Manuel d'utilisateur de Densidées

version 1.3 - 30/06/2010

Philippe Gambette, Hyeran Lee

30 juin 2010

Table des matières

1 Introduction

Densidées calcule la densité des idées d'un texte (au sens de [2] et [5], c'est à dire le nombre moyen d'idées exprimées en 10 mots), utile en particulier pour l'analyse des discours de patients atteint de la maladie d'Alzheimer. Il adapte à la langue française le calcul implémenté pour la langue anglaise dans le logiciel CPIDR [1].

Densidées est un logiciel libre sous licence GPL. Toutefois, si vous l'utilisez, nous vous invitons à citer l'article qui lui est consacré [4]. Cet article contient en particulier une évaluation de l'étiquetage des prédicats réalisé automatiquement par Densidées sur un corpus oral retranscrit de 13939 mots dont 5747 propositions. Les résultats de la version 1.2 sont 2,7% de faux négatifs et 3,1% de faux positifs, soit un taux d'erreur de 0,5% sur le nombre de prédicats.

En cas de problème d'utilisation de Densidées, vous pouvez envoyer votre question à `gambette@lirmm.fr`, accompagnée si possible d'une capture d'écran du résultat de la ligne de commande (son utilisation est détaillée en section 3.3).

2 Installation du programme

Densidées est écrit en Python (version 2.6), il faut donc commencer par télécharger Python sur http://www.python.org/download/releases/2.6/ et l'installer (par exemple sous Windows dans `C:\Python26`). Sous Windows Vista, pour réussir l'installation, il ne faut pas laisser la case cochée par défaut "Install for all users", mais choisir d'installer seulement pour son compte d'utilisateur.

Ensuite, téléchargez Densidées sous forme d'un fichier zip à l'adresse http://code.google.com/p/densidees/. Décompressez-le, par exemple dans le dossier `C:\Densidees`.

Ce dossier contiendra alors en particulier :
- `Densidees.exe`, une interface graphique pour Windows qui permet de lancer le programme sur le texte voulu et de cliquer pour obtenir le résultat,
- `Densidees.py`, le code source du programme en Python, que vous pouvez lancer directement depuis la ligne de commande, et appeler automatiquement dans des scripts pour calculer la densité des idées d'un ensemble de plusieurs fichiers.

Depuis sa version 1.3, Densidées permet de charger directement un texte, et d'effectuer son étiquetage grammatical en faisant appel à TreeTagger par un clic sur le bouton *Taguer !* : ceci nécessite l'installation de :
- TreeTagger : télécharger le programme sur ftp://ftp.ims.uni-stuttgart.de/pub/corpora/tree-tagger-windows-3.2.zip et le décompresser sur le disque dur, par exemple dans `C:\TreeTagger`. Télécharger le fichier de paramètres français ftp://ftp.ims.uni-stuttgart.de/pub/corpora/french-par-linux-3.2.bin.gz et décompresser le fichier `french-par-linux-3.2.bin` qu'il contient dans le dossier `lib` de TreeTagger (par exemple `C:\TreeTagger\lib`). Renommer ce fichier en `french.par`. Il faudra alors indiquer l'adresse `C:\TreeTagger` dans Densidées.
- Perl : pour Windows, télécharger et installer Strawberry Perl (http://strawberryperl.com) par exemple.

3 Utilisation du programme

3.1 Sur un corpus oral

Un mode oral, décrit ci-dessous, est fourni dans Densidées. Toutefois il est également nécessaire, pour l'utilisation de ce mode, de prétraiter les corpus avec un parenthésage qui conduira aux résultats suivants :
- tous les mots entre crochets ne sont comptés ni comme mots ni comme propositions.
- tous les mots entre parenthèses sont comptés comme mots mais pas comme propositions.

Il faut donc utiliser les parenthèses et crochets pour les cas suivants :
- crochets "[]" : mots fragmentés, répétitions exactes, pauses remplies non lexicales ("pff", "bah", "hein", etc.), passages inaudibles.
- parenthèses "()" : idées répétées, phrases inachevées, marqueurs discursifs qui sont des mots en français ("bon", "bien", etc.), noms propres (personnes, villes).

Si la transcription est réalisée exclusivement pour un calcul de la densité des idées avec Densidées, on pourra omettre de transcrire les parties entre crochets pour gagner du temps.

