Utilisation de l'analyse textuelle automatique dans la recherche sur la maladie d'Alzheimer

Hye-Ran Lee [1], *Philippe Gambette* [2], *Melissa Barkat-Defradas* [1]

[1] Laboratoire Praxiling, UMR 5267 – CNRS / Université Montpellier 3
[2] LIRMM, UMR 5506 – CNRS / Université Montpellier 2

1. Introduction

La densité des idées (DI) est le ratio entre le nombre de propositions sémantiques et le nombre de mots dans un texte. Ainsi, elle reflète la qualité informative des propositions langagières d'un texte. La méthode utilisée pour mesurer la DI est l'analyse prédicative (Kintsch, 1974). L'analyse prédicative permet d'extraire les propositions sémantiques dans un texte. Une proposition est constituée d'un prédicat et d'un ou plusieurs argument(s). Cette méthode a montré son utilité dans plusieurs domaines de la psycholinguistique appliquée : la compréhension du texte (Kintsch et al., 1973 ; Kintsch, 1998), la mémoire (Thorson et al., 1984), la qualité de prise de note des étudiants (Takao et al., 2002), le vieillissement (Kemper et al., 2001), la schizophrénie (Covington et al., 2007), la distinction de documents (Covington, 2009).

Aussi, la DI est reconnue comme un indicateur pertinent des fonctions intellectuelles des sujets. La grande étude longitudinale américaine menée sur les membres d'une communauté de religieuses a montré qu'une faible DI observée dans les autobiographies écrites au moment où elles ont formulé leurs v?ux entre 18 et 32 ans est fortement corrélée au développement de la maladie d'Alzheimer cinquante ans plus tard (Snowdon, 1996 ; Snowdon et al., 1997 ; Snowdon et al. 1999 ; Snowdon et al., 2000). Aussi, de nombreuses études ont montré la diminution de la DI dans le discours des patients atteints de la maladie d'Alzheimer (Kemper et al., 1993 ; Lyons et al., 1994 ; Kemper et al., 2001 ; Riley et al., 2005). De même, la DI permet de mesurer l'avancée de la maladie (Kemper et al., 2001 ; Lyons et al., 1994 ; Chand et al, 2007 ; Lee et al., 2009). Ainsi, la DI est un champ d'investigation prometteur pour le diagnostic et la mesure de l'évolution de la maladie d'Alzheimer.

La difficulté majeure pour inclure la DI dans les critères linguistiques pour mesurer la performance langagière des patients atteints de la maladie d'Alzheimer est que son analyse est longue, coûteuse, et fastidieuse. L'analyse manuelle parfois subjective fait donc souvent obstacle à la prise en compte de la DI dans l'étude linguistique de la maladie d'Alzheimer.

Ainsi, il semble indispensable et novateur de développer un outil informatique qui permettrait de mesurer rapidement et objectivement la DI pour la perspective de mise en pratique clinique. Ce besoin nous a poussé à concevoir le logiciel *Densidées*. Densidées est un outil informatique qui calcule automatiquement une approximation de la DI d'un texte en français étiqueté grammaticalement. Cet outil est une adaptation française du logiciel CPIDR 3.2 (Brown et al., 2008). Densidées est un logiciel libre sous licence GPL, téléchargeable sur le site :
http://code.google.com/p/densidees/
Ecrit en Python, il peut être utilisé directement depuis la ligne de commande, appelé automatiquement dans des scripts pour calculer la DI d'un ensemble de plusieurs fichiers. Une interface graphique simple est également proposée sous Windows (Figure 1).

L'idée principale du programme est qu'une proposition correspond typiquement à un verbe, un adjectif, à un adverbe, une préposition, ou une conjonction (Snowdon et al., 1996). Le nom n'est pas une proposition, et les informations sur le verbe comme le temps, l'aspect, la voix, la modalité, ne sont pas prises en compte comme proposition. La première étape est donc d'étiqueter le texte à analyser avec un étiqueteur grammatical, en l'occurrence *TreeTagger* (Schmid, 1994), ceci est possible par exemple avec l'interface web disponible à l'adresse :
http://cele.nottingham.ac.uk/~ccztk/treetagger.php.

