Exercice avec TermoStat : niveau I

Les *dépouilleurs terminologiques* ou *extracteurs de termes* permettent d'identifier des unités clés dans un texte ou une collection de textes. Ce dépouillement peut aider un traducteur ou autre langagier à se familiariser avec les thèmes et le contenu d'un texte, et peut aussi servir de point de départ pour des recherches terminologiques. Cela dit, il est essentiel de reconnaître que l'identification automatique d'unités clés est un processus qui se base entièrement sur des critères formels, et qui ne peut qu'identifier des *candidats termes*. Une évaluation humaine est toujours nécessaire pour évaluer le caractère terminologique des candidats identifiés et leur pertinence dans un contexte donné.

L'identification des *candidats termes* repose généralement directement ou indirectement sur la fréquence des occurrences d'unités, sur leur forme, ou sur les deux à la fois. Ces méthodes d'extraction terminologique sont caractérisées respectivement comme des approches *statistiques, linguistiques* ou *hybrides*. Essentiellement, l'hypothèse de départ suppose que des unités qui sont particulièrement fréquentes, qui ressemblent aux patrons terminologiques typiques (par exemple, NOM + PRÉPOSITION + NOM, ou NOM + ADJECTIF pour le français) ou qui satisfont aux deux conditions sont susceptibles d'être des termes.

TermoStat, créé par Patrick Drouin de l'Observatoire de linguistique Sens-Texte (OLST) de l'Université de Montréal, est un dépouilleur terminologique en ligne qui utilise une méthode hybride — c'est-à-dire qui intègre des méthodes statistique et linguistique — pour identifier des candidats termes. Il prend en compte non seulement la structure des unités (faisant appel à un programme qui s'appelle un *étiqueteur morphosyntaxique* pour cibler des substantifs, des adjectifs et des unités complexes dont ces derniers font partie comme autant de candidats termes), mais il considère aussi les fréquences relatives des unités identifiées dans un corpus d'analyse (le texte/les textes à dépouiller) et un corpus de référence (une collection de textes journalistiques). TermoStat permet ainsi d'identifier des candidats termes complexes et simples à l'aide d'un seul processus de dépouillement.

Drouin utilise le terme *spécificité* pour désigner l'unité de mesure par défaut qui permet d'identifier des candidats termes en calculant la différence entre les fréquences relatives des unités candidates dans les corpus d'analyse et de référence. TermoStat peut toutefois utiliser d'autres mesures pour identifier des termes. Ainsi, il est possible de comparer les résultats de différentes approches pour évaluer leur performance dans un contexte donné.

Pour faire un décompte exact des occurrences de chaque candidat terme, TermoStat utilise un processus qui s'appelle la *lemmatisation*; il ramène à leur forme de base les formes fléchies des candidats termes (par exemple, les formes *ancestraux*, *ancestrale*, et *ancestrales* seraient ramenées à la forme *ancestral* et *droits* à *droit*) et chaque occurrence d'une de ces formes est comptée comme une occurrence de cette forme de base. Pour cette raison, les résultats affichés incluent deux champs : le *candidat au regroupement* (qui est la forme de base ou la suite des formes de base identifiée(s) par TermoStat), et les *variantes orthographiques* (qui sont les formes observées dans le texte lui-même).

Pour en savoir plus sur ces techniques d'extraction terminologique, vous pouvez consulter, entre autres, *Initiation à la traductique (2e édition)* (L'Homme 2008), *Computer-Aided Translation Technology* (Bowker 2002) et *La terminologie : Principes et techniques* (L'Homme 2004). Vous pouvez en savoir plus sur le fonctionnement de TermoStat dans l'article Drouin, P. (2003) « Term Extraction using non-technical corpora as a point of leverage », *Terminology* 9(1): 99–115, et dans d'autres articles sur TermoStat, indiqués dans la page Web www.mapageweb.umontreal.ca/drouinp/. (La thèse de Drouin, qui porte sur la création de l'outil, est aussi accessible sur cette page.)