3.2 Avec l'interface graphique sous Windows

Pour lancer Densidées sous Windows, il faut double-cliquer sur le programme `Densidees.exe` : la fenêtre montrée en figure 1 apparaît alors.

La première étape est d'étiqueter le texte avec TreeTagger. Une première possibilité, si vous avez installé le logiciel sur votre machine comme expliqué en section 2, est de coller (ou ouvrir) le texte que vous voulez étiqueter dans le cadre de gauche de Densidées, puis de cliquer sur le bouton *Taguer !*, après avoir pris soin de renseigner correctement l'adresse du dossier contenant TreeTagger.

Si vous n'avez pas installé TreeTagger, vous pouvez utiliser l'interface web disponible à l'adresse `http://cental.fltr.ucl.ac.be/treetagger/`. **Attention !** Avec cette interface, utilisez bien un clic droit sur le lien vers le résultat pour enregistrer le fichier sur votre machine. Si au contraire vous cliquez sur le lien pour afficher directement dans votre navigateur le texte étiqueté, il est possible que les accents n'apparaissent pas correctement, ce qui causera des erreurs de Densidées.

Après avoir collé le texte étiqueté par TreeTagger dans le cadre de gauche (ou bien ouvert un fichier TXT contenant un texte étiqueté par TreeTagger à l'aide du bouton *Ouvrir le texte*), il suffit de cliquer sur le bouton *Calculer !* pour voir apparaître le résultat dans le cadre de droite. De plus, le résultat s'enregistre dans un fichier portant le même nom que le fichier ouvert dans Densidées, suivi de l'extension `".di.txt"`.

Si rien ne s'affiche dans le cadre de droite au bout de 10 secondes, vous pouvez faire s'afficher les éventuels messages d'erreur en utilisant la ligne de commande comme indiqué ci-dessous. La commande à utiliser est en fait donnée par l'interface graphique de Densidées, au bas de la fenêtre, et vous pouvez la sélectionner, la copier, puis la coller dans la ligne de commande. Par exemple dans la figure 1 il s'agit de :
`"C:\Python26\Python.exe" "C:\These\densidees\Densidees\Densidees.py" "C:\These\densidees\Densidees\Test.txt"`.