Une fois étiqueté, le texte doit simplement être collé dans un fichier, ou dans l'interface graphique de Densidées. Après un clic sur le bouton *Calculer !*, le résultat s'affiche dans le cadre de droite. Pour chaque mot, *W* signifie que le mot analysé rentre dans le nombre total de mot et *P* indique une proposition. À la fin de ce cadre, le nombre total de mots, le nombre total de propositions dans le texte soumis à l'analyse sont donnés ainsi que la DI.

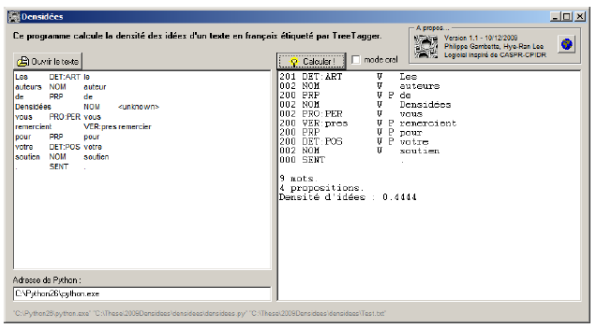

Figure 1. Interface graphique de Densidées sous Windows

Densidées comprend 25 règles dont 7 pour analyser une transcription orale.

2. Méthodes

Pour mesurer la fiabilité de Densidées et voir la relation entre la DI et la maladie d'Alzheimer, nous examinerons les

discours oraux d'une vingtaine de sujets âgés sains et d'une vingtaine de patients atteints de la maladie d'Alzheimer.

Ces deux groupes de sujets sont appariés par âge, sexe, et niveau d'études. Pour obtenir les données orales, nous avons mené des entretiens individuels semi-dirigés. Ces entretiens ont été enregistrés et transcrits. Les données orales ont été soumises à la fois à l'analyse automatique de Densidées et l'analyse manuelle de la DI selon des règles basées sur les études précédentes (Kintsch et al., 1973 ; Kintsch, 1974, Turner et al., 1977 ; Kintsch et al., 1978 ; Le Ny, 1979 ; Le Ny, 1987 ; Tiberghien et al., 1992 ; Kemper et al., 1993 ; Lyons et al., 1994 ; Ghiglione, 1995 ; Coirier et al., 1996 ; Snowdon et al., 1996 ; Rondal, 1997 ; Kintsch, 1998 ; Tiberghien, 1999 ; Duong et al., 2000 ; Kemper et al., 2001 ; Rouet, 2001 ; Blanc et al., 2005 ; Brown et al., 2008). Pour équilibrer le corpus à comparer, seules les 20 premières lignes de chaque transcription ont été analysées.

Le taux d'accord entre l'analyse manuelle des différents examinateurs sera discuté. Nous examinerons également la possible influence des facteurs sociaux comme l'âge, le sexe, et le niveau d'étude sur le résultat de la DI.

3. Conclusion

Bien que Densidées ne permette pas une analyse qualitative fine du discours, il donne un résultat rapide et objectif de la DI. Pour assurer la fiabilité de cet outil, nous étudions l'utilisation de l'étiqueteur commercial *Cordial*.

Nos résultats préliminaires montrent que la DI est plus faible dans la production langagière des patients souffrant de la maladie d'Alzheimer que celle des personnes âgées saines. La mise au point de cet outil permet d'élargir le champ de recherche sur la maladie d'Alzheimer en ouvrant d'autres possibilités de la détection précoce et différentielle de cette maladie via l'analyse linguistique.

Références

Brown, C., Snodgrass, T., Kemper, S., Herman, R. & Covington, M. (2008). Automatic measurement of propositional idea density from part-of-speech tagging. *Behavior Research Methods*, 40(2), 540-545.

Chand, V. & Bonnici, L. (2007). Quantifying language degradation in Alzheimer's disease. *New Ways of Analyzing Variation 36*.

Coirier, P., Caonac'h, D. & Passerault, J. M. (1996). Psycholinguistique textuelles, Paris : Armand Colin.

Covington, M. (2009). Idea Density: A potentially informative characteristic of retrieved documents, IEEE SoutheastCon.

Covington, M., Riedel, W., Brown, C., He, C., Morris, E.,Weinstein, S., et al. (2007). Does ketamine mimic aspects of schizophrenic speech? *Journal of Psychopharmacology*, 21, 338-346.

Ghiglione, R., Kehenbosch, C. & Landré, A. (1995). L'analyse cognitive-discursive, Presse Universitaire de Grenoble.

Kemper, S., Greiner, L., Marquis, J., Prenevost, K. & Mitzner, T. (2001). Language decline across life span: findings from the nun study, *Psychology and Aging*, 16(2), 227-239.