Les exercices qui suivent visent à vous aider à :

- utiliser le dépouilleur terminologique TermoStat;
- évaluer les résultats du dépouillement au niveau de
    - la présentation des résultats,
    - la nature des unités identifiées,
    - la forme de celles-ci,
    - la précision et le rappel du dépouilleur (c.-à-d., bruit et silences dans les résultats), et
    - les mesures prises pour améliorer les résultats ainsi que leurs effets; et
- explorer des fonctions complémentaires de TermoStat.

Préparation

1. Créez un sous-répertoire dans le répertoire *Mes documents*, dans lequel vous pourrez stocker les fichiers utilisés aux fins de ces exercices. (Pour des instructions à ce sujet, consultez le document *Créer un sous-répertoire* sur le site de la CERTT (**ACCÈS PAR NOM D'OUTIL >WINDOWSXP> CRÉER UN SOUS-RÉPERTOIRE**).)
2. Téléchargez un extrait de l'entrée de Wikipedia sur la cosmologie en format .txt à partir du site CERTT (**FICHIERS À TÉLÉCHARGER > WIKIPEDIACOSMOLOGIE (EXTRAITS)**), ainsi qu'une copie de ce tutoriel en format .txt (**FICHIERS À TÉLÉCHARGER > TUTORIEL_TERMOSTAT**)
3. Parcourez le texte de Wikipédia. Identifiez 5 à 10 termes dans le document que vous trouvez pertinents, et prenez-les en note.
4. Ouvrez le fureteur de votre choix, par exemple, Internet Explorer ou Mozilla Firefox (à partir du raccourci sur le Bureau ou du menu **DEMARRER**).
5. Ouvrez la page de TermoStat Web : http://olst.ling.umontreal.ca/~drouinp/termostat_web.
6. Inscrivez-vous et ouvrez une session :
    a. Dans le coin inférieur gauche de la fenêtre d'inscription, localisez l'hyperlien **INSCRIPTION** et cliquez dessus.

b. Entrez les informations demandées et cliquez sur le bouton **INSCRIPTION** une fois terminé. Vous serez automatiquement redirigé vers la page d'accueil,

NOTE : Assurez-vous de prendre en note votre nom d'utilisateur et mot de passe.

c. Entrez votre nom d'utilisateur et votre mot de passe dans les champs indiqués et cliquez sur le bouton **OUVRIR SESSION**.

Identification de candidats termes

1. Cliquez sur le bouton **PARCOURIR** et sélectionnez le fichier à dépouiller (celui que vous venez de télécharger).

NOTE : TermoStat Web ne peut traiter que des corpus de texte en format *Texte brut* (.txt). Assurez-vous que les fichiers que vous analysez sont sauvegardés dans ce format, ou convertissez-les à l'aide de Word ou d'un autre traitement de texte (FICHIER/BOUTON OFFICE > ENREGISTRER SOUS…).

2. À partir de la liste déroulante *Langue*, sélectionnez la langue du fichier (en l'occurrence, **FRANÇAIS**).
3. À droite du champ *Extraction*, cochez les options qui détermineront les méthodes d'extraction pour identifier les termes. (Pour ce premier essai, il est suggéré d'utiliser d'abord les options par défaut : gardez les deux cases radio *Termes simples* et *Termes complexes nominaux* cochées, et dans la liste déroulante à droite de *Termes simples*, gardez la case *Noms* cochée. Notez néanmoins les autres options disponibles.)
4. Cliquez sur le bouton **LANCER L'ANALYSE** pour démarrer l'extraction.

Évaluation de l'extraction

1. Une fois l'analyse terminée, parcourez les résultats de l'onglet *Liste des termes* pour vous familiariser avec les données fournies et l'organisation de celles-ci.

a. La première colonne (*Candidat regroupement*) affiche la forme lemmatisée (de base) identifiée par TermoStat (c'est-à-dire, l'unité ou la suite d'unités qui constituent le candidat dans sa forme canonique après le processus d'étiquetage utilisé par TermoStat).

