Traitement des mots inconnus FIPS

1 Introduction

- Nous allons présenter ici la méthode utilisée pour le traitement des mots inconnus. Pour notre analyseur syntaxique un mot inconnu est un mot qui n'est pas dans le lexique. La possibilité de rencontrer un mot inconnu du lexique est envisageable pour des phrases constituées d'un vocabulaire spécifique (mot ayant une probabilité d'apparence très faible) ou pour des néologismes courants du langage parlé qui sont très utilisés, par exemple par les journalistes. La recherche des mots inconnus nous intéresse fortement pour la lecture automatique d'un texte écrit, car tous les mots sans exception dans ce genre d'application doivent être prononcés. La méthode employée dans notre étude est basée principalement sur l'utilisation de l'information morphologique du mot. Nous allons donc essayer de montrer l'intérêt non négligeable de la morphologie pour le traitement des mots inconnus.

2 les informations disponibles

2.1 La majuscule

- Une suite de lettres qui constitue un mot inconnu est, en français, un porteur potentiel d'une bonne quantité d'informations. La première disponible dans une lecture de gauche à droite est la première lettre du mot qui peut être une majuscule indiquant soit que c'est le premier mot de la phrase (mais ce cas-là ne nous intéresse pas) soit que ce mot est en fait un nom propre. Dans ce dernier cas le nom propre est déterminant au sens où il ne permettra pas d'autre catégorie possible pour ce mot. La seconde volée d'informations peut apparaître sous forme multiple et nous la caractériserons par le terme terminaison de mot.

2.2 La terminaison

2.2.1 Les caractéristiques de la terminaison

- Une terminaison est difficilement identifiable car même si il est évident de savoir où se termine une terminaison (la dernière lettre du mot), il est quasiment impossible de savoir ou commence cette terminaison. Le terme terminaison est donc à prendre ici au sens large car il encadre aussi bien le terme suffixe que la désinence verbale et toutes fins de mots.

- L'idée est d'utiliser la terminaison du mot pour en déduire sa catégorie syntaxique et ses traits. En effet, on sait pour un mot de même famille reconnaître par sa forme sa catégorie syntaxique:

ex 1.1:

- *modifier* est le verbe
- *modification* est le nom
- *modifiable* est l'adjectif

- Dans l'exemple 1.1 on peut identifier le verbe par la terminaison *'er'*, le nom par la terminaison *'tion'* et l'adjectif par la terminaison *'able'*. Bien sûr ceci n'est qu'un exemple et un autre mot pourrait être tout aussi ambigu, si on considère le mot *"boucher"*, il peut soit être un verbe à l'infinitif soit un nom au singulier.

- Si on imagine que l'on propose dans notre analyseur syntaxique les deux solutions pour le mot *"boucher"* on peut espérer que le contexte syntaxique soit suffisant pour identifier laquelle des solutions est la meilleure. Ici la différence entre un verbe et un nom est tel que notre analyse devrait sans problème nous guider dans la bonne direction.

2.2.2 les ressources disponibles

- Une quantité non négligeable de l'information est constituée par notre lexique qui contient plus de 160000 mots. Une étude statistique a été réalisée sur ce lexique, les résultats nous permettent d'attribuer des poids en fréquence absolue aux terminaisons (voir paragraphe sur les statistiques du lexique pour plus de détails). Les informations caractérisantes qui ont été déduites, nous ont permis d'élaborer une stratégie d'utilisation de ces statistiques.

- Les règles de conjugaison à partir du radical sont aussi une bonne source d'information pour les terminaisons verbales.

- Pour les noms et adjectifs on peut aussi indiquer la possible utilisation d'information pour déduire le genre et le nombre par rapport à la terminaison.

- Récapitulatif des sources d'informations:

- Fréquences sur les terminaisons du lexique
- Terminaisons caractéristiques des désinences verbales
- Terminaisons caractéristiques des noms et adjectifs en genre et en nombre

3 La méthode

- On peut élaborer un système de traitement des mots inconnus comme suit: On a en entrée du module global le mot inconnu qui est bien entendu porteur de sa terminaison. Et on veut en sortie une liste (la plus fournie possible) de propositions d'objets lexicaux accompagnés de leurs traits respectifs. La fonction principale est donc *proposition d'objets lexicaux*, nous avons divisé cette dernière en quatre sous-fonctions qui sont:

- traitement des noms inconnus
- filtre élimateur
- traitement specifique des catégories
- mise en ordre de probabilités

3.1 Traitement des noms inconnus

- Tous les mots inconnus commençant par une majuscule seront considérés comme des noms propres. Le mot qui est identifié comme un nom propre doit être affecté de traits le caractérisant, nécessaires pour la suite du traitement. On affecte donc les traits suivants par défaut:

```
La catégorie syntaxique    : Nom
Le Nombre                  : Singulier ou pluriel
Le genre                   : Masculin ou Féminin
La personne                : 3e Personne du singulier ou du pluriel
Le type de Nom             : Nom propre
Le nom provient d'un verbe : Faux
Caractéristiques de nom    : - pays
                             - ville
                             - rivière
                             - montagne
                             - personne
                             - corporation
```

- L'identification d'un nom propre nous aide pour la synthèse bien sûr, mais aussi pour la traduction car un nom propre ne nécessite pas de traduction en général.