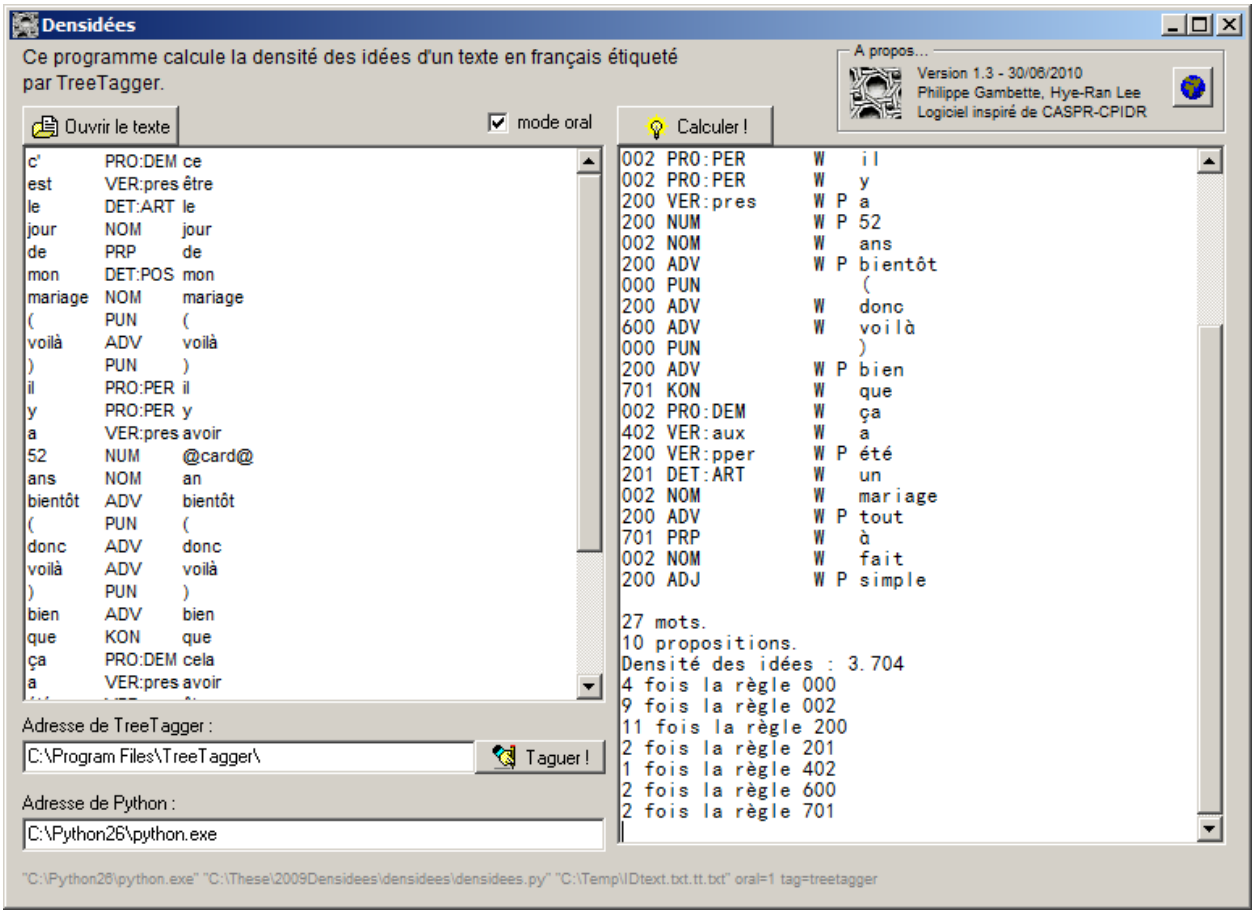

FIGURE 1 – Interface graphique de Densidées sous Windows.

3.3 Directement depuis la ligne de commande Windows

Commencez par ouvrir une fenêtre de ligne de commande en allant dans le *menu Démarrer*, *Exécuter*, en tapant alors `cmd` puis en appuyant sur *Entrée*.

Il faut alors taper une ligne de commande de ce type : `"C:\Python26\Python.exe" "C:\Densidees\Densidees.py" "C:\Densidees\Texte1.txt"`

Cette commande signifie qu'on va appeler le logiciel `Python.exe` pour lui demander de lancer le programme `Densidees.py` sur le fichier texte `Texte1.txt` qui contient le résultat de l'étiquetage TreeTagger d'un texte en français.

Vous pouvez utiliser l'option `oral=1` pour utiliser le calcul de densité des idées en mode oral (traitement des répétitions). Par exemple, la figure 2 montre le résultat de l'utilisation du programme en ligne de commande, avec activation du mode oral.