NOTE : Si seule la forme plurielle d'un candidat terme est présente dans le fichier analysé, elle sera affichée comme telle sous cette colonne.

b. La colonne suivante (*Fréquence*) indique la fréquence du candidat terme identifié dans le fichier analysé.

c. La troisième colonne (*Score (Spécificité)*) affiche le résultat du calcul de *spécificité*, que TermoStat utilise comme unité de mesure par défaut.

NOTE : La valeur de spécificité résulte de la comparaison entre la fréquence de l'unité dans le texte analysé (*corpus d'analyse*) et un corpus de textes généraux (*corpus de référence*). Plus l'indice de spécificité est élevé, plus l'unité est propre au texte (c'est-à-dire, y est particulièrement fréquente) et plus celle-ci est considérée comme étant susceptible d'être un terme.

En cliquant sur le signe plus  sous *Spécificité*, il est possible de réordonner les résultats selon d'autres méthodes de calculs. Pour plus de renseignements sur les différentes approches possibles, cliquez sur l'hyperlien **AIDE** dans le coin supérieur droit de la fenêtre, pour consulter le *Guide de l'utilisateur* : http://olst.ling.umontreal.ca/~drouinp/termostat_web/doc_termostat/doc_termostat.html).

d. La prochaine colonne (*Variantes orthographiques*) affiche les formes de l'unité identifiées par l'outil dans le texte lui-même (les formes non lemmatisées, certains types de variantes orthographiques);

e. La dernière colonne (*Matrice*) affiche les catégories grammaticales des unités qui composent le candidat terme.

2. Cliquez sur un des termes dans la liste *Candidat de regroupement*. La fenêtre *Contextes* s'ouvre pour vous permettre de consulter les différents contextes d'énonciation dans lequel le candidat terme a été identifié, que ce soit au niveau d'une phrase (sous l'onglet *Phrase*) ou des chaînes de caractères adjacentes (sous l'onglet *Concordance*).

   a. Quelle utilité voyez-vous à cette fonction de TermoStat? Cette information peut-elle vous aider à déterminer s'il s'agit bien d'un terme?

   b. Quelles autres informations à propos du candidat terme pouvez-vous identifier à l'aide de cette fonction?

   c. Fermez la fenêtre *Contextes* lorsque vous aurez terminé.

3. De retour à l'onglet *Liste de termes*, cliquez sur les en-têtes des colonnes pour trier les résultats en fonction de différents critères. Évaluez les différences observées au niveau de l'ordre des candidats termes identifiés.

   a. Y en a-t-il qui sont identifiés comme intéressants selon un critère (par exemple, la spécificité), mais pas selon un autre (par exemple, la fréquence)?

   b. Quelle est l'utilité de trier les résultats selon la colonne *Candidat regroupement*? Quand et dans quels buts serait-il utile de le faire ? Croyez-vous que l'utilité de ce tri serait la même dans une autre langue telle que l'anglais?

4. Évaluez les résultats de l'extraction selon les critères établis en introduction :

   a. Quelle est la forme des unités identifiées? Leur catégorie grammaticale ou leur structure?

   b. Quelle est la fréquence des unités identifiées? Les candidats identifiés sont-ils toujours fréquents?

   c. Quelle est la précision de l'extraction? Combien des candidats proposés sont à votre avis vraiment des termes? Des candidats qui ne sont pas, strictement parlant, des termes, peuvent-ils néanmoins être utiles pour des traducteurs?

   d. Quel est le rappel de l'extraction? Les résultats présentent-ils tous les termes que vous aviez identifiés dans le texte? Si ces derniers apparaissent dans la liste de TermoStat, quel est leur rang? Ce rang, correspond-il à celui que vous lui auriez attribué selon l'importance du terme? S'ils n'apparaissent pas dans la liste de TermoStat, pourquoi ont-ils été exclus, à votre avis?

   e. D'après vous, l'identification des formes de termes complexes par TermoStat peut-elle présenter des difficultés? Quelles pourraient-être les sources de ces difficultés? Malgré les problèmes, ces résultats sont-ils utiles?

   f. Quel est l'effet de la lemmatisation sur la présentation des résultats? Sur la forme proposée pour le candidat? Sur la mesure de la fréquence?

   g. La présentation des résultats est-elle efficace et conviviale?

   h. Voyez-vous d'autres complications avec les candidats proposés par TermoStat? En regardant ces candidats ou leurs contextes, pouvez-vous identifier les sources de ces difficultés?

