linguistique important du système élaboré. Elle est effectuée à l'aide des corrélations binaires et il est prévu de l'améliorer en utilisant les corrélations ternaires.

COMMENTAIRES ET RESULTATS EXPERIMENTAUX

COMMENTAIRE 1.

On a ci-dessous un extrait de l'échantillon linguistique tel qu'il est codé manuellement, à l'exception de la numération qui est automatique.

Pour l'instant 4000 mots ont été codés de cette façon.

Cette première opération est la seule opération manuelle.

445	LA	----->	ARTD
446	PREMIERE	----->	ADJA
447	FOIS	----->	SUBS
448	!	----->	RESP
449	IL	----->	PROI
450	NOUS	----->	PROV
451	FAUT	----->	VTI
452	D'ABORD	----->	ADVA
453	CONSTATER	----->	V
454	QUE	----->	RECS
455	CES	----->	ARTG
456	CALCULATEURS	----->	SUSU
457	»	----->	VIPO
458	AVEC	----->	PREP
459	L'	----->	ARTD
460	IMMENSE	----->	ADJA

COMMENTAIRE 2.

Ci-après, figurent un certain nombre de mots avec les propriétés accumulées correspondantes.

On trouve de même les probabilités absolues d'apparition des mots et les probabilités relatives des propriétés lorsqu'on a tiré le mot. (Les probabilités ont été prises en cours d'apprentissage sur 816 mots).

Ainsi pour « D' » on trouve	Probabilité = $0,0073439 = \frac{6}{816}$
pour la valeur PREA	on trouve $0,16 = \frac{1}{6}$
pour la valeur ARTI	on trouve $0,33 = \frac{2}{6}$
pour la valeur PREP	on trouve $0,50 = \frac{3}{6}$

6 étant le nombre de fois où « D' » est rencontré.


```

D'
-----
CE MOT APPARAÎT      6. FOIS  PROBABILITE : 0.0073439
- PREA :             1.        PROBABILITE : 0.1666666
- ARTI :             2.        PROBABILITE : 0.3333333
- PREP :             3.        PROBABILITE : 0.5000000
D'ABORD
-----
CE MOT APPARAÎT      2. FOIS  PROBABILITE : 0.0024480
- ADVA :             1.        PROBABILITE : 0.5000000
- ADVP :             1.        PROBABILITE : 0.5000000
DANS
-----
CE MOT APPARAÎT      2. FOIS  PROBABILITE : 0.0024480
- PREP :             2.        PROBABILITE : 1.0000000
DE
-----
CE MOT APPARAÎT      26. FOIS  PROBABILITE : 0.0318237
- PREA :             9.        PROBABILITE : 0.3461538
- ARTI :             1.        PROBABILITE : 0.0384615
- PREP :            16.        PROBABILITE : 0.6153846

```

COMMENTAIRE 3.

On trouve ci-dessous la syntaxe des transitions binaires pour lesquelles au moins une des deux listes des mots de la source possède au plus un élément. (Par exemple si la source a la forme ARTD * VT elle ne figure pas dans le répertoire que nous donnons ici).

Lorsqu'il y a plusieurs résolutions possibles, on trouve un message: AMBIGUITE, qui facilite le dépouillement des résultats.

```

27          ADJ * PREA PREP PROV :
                -ADJ * PREA
28          ADJ * PREA PREP RECC :
                -ADJ * PREP
29          ADJ * RECS RENP RGSU RGSV RIF RSUM VIPF
                VIPO VIRI :
***** AMBIGUITE
                -ADJ * RECS
                -ADJ * RENP
                -ADJ * RIF
                -ADJ * VIPF
                -ADJ * VIPO
30          ADJ * RECS RENP RGSU RGSV RSUM :
                -ADJ * RECS
31          ADJ * RECS RENP RGSV RSUM :
                -ADJ * RENP
32          ADJ * RECS RGSV :
                -ADJ * RECS

```

33 AD ADJP * ADVP PREC RECC RECS :
-ADJ * RECS

34 ADJ ADJP * ARTI PREA PREP :
-ADJP * PREP

35 ADJ ADJP * IVTE VTE :
-ADJP * VTE

36 ADJ ADJP * PONF RECS RESP :
-ADJ * RECS

37 ADJ ADJP * RECS RENP RGSU RGSV RIF RSUM
VIPF VIPO VIRI :

***** AMBIGUITE

-ADJ * RECS
-ADJ * RECS

38 ADJA * ADJP SUBS :
-ADJA * SUBS

39 ADJA * RECS RENP RGSU RGSV RSUM :
-ADJA * RENP

40 ADJA * SUBS SUSU :

***** AMBIGUITE

-ADJA * SUBS
-ADJA * SUSU

41 ADJA * SUBS SUSU VPE :
-ADJA * SUSU

42 ADJA ADJC ADVL ADVP * PREA :

***** AMBIGUITE

-ADVL * PREA
-ADV * PREA

43 ADJA ADJC ADVL ADVP * SUBS :

***** AMBIGUITE

-ADJA * SUBS
-ADJC * SUBS

COMMENTAIRE 4.