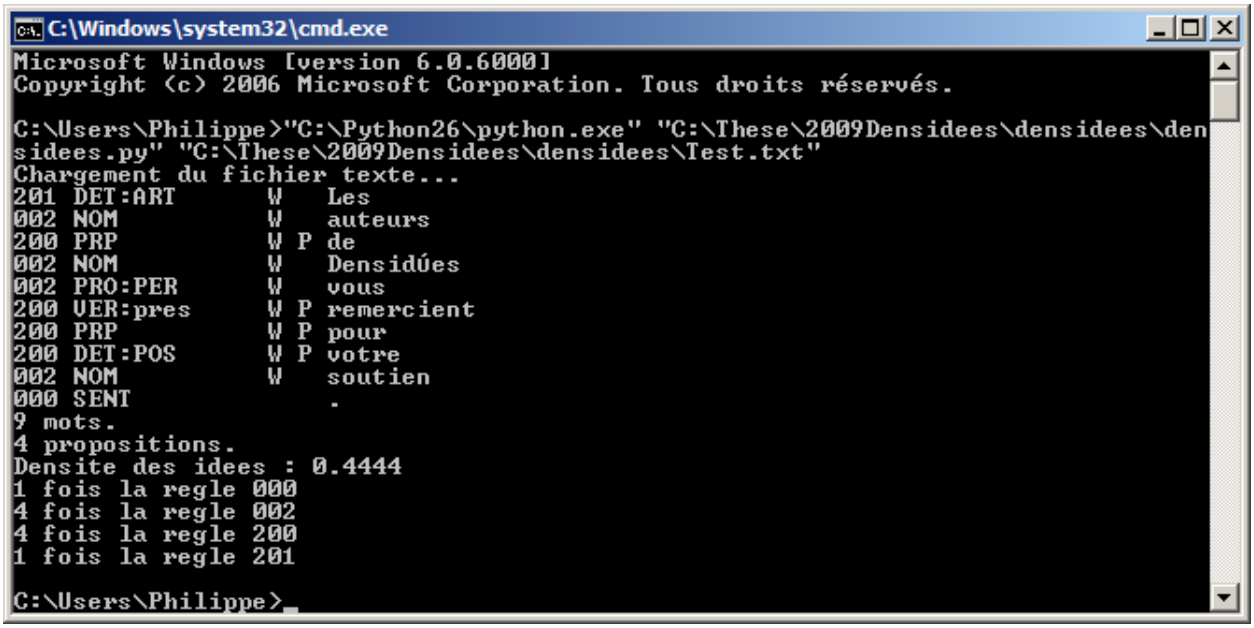

FIGURE 2 – Utilisation de Densidées en ligne de commande sous Windows.

3.4 Sur plusieurs fichiers : le mode “invisible”

Si vous voulez calculer la densité des idées de plusieurs fichiers, Densidées vous permet, en ligne de commande, de n’afficher que les données suivantes, séparées par des point-virgules :
– nom de fichier
– nombre de mots
– nombre de propositions
Il faut pour cela utiliser l’option `visible=0` comme montré en figure 3.

Cela vous permettra de construire très facilement un tableau de résultats, en enregistrant la sortie des fichiers dans un document au format CSV, que vous pourrez ensuite ouvrir dans le logiciel Excel. Pour cela, sous Windows, imaginons que nous voulons avoir dans un tableau Excel la densité des idées de trois textes `Texte1.txt`, `Texte2.txt` et `Texte3.txt`. On crée un fichier texte contenant les 3 commandes suivantes :

```
"C:\Python26\python.exe" "C:\Densidees\Densidees.py" visible=0 "C:\Densidees\Texte1.txt" > C:\Densidees\resultats.csv
"C:\Python26\python.exe" "C:\Densidees\Densidees.py" visible=0 "C:\Densidees\Texte2.txt" >> C:\Densidees\resultats.csv
"C:\Python26\python.exe" "C:\Densidees\Densidees.py" visible=0 "C:\Densidees\Texte3.txt" >> C:\Densidees\resultats.csv
```

et on l’enregistre avec l’extension “.bat”, par exemple dans `"C:\Densidees\script.bat"`. En double cliquant dessus, les trois commandes sont exécutées. La première crée un fichier `resultats.csv` (ouvrable dans Excel) où elle indique son résultat sur la première ligne, les deux suivantes ajoutent leur résultat sur les lignes suivantes. Une formule Excel permettra de calculer la densité des idées à partir des valeurs de la deuxième et troisième colonne (nombre de mots et nombre de propositions).