3.2 Filtre éliminateur

- Les différentes fréquences sur les terminaisons ne sont pas facilement exploitable, en effet il est difficile de savoir si une terminaison qui a une fréquence de 1000 pour les verbes et 100 pour les noms a plus de chance d'être un verbe et alors éliminer la solution du nom. En fait nous avons relevé 2 types de terminaisons utilisables pour le traitement des mots inconnus:

- Les terminaisons solitaires
- Les terminaisons nulles

- En fait les deux sont très liées, une terminaison dite solitaire est une terminaison qui a une seule fréquence différente de zéro. Alors qu'une terminaison dite nulle est une terminaison qui a au moins une fréquence égale à zéro.

- La terminaison *iez* (typique d'un verbe) est solitaire, voici ses valeurs:

```
Verbe: 12232
Nom  :     0
Adj  :     0
Adv  :     0
```

- La terminaison *iste* est dite nulle, voici ses valeurs:

```
Verbe:    18
Nom  :   128
Adj  :    26
Adv  :     0
```

3.2.1 La terminaison solitaire

- L'avantage de la terminaison solitaire est de garder une solution unique, mais cela implique des précautions surtout que la stratégie employée va de la plus grande (maximum 8 lettres) à la plus petite (2 lettres au minimum) terminaison pour un mot. En fait si on ne prend garde on risque de se retrouver avec des catégories possibles éliminées, c'est pour cela que l'on met des seuils minimums pour indiquer qu'une terminaison peut être solitaire. Ces seuils varient suivant la catégorie syntaxique, par défaut nous avons mis 2 pour les adverbes, 5 pour les adjectifs, 10 pour les noms et 15 pour les verbes. Bien entendu ces valeurs sont très subjectives et seuls les résultats nous permettent de mieux régler ces paramètres.

3.2.2 la terminaison nulle

- Elle est la plus courante et nous l'utilisons dans le cas où aucune terminaison solitaire n'a été identifiée. Toutes les fréquences égales à zéro provoquent l'élimination de la catégorie correspondante. En effet il est plus valable d'éliminer une fréquence nulle, qu'une fréquence qui est même 10000 fois plus petite que la plus grande.

3.3 Traitements spécifiques des catégories

- Les terminaisons sont plus ou moins caractérisantes suivant les catégories syntaxiques. Dans les cas extrêmes nous leur attribuons des valeurs par défaut.

3.3.1 Les désinences verbales

- Les terminaisons pour les verbes sont porteuses de beaucoup d'informations. On peut apprendre sur le verbe temps, mode, genre, nombre et personne. De plus les désinence verbales sont peu ambigües, à une terminaison correspond en général un seul temps, même s'il existe quelques exceptions qui sont détaillées dans la liste des désinences verbales. On trouve donc dans le programme, pour une désinence verbale, les champs suivants:

```
terminaison   : ARRAY [0..7] OF CHAR;
nombre        : NumberSet;
genre         : GenderSet;
personne      : PersonSet;
temps         : TempsType;
mode          : ModeSet;
```

- La liste des désinences verbales est classée de la plus grande (en nombre de caractères) vers la plus petite, pour nous permettre de reconnaître les terminaisons les plus caractéristiques. Une désinence reconnue est une désinence qui correspond exactement avec la terminaison du mot inconnu. Si aucune désinence n'a été trouvée, ce module retourne des valeurs par défaut qui sont pour le verbe:

```
nombre      := NumberSet{singular,plural};
genre       := GenderSet{masculine,feminine};
personne    := PersonSet{1..6};
temps       := presentT;
mode        := ModeSet{indicatif..conditionnel};
```

3.4 mise en ordre de probabilités

- La mise en ordre de probalités se fait suivant le poids des fréquences. Il est affiché à l'écran pour donner une information complémentaire à l'utilisateur dans le choix de l'analyse la plus probable.

4 Études complémentaires

- Ici sont détaillées les études qui ont été faites préalablement dans le cadre de la réalisation d'une solution pour le traitement des mots inconnus. Bien entendu tous les résultats ne nous ont pas servi, mais pouront être utilisés pour améliorer le traitement des mots inconnus.

4.1 les suffixes

- voici une liste de suffixes qui sont répertoriés trouvés dans le Grevisse.

- 58 pour les noms et adjectifs:

Catégorie	N	Adj
able		X
ade	X	
age	X	
aie	X	
aille	X	
aire	X	X
ais	X	X
aison	X	
an	X	
ant	X	
ard	X	X
asse	X	X
at	X	
âtre		X
aud	X	X
e	X	
é	X	X

- A remarquer pour les noms et adjectifs que le tableau ci-dessus est valable avec des *s* en fin de mot pour les mêmes mots mais au pluriel.

Catégorie	N	Adj
eau	X	
ée	X	
el		X
al		X
ement	X	
ence		X
ent	X	X
esque	X	X
esse	X	
ise	X	
et	X	
eur	X	
euse	X	
eux	X	X
aque		X
iaque		X

- Pour les adverbes on trouve le suffixe *ment*, seul candidat apte à construire des adverbes à partir d'autres mots.