Dans la page on trouve la liste des transitions ternaires pour lesquelles (comme pour le commentaire 3 relatif aux transitions binaires) au moins une des deux listes des mots de la source possède plus d'un élément. S'il y a ambiguïté, elle est marquée d'un message AMBIGUITE.

87 ADJA * SUBS * PONF RECS RESP :
***** AMBIGUITE

-ADJA * SUBS * PONF
-ADJA * SUBS * RECS
-ADJA * SUBS * RESP

88 ADJA * SUBS * PREA PREP :
-ADJA * SUBS * PREP

89 ADJA * SUBS * RECS RENP RGSU RGSV RIF RSUM
VIPF VIPO VIRI :

***** AMBIGUITE

-ADJA * SUBS * RENP
 -ADJA * SUBS * RIF
 -ADJA * SUBS * VIPF
 90 ADJA * SUBS * RECS RENP RGSU RGSV RSUM :
 -ADJA * SUBS * RENP
 91 ADJA * SUBS SUSU * ADJP :
 -ADJA * SUBS * ADJP
 92 ADJA * SUBS SUSU * ARTI PREA :
 -ADJA * SUBS * PREA
 93 ADJA * SUBS SUSU * ARTI PREA PREP :
 ***** AMBIGUITE
 -ADJA * SUBS * PREP
 -ADJA * SUSU * PREA
 94 ADJA * SUBS SUSU * PONF RECS RESP :
 -ADJA * SUBS * RECS
 95 ADJA * SUBS SUSU * PREA PREP RECC :
 -ADJA * SUBS * PREP
 96 ADJA * SUBS SUSU * VT
 -ADJA * SUSU * VT
 97 ADJA * SUBS SUSU VPE * ARTI PREA :
 -ADJA * SUSU * PREA
 98 ADJA * SUSU * ARTI PREA PREP :
 -ADJA * SUSU * PREP
 99 ADJA * SUSU * RECS RENP RGSU RGSV RIF RSUM
 VIPF VIPO VIRI :
 -ADJA * SUSU * RIF
 -ADJP * PROV * VT
 144 ADJP * ARTD PROV * SUBS
 -ADJP * ARTD * SUBS
 145 ADJP * ARTI PREA * ADJA ADPR SUJT :
 -ADJP * PREA * ADJA
 146 ADJP * ARTI PREA * ADJP SUBS SUSU :
 -ADJP * PREA * SUBS
 147 ADJP * ARTI PREA * SUBS :
 -ADJP * PREA * SUBS
 148 ADJP * ARTI PREA * SUBS SUSU :
 -ADJP * PREA * SUBS
 149 ADJP * ARTI PREA PREP * ADPR ARTI :
 -ADJP * PREP * ARTI

COMMENTAIRE 5.

On trouve ci-dessous le résultat de l'analyse de phrases constituées avec les mots du dictionnaire avec des valeurs qui peuvent être différentes de celles connues du dictionnaire et dans des combinaisons qui sont différentes du texte utilisé pour l'apprentissage.

LE	(D):ARTD
CALCULATEUR	(D):SUSU
ELECTRONIQUE	(D):ADJP
EST	(D):VTE
BIEN	(D):ADVA ADVI ADVP SUBS
BON	(D):ADJA
POUR	(D):PREP
L	(D):ARTD
HOMME	(D):SUBS
.	(D):*****INCOMPATIBILITE*****
S'	(D):RECS RESP
IL	(D):PROP
NE	(D):NE
LE	(D):ARTD PROV
POSSEDAIT	(D):VT
PAS	(D):ADVP
.	(D):RECS
IL	(D):PROP
NE	(D):NE
SE	(D):PROV
SERAIT	(D):IVTE VTE
PAS	(D):ADVP
POSE	(D):ADJ
SUR	(D):PREP
LA	(D):ARTD
LUNE	(D):SUBS
.	(D):RECS
LA	(D):ARTD
PASSION	(D):SUSU
ET	(D):RECS RSUM
LA	(D):ARTD
COLERE	(D):SUSU
DES	(D):PREA
GENS	(D):SUBS
CREENT	(D):VT
UN	(D):ADPR ARTI
EMBOUTEILLAGE	(D):SUSU
EXTRAORDINAIRE	(D):ADJA
SUR	(D):PREP
LES	(D):ARTD
MARCHES	(D):SUBS
.	(D):RECS
MEME SI	(D):LREC
BIENTOT	(D):ADVL
LES	(D):ARTD
HOMMES	(D):SUBS
VENAIENT	(D):VT
SUR	(D):PREP
LA	(D):ARTD
LUNE	(D):SUBS
.	(D):RECS
LEUR	(D):ARTP
AMBITION	(D):SUSU
SERAIT	(D):IVTE VTE
LA	(D):ARTD PROV