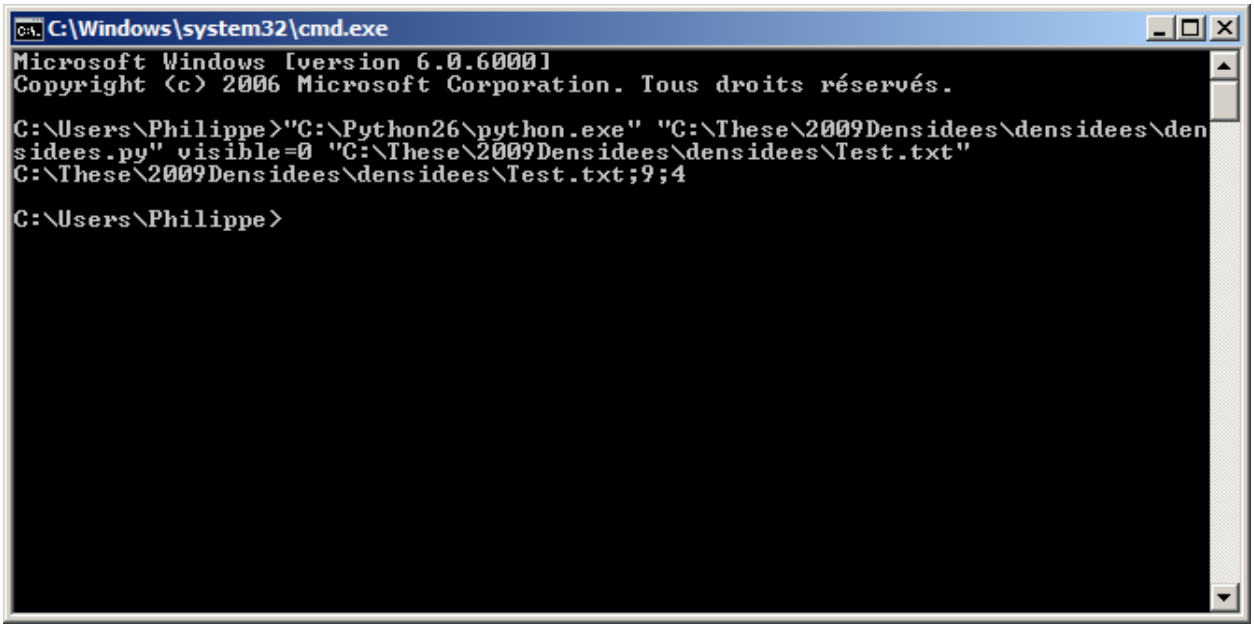

FIGURE 3 – Utilisation du mode "invisible" de Densidées en ligne de commande sous Windows.

3.5 Règles appliquées

Les règles suivantes sont appliquées pour étiqueter les éléments du texte étiqueté par TreeTagger en mots et propositions.

- **001 Interjections :**
  Interjections non reconnues par TreeTagger => pas mot, pas proposition
- **002 Ponctuation et symboles :**
  Signe de ponctuation, symbole => pas mot
- **020 Répétition ou correction d'un mot (mode oral) :**
  A A ou préfixe-de-A A => premier A : pas mot, pas proposition
- **023 Répétition ou correction de 2 mots (mode oral) :**
  A B A B ou préfixe-de-A préfixe-de-B A B => premier A et premier B : pas mot, pas proposition
- **024 Répétition ou correction de 3 mots (mode oral) :**
  A B C A B C ou préfixe-de-A préfixe-de-B préfixe-de-C A B C => premier A, premier B et premier C : pas mot, pas proposition
- **101 Est-ce que (mode oral) :**
  Rien n'est compté comme proposition
- **102 Clivages :**
  "c'est" + au plus 5 mots + "que" ou "qui" : "être" non compté comme proposition
- **200 Etiquetage basique des propositions :**
  Les tags correspondant à des propositions sont marqués comme propositions
  KON, NUM, DET*, PRP*, ADJ, PRO:POS, PRO:IND, ADV, VER*, PRO:REL
- **054 Déterminants démonstratifs (étiquetés pronoms démonstratifs) comptés comme proposition :**

"cet", "cette", "ces" -> comptés comme proposition
"ça" pas compté comme proposition

- **201 Déterminants non propositions :**
  lemme = "un" ou "le" ou "du" => pas proposition
  Attention à "du" = "de le", voir règle 202
- **202 Complément du nom introduit par "du" :**
  NOM du NOM => "du" est une proposition
- **203 Soit soit :**
  soit + 1 à 3 mots + soit : seul le premier "soit" est compré comme proposition
- **204 Conjonctions "ou" ou "et" superflues avant adverbe :**
  adverbes après "et" : "puis", "alors", "donc", "ensuite", "finalement" => "et" pas proposition
  adverbes après "ou" : "alors" "bien" => "ou" pas proposition
- **206 "de" non proposition :**
  "de" n'est pas proposition après "falloir", "agir", "arriver", "paraître"
- **207 "que" non proposition :**
  "que" n'est pas proposition après "falloir", "sembler", "arriver", "paraître"
- **208 Comparatifs :**
  "autant" ou "moins" ou "plus" + <3 mots + "que" : "que" non proposition
- **210 Oui et non (mode oral) :**
  oui et non : pas proposition
- **211 Négation :**
  "aucun" "guère" "jamais" "nul" "pas" "plus" "point" "que" "rien" précédé (à distance 1, 2 ou 3) par "ne" : seul "ne" proposition
- **212 Négation suivie de "de" :**
  "de" n'est pas une proposition si précédée par négation
- **213 Futur proche :**
  lemme="aller" + infinitif = futur proche : aller n'est pas une proposition
- **214 Si ... alors :**
  "si" + 1 à 9 mots + "alors" : ne pas compter "alors" comme proposition, seulement "si".
- **301 Verbes de liaison :**
  Verbe de liaison pas proposition si suivi d'un adjectif ou d'un adverbe
- **302 Verbe être suivi d'une préposition :**
  "être" non proposition si suivi d'une préposition
- **402 Auxiliaire :**
  AUX + VERBE => une seule proposition
- **405 Auxiliaire avec mot interposé :**
  AUX + mot + VERBE => une seule proposition
- **500 Passif :**
  participe passé + "par" => "par" non proposition item **509 "à" + infinitif :**
  "à" + infinitif => "à" non proposition item **510 Gérondif :**
  "en" + "participe présent" => "en" non proposition
- **512 Verbes suivis d'une préposition naturelle :**
  "à" non prop si précédé de aller, voyager
  "de" non prop si précédé de venir
- **600 Marqueurs discursifs (mode oral) :**