4.2 Les désinences pour les temps simples

- La désinence étant la variation dans la finale, nous allons observer cette dernière et en déduire les temps qui peuvent y être associés.

```
nPS: n Personne du singulier
nPP: n Personne du Pluriel

PréSu : Présent du Subjonctif
PréId : Présent de l'indicatif
PréIp : Présent de l'impératif
PréIf : Présent de l'infinitif
ImpId : Imparfait de l'indicatif
PasSi : Passé Simple
ImpSu : Imparfait du subjonctif
ParPr : Participe Présent
Géron : Gérondif
ParPa : Participe Passé
FutSi : Futur Simple
ConPr : Conditionnel Présent

e        -> PréId 1PS, 3PS
         -> PréSu 1PS, 3PS
         -> PréIp 2PS
```

Catégorie	N	Adj
ible		X
ie	X	
ième	X	X
ien	X	X
ier	X	X
if		X
ille	X	
in	X	X
ine	X	X
ique	X	X
is	X	
isant	X	
isme	X	
issime		X
iste	X	X
itude	X	
o	X	X
oir	X	
on	X	
ot	X	X
té	X	
tion	X	
toire		X
u		X
ule	X	
ure	X	

```
es      -> PréId 2PP
        -> PréSu 2PP
ons     -> PréId 1PP
        -> PréIp 1PP
ez      -> PréId 2PP
        -> PréIp 2PP
ent     -> PréId 3PP
        -> PréSu 3PP
ions    -> PréSu 1PP
        -> ImpId 1PP
iez     -> PréSu 2PP
        -> ImpId 2PP

ais     -> ImpId 1PS, 2PS
ait     -> ImpId 3PS
aient   -> ImpId 3PP

ai      -> PasSi 1PS
as      -> PasSi 2PS
a       -> PssSi 3PS
```

```
âmes     -> PasSi 1PP
âtes     -> PasSi 2PP
èrent    -> PasSi 3PP
is       -> PasSi 1PS, 2PS
it       -> PasSi 3PS
îmes     -> PasSi 1PP
îtes     -> PasSi 2PP
irent    -> PasSi 3PP
us       -> PasSi 1PS, 2PS
ut       -> PasSi 3PS
ûmes     -> PasSi 1PP
ûtes     -> PasSi 2PP
urent    -> PasSi 3PP
ins      -> PasSi 1PS, 2PS
int      -> PasSi 3PS
înmes    -> PasSi 1PP
întes    -> PasSi 2PP
inrent   -> PasSi 3PP

asse     -> ImpSu 1PS
asses    -> ImpSu 2PS
ât       -> ImpSu 3PS
assions  -> ImpSu 1PP
assiez   -> ImpSu 2PP
assent   -> ImpSu 3PP
isse     -> ImpSu 1PS
isses    -> ImpSu 2PS
ît       -> ImpSu 3PS
issions  -> ImpSu 1PP
issiez   -> ImpSu 2PP
issent   -> ImpSu 3PP
usse     -> ImpSu 1PS
usses    -> ImpSu 2PS
ût       -> ImpSu 3PS
ussions  -> ImpSu 1PP
ussiez   -> ImpSu 2PP
ussent   -> ImpSu 3PP
insse    -> ImpSu 1PS
insses   -> ImpSu 2PS
înt      -> ImpSu 3PS
inssions -> ImpSu 1PP
inssiez  -> ImpSu 2PP
inssent  -> ImpSu 3PP

er       -> PréIf
ir       -> PréIf
oir      -> PréIf
re       -> PréIf

ant      -> ParPr
         -> Géron
```

```
é        -> ParPs 1Ps, 2PS, 3PS, 1PP, 2PP, 3PP
ée       -> ParPs 3PS
és       -> ParPs 1PP, 2PP, 3PP
ées      -> ParPs 1PP, 2PP, 3PP
i        -> ParPs 1Ps, 2PS, 3PS, 1PP, 2PP, 3PP
ie       -> ParPs 3PS
is       -> ParPs 1PP, 2PP, 3PP
ies      -> ParPs 1PP, 2PP, 3PP
u        -> ParPs 1Ps, 2PS, 3PS, 1PP, 2PP, 3PP
ue       -> ParPs 3PS
us       -> ParPs 1PP, 2PP, 3PP
ues      -> ParPs 1PP, 2PP, 3PP

rai      -> FutSi 1PS
ras      -> FutSi 2PS
ra       -> FutSi 3PS
rons     -> FutSi 1PP
rez      -> FutSi 2PP
ront     -> FutSi 3PP

rais     -> ConPr 1PS, 2PS
rait     -> ConPr 3PS
rions    -> ConPr 1PP
```

- Si on part de la forme terminale du mot, on remarque qu'il n'y a pas trop d'ambiguïté sauf entre le présent du subjonctif et le présent de l'indicatif et pour la terminaison *iez* et *ions*.

4.3 Les statistiques

- On se propose de se servir du lexique du français du LATL pour construire une base de données statistiques où apparaît pour chaque terminaison un nombre en fonction de la catégorie syntaxique du mot.

4.3.1 La méthode

- On parcourt le lexique du français en entier en se fixant une limite raisonnable quant à la taille maximale d'une terminaison (6). A chaque nouvelle suite de caractères, on crée un nouvel emplacement dans la liste des terminaisons qui pointe sur une carte des catégories où sera indiquée, pour chaque catégorie, une fréquence.