Kemper, S., LaBarge, E., Ferraro, R., Cheung, H. & Storandt, M. (1993). On the preservation of syntax in Alzheimer's disease, *Archives of Neurology*, 50, 81–86.

Kintsch, W. & Keenan, J. (1973). Reading rate and retention as a function of the number of propositions in the base structure of sentences, *Cognitive Psychology*, 5 (3), 257-274.

Kintsch, W. & Van Dijk, T. (1978). Toward a model of text comprehension and production, *Psychological Review*, 85, 363-394.

Kintsch, W. (1974). The representation of meaning in memory, Hillsdale, NJ: Erl-baum.

Kintsch, W. (1998). Comprehension: a paradigm for cognition, Cambridge, MA: Cambridge University Press.

Le Ny J. F. (1979). La sémantique psychologique, Paris : PUF.

Le Ny, J. F. (1987). La sémantique psychologie, In J. A. Rondal & J. P. Thibaut (Eds.), *Problèmes de psycholinguistique*, 13-42. Bruxelles : Pierre Mardaga.

Lee, H. & Barkat-Defradas, M. (2009). La densité des idées : un modèle d'analyse du discours pertinent pour le diagnostic précoce de la maladie d'Alzheimer ? 8ème *Rencontres Jeunes Chercheurs en Parole*, Avignon, 16-18 Novembre.

Lyons K., Kemper S., LaBarge E., Ferraro F. R., Balota D. & Storandt M. (1994). Oral language and Alzheimer's disease: A reduction in syntactic complexity, *Aging and Cognition*, 1(4), 271–281.

Mizner, T. (2001). Langage decline across life span : findings from the nun study, *Psychology and Aging*, 16(2), 227-239.

Riley, K., Snowdon, D., Desrosiers, M., Markesbery, W. (2005). Early life linguistic ability, late life cognitive function, and neuropathology : findings from the Nun Study. *Neurobiology of Aging*, 26, 341-347.

Rondal, J. R. (1997). L'évaluation du langage, Bruxelles : Mardaga.

Rouet, J. F. (2001). Les activités documentaires complexes : aspects cognitifs et développementaux, *Rapport pour l'Habilitation à Diriger des Recherches*, Poitiers : Laboratoire Langage et Cognition.

Schmid, H. (1994). Probabilistic Part-of-Speech tagging using decision trees. In D., Jones & H., Somers (Eds.), *New Methods in Language Processing*, 154-164, London : UCL Press.

Snowdon, D. (1997). Aging and Alzheimer's disease : lessons from the Nun Study, *Gerontologist*, 37, 150-156.

Snowdon, D., Kemper, S., Mortimer, J., Greiner, L., Wekstein, D. & Markesbery, W. (1996). Linguistic ability in early life and cognitive function and Alzheimer's disease in late life : findings from the Nyun Study, *JAMA*, 275, 528-532.

Snowdon, D., Geiner, L. & Markesbery, W. (2000). Linguistic Ability in Early Life and the Neuropathology of Alzheimer's disease: Findigs from the Nun Study, *Annals of the New York Academy of Sciences*, 903, 34-38.

Snowdon, D., Greiner, L., Kemper, S., Nanayakkara, N. & Mortimer, J. (1999). Linguistic ability in early life and longevity: Findings from the Nun Study. In J. M., Robine, B., Forette, C., Franschesci, & M., Allard (Eds.), *The paradoxes of Longevity,* 103-113. Amsterdam: Springer.

Takao, A., Prothero, W. & Kelly, G. (2002). Applying argumentation analysis to assess the quality of university oceanography students' scientific writing. *Journal of Geoscience Education*, 50. 40-48.

Thorson, E. & Snyder, R. (1984). Viewer recall of television commercials: Prediction from the propositional structure of commercial scripts. *Journal of Marketing Research*, 21, 127-136.

Tiberghien, G. (1999). La mémoire, In *Neuropsychologie humaine*, X. Seron, & M. Jeannerod (Dir.), 255-281, Bruxelles : Mardaga.

Tiberghien, G., Roulin, J. L. & Beauvois, J. L. (1992). Manuel d'études pratiques de psychologie 2: études pratiques, Paris: PUF.

Turner, K. & Greene, E. (1977). The construction and use of a propositional text base, *Technical report*, 63, University of Colorado.