expressions qui ne sont pas proposition
- **601 Marqueur discursif avec "bien" (mode oral) :**
  "bien" n'est alors pas proposition
- **602 Marqueur discursif avec "donc" (mode oral) :**
  "donc" n'est alors pas proposition
- **701 Mots composés**
  "donc" n'est alors pas proposition
- **702 Mots composés avec "par"**
  expressions qui ne correspondent qu'à une seule proposition
- **703 Mots composés avec "avoir"**
  expressions qui ne correspondent qu'à une seule proposition

En mode oral, tout ce qui se trouve entre parenthèses n'est pas compté comme proposition. Le comptage des mots est en revanche effectué également à l'intérieur des parenthèses. Les parenthèses serviront donc à entourer des phrases incomplètes qui ne correspondent à aucune idée.

3.6 Correspondances entre règles CPIDR et règles Densidées

CPIDR	Densidées	Commentaire
000		
001		
002	002	
003		Succession de deux entiers regroupés en 1 : gérée par TreeTagger et Cordial
004		Fractions, pourcentages : gérés par TreeTagger et Cordial
020	020	
021		
022		
023	023	
050		
054	054	
101	101	Interrogation "est-ce que"
200	200	
201	201	
202	202	Désactivée dans CPIDR
203	203	
204	204	
206		
207		
210		
211		
212	211, 212	
213	213	
214	214	
225		
230		
301	301	
302	302	
310		
311		
401		
402	402	
405	405	
510	509,510	
511		
512	512	Désactivée dans CPIDR
610		
632		
634	600, 601, 602, 701, 702, 703	

3.7 Modification du code source

Vous pouvez modifier le code du programme Python, pour supprimer des règles, en ajouter ou en modifier.

Le principe est d'ouvrir le fichier `Densidees.py` dans un éditeur de texte adapté (par exemple Notepad2).

Les règles apparaissent alors au milieu du fichier. `text[i]` concerne le i-ième élément du texte, et contient plusieurs informations :
- `text[i]["word"]` : la forme graphique de cet élément dans le texte.
- `text[i]["tag"]` : l'étiquette donnée par TreeTagger à l'élément.
- `text[i]["lemma"]` : le lemme de l'élément, d'après la lemmatisation de TreeTagger.
- `text[i]["rule"]` : contient le numéro de la dernière règle appliquée pour déterminer si le mot est une proposition.
- `text[i]["isProp"]` : contient "P" si l'élément est une proposition, " " sinon.
- `text[i]["isWord"]` : contient "W" si l'élément est un mot, " " sinon.

4 Licence

4.1 Citation

Bien que Densidées soit un logiciel libre sous licence GPL, nous aimerions que vous fassiez référence à l'article suivant si vous l'utilisez dans une publication :

Hyeran Lee, Philippe Gambette, Constance Thuillier & Elsa Maillé. Densidées : calcul automatique de la densité des idées dans un corpus oral. *Actes de la douzième Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL 2010)*, http://halshs.archives-ouvertes.fr/halshs-00495768/fr/, 2010 [4].