4.3.2 Le développement

- Pour accéder plus rapidement aux terminaisons nous créons une table des terminaisons à 2 dimensions qui est de taille 256 (pour le premier caractère) par la taille max de la terminaison, et qui pointe pour chaque terminaison finale (même lettre en fin de mot) sur une liste de terminaisons de même terminaison finale. Comme illustré ci-dessous:

```
tableterm[0][2]
          .
          .
tableterm[81][2] ab -> ac -> ad ...
          .
          .
          .
          .
tableterm[256][2]
          .
tableterm[0][3]
          .
          .
tableterm[81][3] abc -> aca -> adi ...
          .
          .
          .
          .
tableterm[256][3]
```

- On a donc les chaînes de caractères inversées par rapport à notre sens de lecture. Dans l'exemple ci-dessus on peut avoir les mots suivants *baba, décida, harmonica* ...

- Chaque élément de cette liste pointe sur une structure contenant toutes les catégories syntaxiques.

```
CatFreq = RECORD
            nomFreq      : CARDINAL;
            adjectifFreq : CARDINAL;
            adverbeFreq  : CARDINAL;
            verbeFreq    : CARDINAL;
            préposition  : CARDINAL;
            determinant  : CARDINAL;
            conjonction  : CARDINAL;
            interjection : CARDINAL;
          END;
```

4.4 Les résultats

4.4.1 Nos intuitions par rapport aux résultats

-Ici suit une liste de recherche "à la main" mise en rapport aux statistiques créée à partir du lexique. Le but de la recherche est de vérifier nos intuitions sur des terminaisons qui sembleraient caractérisantes.

* Terminaisons sur le temps donc pour les verbes:

```
er       -> re       -> 2594 V  396 N  67 Adj 5 Adv +
e        -> e        -> 11819 V  6171 N 3740 Adj  54 Adv  44 Conj 28 Det 5 P ?
es       -> se       -> 15139 V  5698 N 3672 Adj  11 Det 4 Adv ?
ons      -> sno      -> 18213 V  1140 N  19 Adj +
ez       -> ze       -> 18204 V 4 N ++
ait      -> tia      -> 6086 V  16 N  11 Adj ++
ais      -> sia      -> 7099 V  24 N  12 Adj ++
aient    -> tneia    -> 6077 V *
ions     -> snoi     -> 12232 V 845 N +
iez      -> zei      -> 12232 V *
ai       -> ia       -> 5622 V  14 N +
as       -> sa       -> 5621 V  113 N  4 Adj  1 Adv ?
a        -> a        -> 5623 V  96 N  4 Adv 2 Adj ?
âmes     -> semâ     -> 2584 V  1 N ++
âtes     -> setâ     -> 2586 V  1 N ++
èrent    -> tnerè    -> 2639 V *
erais    -> siare    -> 2589 V  1 N ++
erait    -> tiare    -> 2589 V *
erions   -> snoir    -> 3582 V *
eraient  -> tneiare  -> 2589 V *
eriez    -> zeire    -> 2589 V *
erai     -> iare     -> 2589 V  1 N ++
eras     -> sare     -> 2589 V *
era      -> are      -> 2589 V *
erons    -> snore    -> 2589 V  12 N ++
erez     -> zere     -> 2589 V *
eront    -> tnore    -> 2589 V *
é        -> é        -> 2585 V  628 Adj  556 N  5 Adv  2 P ?
és       -> sé       -> 2585 V  619 Adj  528 N  ?
ées      -> seé      -> 2590 V  621 Adj  187 N ?
ie       -> ei       -> 570 V  511 N  69 Adj  1 Adv ?
ies      -> sei      -> 445 N  388 N  67 Adj ?
ient     -> tnei     -> 6295 V  8 N  4 Adj ++
iait     -> tiai     -> 139 V *
iais     -> siai     -> 139 V *
iions    -> snoii    -> 278 V *
iiez     -> zeii     -> 278 V *
iaient   -> tneiai   -> 139 V *
iai      -> iai      -> 138 V *
ias      -> sai      -> 344 V  4 N ++
ia       -> ai       -> 138 V  16 N ++
iâmes    -> semâi    -> 138 V *
iâtes    -> setâi    -> 138 V *
ièrent   -> tnerèi   -> 140 V *
ierai    -> iarei    -> 180 V *
ieras    -> sarei    -> 180 V *
iera     -> arei     -> 180 V *
ierons   -> snorei   -> 180 V *
```

```
ierez    -> zerei    -> 180 V *
ieront   -> tnorei   -> 180 V *
ierais   -> siarei   -> 180 V *
ierions  -> snoirei  -> 180 V *
ieriez   -> zeirei   -> 180 V *
ieraient -> tneiarei -> 180 V *
ié       -> éi       -> 138 V  21 Adj  13 N +
iés      -> séi      -> 138 V  21 Adj  18 N +
iées     -> seéi     -> 138 V  21 Adj  6 N +
ir       -> ri       -> 273 V  101 N  5 Adj 2 Adv ?
is       -> si       -> 7099 V 220 N  114 Adj  11 Adv +
it       -> ti       -> 6692 V 100 N  41 Adj 3 Adv +
issons   -> snossi   ->  207 V 2 N  1 Adj ++
issez    -> zessi    ->  207 V *
issent   -> tnessi   ->  528 V *
issait   -> tiassi   ->  205 V *
issais   -> siassi   ->  205 V *
issions  -> snoissi  ->  731 V  9 N ++
issiez   -> zeissi   ->  731 V *
issaient -> tneiassi ->  284 V *
îmes     -> semî     ->  329 V *
îtes     -> setî     ->  331 V  1 N ++
irent    -> tneri    ->  345 V *
irai     -> iari     ->  265 V *
iras     -> sari     ->  265 V *
ira      -> ari      ->  265 V *
irons    -> snori    ->  264 V  1 N ++
irez     -> zeri     ->  264 V *
iront    -> tnori    ->  249 V *
irais    -> siari    ->  263 V *
irait    -> tiari    ->  265 V *
irions   -> snoiri   ->  280 V *
iriez    -> zeiri    ->  280 V *
iraient  -> tneiari  ->  265 V *
oir      -> rio      ->  68 N  27 V  1 Adv ?
eut      -> tue      ->  1 V  1 N #
ois      -> sio      ->  38 N 13 Adj 5 Adv 21 V #
oit      -> tio      ->  13 V  13 N 3 Adj 1 Adv #
evons    -> snove    ->  14 V *
evez     -> zeve     ->  14 V *
eus      -> sue      ->  2 V +
euvent   -> tnevue   ->  2 V +
oivent   -> tnevio   ->  5 V ++
oyais    -> siayo    ->  26 V *
oyait    -> tiayo    ->  26 V *
yerai    -> iarey    ->  1 V +
yais     -> siay     ->  57 V *
yait     -> tiay     ->  58 V *
oirai    -> iario    ->  4 V +
us       -> su       ->  174 V  106 N  99 Adj  11 Adv ?
ut       -> tu       ->  59 V  41 N  5 Adv  4 Adj ?
```

```
oie      -> eio      ->  45 V  10 N ?
u        -> u        ->  239 N  114 V  102 Adj 2 Adv ?
ant      -> tna      ->  3029 N  266 Adj  144 N  10 Adv +

* Terminaisons pour les noms:

iste     -> etsi     -> 128 N  26 Adj  18 V +
istes    -> setsi    -> 130 N  26 Adj  9 V +
eur      -> rue      -> 552 N  60 Adj  3 V  1 Adv +
eurs     -> srue     -> 509 N  59 Adj  1 Adv +
ard      -> dra      -> 47 N  16 Adj +
ards     -> sdra     -> 47 N  15 Adj +


* Terminaisons pour les adjectifs:

al       -> la       -> 189 Adj  79 N  2 Adv  1 V +
ale      -> ela      -> 189 Adj  57 N  25 V +
ales     -> sela     -> 189 Adj  55 N  13 V +
aux      -> xua      -> 190 N  181 Adj  4 V ?
eux      -> xue      -> 265 Adj  44 N  1 Adv  1 V +
euse     -> esue     -> 279 Adj  149 N +
euses    -> sesue    -> 280 Adj  149 N +
able     -> elba     -> 236 Adj  26 N  4 V ++
ables    -> selba    -> 235 Adj  14 N  2 V ++

* Terminaisons pour les adverbes:

ement    -> tneme    -> 277 N  206 Adv  4 Adj  2 V #
iment    -> tnemi    -> 18 V  16 N  7 Adv ?
ément    -> tnemé    -> 20 Adv  5 N  1 Adj +
emment   -> tnemme   -> 19 Adv ++
amment   -> tnemma   -> 15 Adv  2 V +
```