MEME	(D):ADJA ADJC ADVL ADVP
.	(D):PONF RECS RESP
VOILA	(D):IVTE
DES	(D):ARTI PREA
AUTOS	(D):SUBS
.	(D):RESP
LEUR	(D):ARTP PROV
USURE	(D):SUBS
ETAIT	(D):VTE
FABULEUSE	(D):ADJA
.	(D):RECS
LE	(D):ARTD
COMPTABLE	(D):SUSU
SOUFFRE	(D):VT
QUAND	(D):RECS
LES	(D):ARTD PROV
AUTOMATES	(D):SUBS
TRICHEURS	(D):SUBS
INTERROGENT	(D):VT
LA	(D):ARTD
BOULETTE	(D):SUBS
.	(D):RECS
LES	(D):ARTD
PARVENUS	(D):ADJ
SONT	(D):VTE
INQUIETS	(D):ADJ
DE	(D):PREP
LA	(D):ARTD
DECISION	(D):SUBS
DE	(D):PREP
RALENTER	(D):VP
L	(D):V
ECONOMIE	(D):SUBS
.	(D):RECS
LES	(D):ARTD
MORALISTES	(D):SUSU
SCANDALISES	(D):ADJ
,	(D):RIF
GROSSES	(D):ADJA
BETES	(D):SUBS
BIEN	(D):ADVI
EN	(D):PREA
PEINE	(D):SUBS
DE	(D):PREP
RESOUDRE	(D):V
LES	(D):ARTD
PROBLEMES	(D):SUBS
DE	(D):PREP
L'	(D):ARTD
HOMME	(D):SUBS
.	(D):RGSU
DIRIGENT	(D):VT
LE	(D):ARTD
MONDE	(D):SUBS
.	(D):RECS

COMMENTAIRE 6.

Analyse par défaut.

Dans les pages suivantes, on présente un essai d'analyse par défaut, effectuée sur un texte de 1000 mots, à l'aide des corrélations binaires.

Les mots des lignes 2 et 4 sont analysés par défaut et les mots des lignes 21, 22, 25, 30, sont analysés par défaut. On peut le constater en vérifiant dans la liste de comparaison, les mots qui figurent ou non dans le dictionnaire.

Le mot MOTEUR à la ligne 22 est analysé par défaut. Mais comme le nombre de valeurs trouvées est supérieur à 20, on a considéré que l'analyse par défaut présentait moins d'intérêt et on a mis des points de suspension.

1	LE	:	ARTD
2	CALCULATEUR	:	SUBS SUSU
3	PEUT	:	VT
4	TROUVER	:	ADVL ADVP ADVQ RECS RENP RESP RIF V VIPO
5	DANS	:	PREP
6	LE	:	ARTD
7	PROGRAMME	:	SUSU
8	D'	:	PREA
9	INSTRUCTIONS	:	SUBS
10	CHARGE	:	ADJP
11	EN	:	*****INCOMPATIBILITE*****
	PREP		
	PREA		
12	MEMOIRE	:	SUBS
13	LA	:	ARTD
14	MANIERE	:	SUBS
15	DONT	:	PREC
18	EN	:	PREA PREP
19	SONT	:	VTE
20	LE	:	ARTD
21	VERITABLE	:	ADJA ADPR ADVI ADVS SUBS SUJT SUSU
22	MOTEUR	:	ADJ ADJP ADPR ADVI ADVL ADVP NPRO PROG RECS REMP RESP RIF SUBS SUSU V VANT VIPF VIPO VPE etc.
23	.	:	RECS
24	UNE	:	ARTI
25	QUESTION	:	ADJA ADVI SUBS SUSU
26	NOUVELLE	:	ADJ ADJP
27	N'	:	NE
28	Y	:	PROV
29	SERA	:	VTE
30	TRAITÉE	:	ADJ ADVI ADVL ADVP PREP PROG RECS SUBS V VAUX VPE VTE VTI VTSU VVT
31	QUE	:	ADVP PREC RECC RECS
32	SI	:	RECS