Le poster suivant montre le lien entre la densité des idées calculée par Densidées et la maladie d'Alzheimer :

Hyeran Lee, Philippe Gambette & Melissa Barkat-Defradas. Utilisation de l'analyse textuelle automatique dans la recherche sur la maladie d'Alzheimer. Poster au Colloque international des jeunes chercheurs en Didactique des Langues et en Linguistique (CEDIL 2010), http://www.lirmm.fr/~gambette/2010LeeGambetteBarkatPoster.pdf, 2010 [3].

4.2 Licence

Densidées v.1.3 - 30/06/2010
http://code.google.com/p/densidees/

License, or (at your option) any later version.

Densidées is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU General Public License for more details.

You should have received a copy of the GNU General Public License along with Densidées. If not, see http://www.gnu.org/licenses/.

5 Historique des versions

2010/06/30 1.3 :
– appel automatique de TreeTagger depuis Densidées
– mode "invisible" permettant de n'afficher que le résultat, en ligne de commande
– calcul de la densité des idées comme ratio pour 10 mots selon la formule traditionnelle
– amélioration des règles 208, 301 et 701
2010/03/07 1.2 :
– prétraitement des auxiliaires si pas fait par TreeTagger
– 35 règles 001, 002, 020, 023, 024, 101, 102, 200, 054, 201, 202, 203, 204, 206, 207, 208, 210, 211, 212, 213, 214, 301, 302, 402, 405, 500, 509, 510, 512, 600, 601, 602, 701, 702, 703
2009/12/12 1.1 :
– interface graphique
– mode oral
– 27 règles 002, 020, 023, 024, 200, 054, 201, 202, 203, 204, 206, 207, 210, 211, 212, 213, 214, 301, 302, 402, 405, 500, 512, 600, 601, 602, 701
– affichage final du nombre de chacune des règles utilisées
2009/11/21 1.0 :
– 7 règles 002, 003, 200, 201, 301, 302, 402
– texte étiqueté par TreeTagger en entrée du programme

6 Remerciements

Nous remercions le LIRMM et l'association Contact, partenaires de la journée Osidmesh [1] à l'origine de la collaboration qui a débouché sur la conception de ce logiciel.

Elsa Maillé et Constance Thuillier ont également contribué à l'ajout de règles dans ce logiciel, ainsi qu'à son évaluation par l'étiquetage manuel de corpus de test.

L'école doctorale I2S, l'ATALA, l'école doctorale 58 et le laboratoire Praxiling ont participé au financement de la participation d'Hyeran Lee au colloque RECITAL 2010 pour présenter le logiciel à la communauté francophone de traitement automatique du langage.

Si vous vous demandez d'où provient l'icône de Densidées, allez voir sur cette page http:

1. http://www.lirmm.fr/~semindoc/Osidmesh.html

//philippe.gambette.free.fr/Photos/200907_Liban.htm, ou pensez à lever les yeux le jour où vous visiterez Beyrouth.

Références

[1] Cati Brown, Tony Snodgrass, Susan J. Kemper, Ruth Herman, and Michael A. Covington. Automatic measurement of propositional idea density from part-of-speech tagging. *Behavior Research Methods*, 40(2):540–545, 2008. http://www.ai.uga.edu/caspr/BrownSnodgrassKemperHermanCovington2008.pdf, free software CPIDR available from http://www.ai.uga.edu/caspr.

[2] Walter Kintsch. *The representation of meaning in memory*. John Wiley & Sons, 1974.

[3] Hyeran Lee, Philippe Gambette, and Melissa Barkat-Defradas. Utilisation de l'analyse textuelle automatique dans la recherche sur la maladie d'alzheimer, 2010. Poster au Deuxième Colloque international des jeunes chercheurs en Didactique des Langues et en Linguistique (CEDIL2010), http://www.lirmm.fr/~gambette/2010LeeGambetteBarkatPoster.pdf.

[4] Hyeran Lee, Philippe Gambette, Elsa Maillé, and Constance Thuillier. Densidées : calcul automatique de la densité des idées dans un corpus oral. In *Actes de la douzième Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL 2010)*, 2010. http://halshs.archives-ouvertes.fr/halshs-00495768/fr/.

[5] Althea Turner and Edith Greene. The construction and use of a propositional text base, 1977. Tech. Report 63, Boulder: University of Colorado, Institute for the Study of Intellectual Behavior, http://ics.colorado.edu/techpubs/pdf/77-63.pdf.