- Pour mieux estimer nos résultats nous avons remarqué quelques points que nous indiquons à l'aide des symboles suivants:

- # indique une intuition complètement fausse

- * indique une intuition totalement justifiée

- ? indique une terminaison semi-caractéristique

- + indique que la terminaison semi-caractéristique est presque identifiée

4.4.2 Sur les intuitions

- Les symboles sont présents à titre d'information, pour consulter plus rapidement ces exemples, ils sont trop subjectifs pour leur porter une attention ayant pour but une utilisation quelconque.

- Ce qui apparaît clairement dans ces résultats, c'est la plus grande reconnaissance de verbe que toute autre catégorie syntaxique. Ayant reconnu une terminaison de verbe, il est facile en général d'en déduire le temps et la personne du mot auxquels la terminaison correspond (pour cela il faudra bien entendu lier ces terminaisons au temps et à la personne).

ex:

```
irions   -> snoiri   ->  280 V *
```

- Dans cet exemple on a, à coup sûr, affaire à un verbe au conditionnel à la première personne du pluriel.

- On peut noter aussi que plus un graphème est grand plus il est caractérisant du point de vue syntaxique mais moindre sera la fréquence associée. On en déduit le principe suivant, qui est que l'on cherchera en premier lieu la plus grande terminaison possible pour un mot.

ex: pour le mot *"mangerais"*, on pourrait en déduire les terminaisons suivantes:

```
ais      -> sia      -> 7099 V  24 N  12 Adj ++
erais    -> siare    -> 2589 V  1 N ++
```

- Bien sûr seule la seconde est interressante, car plus grande, la première en fait ne correspondrait qu'au mot *mangeais*.

4.4.3 les solitaires

- Nous avons créé un fichier texte qui a permis d'identifier les terminaisons solitaires que l'on pourrait pratiquement appeler terminaisons caractéristiques, seulement ceci est vrai pour des fréquences solitaires d'une certaine valeur. Cette valeur n'est en fait pas identifiable pour garantir une terminaison réellement caractéristique, mais nous nous efforcerons de trouver un seuil qui se rapproche de la réalité.

- Nous constatons que les fréquences solitaires les plus élevées correspondent à des verbes. Ici nous ne pouvons pas donner tous les résultats, à cause de la taille du fichier, mais nous allons nous efforcer d'identifier les cas remarquables.

- On peut remarquer le cas de la terminaison 'iez' qui obtient une fréquence solitaire énorme de l'ordre de 12232, qui identifie un verbe.

- Nous allons aussi mentionner le cas des fréquences supérieures à 1000, à titre d'information.

- La première terminaison caractéristique ayant une catégorie syntaxique différente des verbes est la terminaison *'ements'* pour les noms, avec une fréquence de 263. Ensuite on arrive aux terminaisons *'eries'* et *'erie'*, toujours pour les noms, avec une fréquence de 131.

- Voici une liste des fréquences solitaires pour des terminaisons relatives à des noms, des adjectifs ou des adverbes, avec une fréquence supérieure à 50.

- On remarque qu'il n'y a toujours pas d'adjectif ou d'adverbe. En fait le premier adjectif apparaît à la fréquence de 23 pour la terminaison *'nnelles'* et est suivie, pour le tableau suivant des autres terminaisons pour les adjectifs.

- Pour les adverbes il n'existe pas de fréquences solitaires caractéristiques.

zei	12232	iez
tneia	6077	aient
zeis	4179	siez
zeir	3582	riez
snoir	3582	rions
zeiss	3578	ssiez
tneiar	3309	raient
zer	3308	rez
tness	3294	ssent
zeissa	2643	assiez
tnerè	2639	èrent
tnessa	2613	assent
zere	2589	erez
zeire	2589	eriez
tnore	2589	eront
tneiare	2589	eraient
tiare	2589	erait
snoire	2589	erions
sare	2589	eras
are	2589	era

- On en déduit donc assez logiquement que les verbes puis ensuite les noms sont plus facilement identifiables. Loin derrière arrivent les adjectifs qui sont peu caractéristiques.

4.4.4 les hautes fréquences

- Ces statistiques correspondent en fait aux terminaisons les plus courantes en français. Voici les douze plus importantes fréquences pour une terminaison donnée:

- là encore les fréquences sont très élevées pour les verbes par rapport aux autres catégories syntaxiques.

- Ces dernières statistiques ne sont pas en fait d'un grand secours car elles ne permettent pas d'identifier clairement des terminaisons caractéristiques. En effet, en général il y a trop

stneme	263	ements
seire	131	eries
eire	131	erie
sems	121	smes
semsi	106	ismes
noitc	90	ction
snoitar	75	retions
srio	58	oirs
stnemes	53	sements
snoitac	52	cations
noiti	52	ition

ellenn	23	nnelle
ellenno	22	onnelle
seuqigo	20	ogiques
sesueut	18	tueuses
esueut	18	tueuse
lennoit	15	tionnel
stnel	14	lents
semèi	14	ièmes
emèi	14	ième
selbati	13	itables
selbass	13	ssables
selban	13	nables
elbati	13	itable
elbass	13	ssable
elban	13	nable
xuell	12	lleux

Catégories	N	Adj	Adv	V	P	Det	Con	Int	Oth	Total
s	11059	5594	54	49340	9	28	6	0	0	s
t	1155	507	313	31235	7	6	8	0	0	t
tn	548	379	286	21382	7	2	5	0	0	nt
sn	1473	130	3	18342	0	5	1	0	0	ns
z	22	0	1	18217	0	0	0	0	0	z
sno	1140	19	0	18213	0	2	0	0	0	ons
ze	4	0	1	18204	0	0	0	0	0	ez
tne	376	95	273	15187	0	0	3	0	0	ent
se	5698	3672	4	15139	0	11	0	0	0	es
zei	0	0	0	12232	0	0	0	0	0	iez
snoi	845	1	0	12232	0	2	0	0	0	ions
e	6171	3740	54	11819	5	28	44	0	0	e

peu d'écart entre la plus grande fréquence et la deuxième plus grande fréquence.

4.4.5 Essai sur les fréquences relatives

- Pour l'instant nos études statistiques portent sur des fréquences absolues. On sait pourtant que le nombre total de verbes est beaucoup plus important que le nombre total d'adverbes. Dans le but de comparer ce qui est comparable nous allons ramener les fréquences des différentes catégories à un poids identique.

- Sur 164225 mots on a:

Pour la terminaison 's' sur ses 66036 mots on a:

- Ce résultat indique donc que l'on a 74 % de chance d'avoir à faire à un verbe si on trouve un mot se terminant par un 's'. A première vue ce résultat semble un peu étrange, notre intuition nous aurait plus porté à penser que le 's' était la marque préférée en nombre d'un

127923	Verbes	77.89 %
23548	Noms	14.39 %
11965	Adjectifs	7.29 %
517	Adverbes	0.31 %
119	Conjonctions	0.07 %
109	Déterminants	0.07 %
43	Prépositions	0.03 %
1	Interjection	0.00 %
0	Autres	0.00 %

49340	Verbes	74.72 %
11059	Noms	16.75 %
5594	Adjectifs	8.47 %
54	Adverbes	0.08 %
28	Déterminants	0.04 %
9	Prépositions	0.01 %
6	Conjonctions	0.01 %

nom ou d'un adjectif. Mais le résultat semble assez logique quand on considère le nombre important de verbes existants (avec les formes fléchies). C'est pour cela que nous allons considérer par la suite que nous avons autant de chance d'avoir un verbe, un adjectif ou toute autre catégorie lors d'un traitement de mot inconnu. Nous allons donc calculer la fréquence relative.

Si on considère seulement les catégories syntaxiques, on a:

```
38.6 %  des verbes qui se terminent par un 's'
46.96 % des noms qui se terminent par un 's'
46.75 % des adjectifs qui se terminent par un 's'
10.44 % des adverbes qui se terminent par un 's'
25 %    des déterminants qui se terminent par un 's'
20 %    des prépositions qui se termine par un 's'
 5 %    des conjonctions qui se terminent par un 's'
```

- Ceci nous indique que l'on a plus de chance de rencontrer un nom ou un adjectif qu'un verbe. En fait le mieux est d'utiliser les deux résultats précédents pour n'en faire qu'un.

4.4.6 Essai sur les fréquences moyennes terminales

- On appelle fréquence moyenne terminale pour une catégorie syntaxique, la moyenne des fréquences de chaque graphème constituant la terminaison.

- Nous allons calculer les fréquences moyennes terminales pour différents mots et pour chacune des catégories intéressantes:

- le mot, *"vraiment"* qui est un adverbe.

Catégories	N	Adj	Adv	V	P	Det	Con	Int	Oth	Total
aiment	0	0	0	1	0	0	0	0	0	1
iment	16	0	7	18	0	0	0	0	0	41
ment	318	6	270	87	0	0	2	0	0	683
ent	376	95	273	15187	0	0	3	0	0	15934
nt	548	379	286	21382	7	2	5	0	0	22609
t	1155	507	313	31235	7	6	8	0	0	33231
Total	2413	987	1149	67910	14	8	18	0	0	72499
En %	3.33	0.74	1.58	93.28	0.01	0	0.02	0	0	100%

- Ici les pourcentages ne sont pas interprétables.

En pourcentage moyen:

Catégories	N	Adj	Adv	V	P	Det	Con	Int	Oth	Total
aiment	0	0	0	100	0	0	0	0	0	1
iment	39.02	0	17.07	43.90	0	0	0	0	0	41
ment	46.56	0.87	39.53	12.73	0	0	0.29	0	0	683
ent	2.35	0.59	1.71	95.31	0	0	0.01	0	0	15934
nt	2.42	1.67	1.26	94.57	0.01	0	0.01	0	0	22609
t	3.47	1.52	1.88	93.99	0.01	0.01	0.02	0	0	33231
total	15.63	0.77	10.24	73.42	0	0	0.05	0	0	100

- On voit bien que la moyenne des fréquences terminales est bien plus caractéristique, car on voit bien apparaître les trois principaux candidats qui sont le verbe, le nom et l'adverbe. Même si l'adverbe n'arrive qu'en troisième position sa moyenne est suffisamment importante pour être caractéristique.

Le mot *"partaient"* qui est un verbe

Catégories	N	Adj	Adv	V	P	Det	Con	Int	Oth	Total
rtaient	0	0	0	27	0	0	0	0	0	27
taient	0	0	0	398	0	0	0	0	0	398
aient	0	0	0	6077	0	0	0	0	0	6077
ient	8	4	0	6295	0	0	0	0	0	6307
ent	376	95	273	15187	0	0	3	0	0	15934
nt	548	379	286	21382	7	2	5	0	0	22609
t	1155	507	313	31235	7	6	8	0	0	33231
Total	2087	985	872	80601	14	8	16	0	0	84583
En %	2.46	1.16	1.03	95.29	0.01	0.01	0.01	0	0	100

En pourcentage moyen:

- On remarque que le résultat est déjà plus parlant, car on a affaire à trois fréquences solitaires à la suite qui augmentent la moyenne pour les verbes.

Catégories	N	Adj	Adv	V	P	Det	Con	Int	Oth	Total
rtaient	0	0	0	100	0	0	0	0	0	27
taient	0	0	0	100	0	0	0	0	0	398
aient	0	0	0	100	0	0	0	0	0	6077
ient	0.13	0.06	0	99.80	0	0	0	0	0	6307
ent	2.36	0.59	1.71	95.37	0	0	0.02	0	0	15934
nt	2.42	1.67	0.01	94.57	0.03	0.01	0.02	0	0	22609
t	3.47	1.52	0.94	93.99	0.02	0.02	0.02	0	0	33231
Total	1.20	0.54	0.38	97.67	0	0	0.01	0	0	100

Le mot *"dentiste"* qui est un nom

Catégories	N	Adj	Adv	V	P	Det	Con	Int	Oth	Total
ntiste	2	0	0	0	0	0	0	0	0	2
tiste	11	0	0	0	0	0	0	0	0	11
iste	128	26	0	18	0	0	0	0	0	172
ste	148	44	0	56	0	0	0	0	0	248
te	751	572	9	827	0	9	1	0	0	2169
e	6171	3740	54	11819	5	28	44	0	0	21861
Total	7211	4382	63	12720	5	37	45	0	0	24463
En %	29.48	19.96	0.26	52.00	0.02	0.15	0.18	0	0	100

En pourcentage moyen:

Catégories	N	Adj	Adv	V	P	Det	Con	Int	Oth	Total
ntiste	100	0	0	0	0	0	0	0	0	2
tiste	100	0	0	0	0	0	0	0	0	11
iste	74.42	15.12	0	10.46	0	0	0	0	0	172
ste	59.68	17.74	0	22.58	0	0	0	0	0	248
te	34.62	26.37	0.41	38.13	0	0.41	0.05	0	0	2169
e	28.23	17.11	0.24	54.06	0.02	0.13	0.20	0	0	21861
Total	66.16	12.72	0.11	20.87	0	0.09	0.04	0	0	100

- On déduit de cet exemple que cette terminaison caractérise bien le nom. Par contre il apparaît aussi que la prise en compte de la terminaison finale (la dernière lettre du mot) a plus tendance à brouiller nos résultats qu'à nous aider dans notre recherche. En effet, on imagine mal qu'une seule lettre, aussi rare soit elle, puisse caractériser le mot au niveau syntaxique.

4.5 conclusion sur les résultats

- La difficulté de l'utilisation pratique des statistiques réside dans le fait qu'elle ne sont que des statistiques, qu'elles ne permettent d'identifier des généralités que si on y associe des limites. Mais comment choisir ces limites, par rapport à d'autres statistiques, à nos intuitions, à des intuitions de spécialistes...? La question reste posée et la solution du spécialiste semble encore

la meilleure alternative. Mais en général le spécialiste ne pourra pas faire un choix sur ses limites, car il y verra du bricolage, une tendance un peu trop aléatoire et le non traitement de cas spécifique.

- Ici le traitement par statistiques est comme une bouée de secours. En effet elle est activée en dernier recours pour l'analyse lexicale, lorsqu'un mot n'est pas reconnu. Elle n'a donc pas une place prédominante et par conséquent n'influence pas nos résultats lors d'une reconnaissance totale des mots.

- En premier lieu, on peut dire que ces statistiques ne nous apportent guère d'information sur des catégories à petites populations comme les déterminants, les conjonctions, les prépositions ou les interjections. En résumé, les statistiques nous renseignent sur plusieurs aspects du mot. En général on récupère plusieurs fréquences, c'est-à-dire plusieurs catégories possibles qui si elles ne nous renseignent pas facilement sur la meilleure catégorie, nous précisent que telle catégorie n'est pas possible (pour une fréquence nulle). La deuxième information essentielle est l'existence d'une fréquence solitaire qui à forte fréquence identifie une terminaison caractéristique. La troisième nous permet d'ordonner par ordre de probabilité les catégories possibles en utilisant la fréquence moyenne terminale. On peut aussi envisager d'utiliser la fréquence relative, mais seuls des tests sur un corpus nous permettra d'en justifier son utilisation.

4.6 Les résultats

- Nous avons essayé de vous montrer ici les résultats obtenus avec des mots inconnus que nous avons soumis à notre analyseur FIPS.

```
>les enfants glupent à la mer.
[CP [TP[DP les [NP enfants ]][T' glupent [PP à [DP la [NP mer ]]]]]]
score :      15

**** mot reconnu !: glupent

> les glupes portent des chemises.

[CP [TP[DP les [NP glupes ]][T' portent [DP [PP des [DP [NP chemises ]]]]]]
score :       5

**** mot reconnu !: glupes

> les oiseaux volent glupement dans le ciel.
[CP [TP[DP les [NP oiseaux ]]i[T' volent [DP ei][AdvP glupement ][AdvP [PP dans [DP le [
el ]]]]]]]
score :      35

**** mot reconnu !: glupement

> les avions sont glupaux en montagne.

[CP [TP[DP les [NP avions ]]i[T' sont [FP[DP ei][F' [AP[DP ei][A' glupaux ]]]][AdvP [PP
P montagne ]]]]]]
```

```
score :      35

**** mot reconnu !: glupaux
```

5 Conclusion

- Comme on peut le voir, cette première approche nous a permis d'obtenir des résultats assez satisfaisants. Les mots inconnus avec des terminaisons évidentes (pour un être humain) sont en principe bien reconnus après activation de l'analyse syntaxique. Pourtant quelques problèmes subsistent et certaines améliorations devront être apportées pour des applications comme la traduction automatique et la synthèse de la parole. En effet, il arrive parfois pour certains exemples que l'analyse ne donne aucun résultat, car nous avons éliminé des possibilités lexicales. Il faut sûrement envisager une boucle de retour qui repropose ces formes lexicales à l'analyseur tant que l'analyseur n'a pas de solution. De plus, ici nous avons étudié le cas où l'analyseur est abandonné à lui même (fonctionnant automatiquement). Pourtant, de la même façon que nous avons de l'intéraction avec l'analyseur syntaxique, il faudra proposer à l'utilisateur une possibilité d'intéraction lexicale avec les mots inconnus, en proposant à cette personne une gamme de solutions par ordre de préférence. Pour terminer, on peut dire que le traitement des mots inconnus devra évoluer et évoluera au fûr et à mesure du développement de notre recherche, tout comme évolue la constitution de nos lexiques, ceci pour améliorer de plus en plus la robustesse de nos analyses.

References

[1] Grevisse, Maurice(86). Le bon usage: Grammaire française, 12e edition, Gembloux-Duculot.

[2] Robert, Paul(89). Le petit Robert 1: Dictionnaire alphabétique et analogique de la langue française.

TABLE DES MATIERES