Évaluation des fonctions complémentaires

Dans cette section, vous pouvez explorer rapidement les autres onglets de la page des résultats, qui présentent a) les candidats termes extraits en différents formats, b) les liens entre différents candidats, et c) les liens entre les candidats et certaines autres unités dans le texte analysé.

1. L'onglet *Nuage* présente les 100 termes dont la valeur de spécificité est la plus élevée. Leur présentation visuelle varie en taille en fonction de cette valeur.
    a. Cliquez sur un des termes au hasard. Que se passe-t-il?
    b. Fermez la fenêtre *Contextes* une fois que vous aurez terminé de la consulter.
2. L'onglet *Statistiques* affiche le nombre de candidats retenus pas TermoStat selon leur structure (les catégories grammaticales qui composent le terme).
    a. Cliquez sur le chiffre à droite de quelques-unes des matrices identifiées par TermoStat. Un encadré affiche alors certains des candidats termes de cette catégorie.
    b. Cette fonction facilite-elle le regroupement et l'identification possible des termes simples ou complexes présents dans le texte analysé?
    c. Le pourcentage indiqué à droite du chiffre peut-il, lui aussi, présenter une information utile au traducteur quant à la terminologie d'un domaine donné?
3. L'onglet *Structuration* présente un tableau des candidats termes simples, et, dans la colonne de droite, de candidats termes complexes qui les incluent. (Par exemple, le terme *modèle* renvoie à deux autres candidats termes complexes : *modèle cosmologique* et *modèle standard*.)
    a. D'après vous, cette manière de présenter les liens entre candidats termes facilite-elle l'identification possible des termes simples ou complexes présents dans le texte analysé?
    b. À partir de l'onglet *Structuration*, cliquez sur l'icône jaune à droite du terme *modèle*.
    c. La fenêtre *Décomposition* s'ouvre, présentant les différents types de relation de ce terme avec d'autres.
    d. Cliquez ensuite sur l'icône rouge et bleue à droite de *modèle*, pour visualiser la représentation graphique (graphe) de ces relations.
    e. Fermez les différentes fenêtres qui se sont ouvertes à cette étape pour revenir à l'onglet *Liste des termes*.
4. L'onglet *Bigramme* présente les couples verbe + nom qui sont souvent observés dans le texte analysé.
    a. Lancez une nouvelle analyse, cette fois du fichier *Tutoriel_Termostat.txt* (**FICHIERS À TÉLÉCHARGER > TUTORIEL_TERMOSTAT**).
    b. Cliquez sur l'onglet *Bigrammes*. L'onglet s'ouvre et affiche un couple verbe + nom : *cliquer* et *bouton*.
    c. Pour voir la décomposition et la représentation graphique de ce couple, cliquez sur les icônes et , tel qu'indiqué ci-dessus.
    d. Quelles informations utiles sur les candidats termes peuvent être mises en évidence par cet affichage? Pour qui et sous quelles conditions pensez-vous qu'elles pourraient être utiles?

Dernières étapes

1. Pour sauvegarder les résultats, vous pouvez cliquer sur l'icône d'enregistrement () dans le coin supérieur droit de la page des résultats. Cette fonction vous permet de sauvegarder le fichier en format .txt avec des tabulations entre les colonnes, et de l'ouvrir par la suite dans un tableur tel que Microsoft Excel, par exemple.

NOTE : Si vous avez terminé avec l'analyse d'un corpus et désirez en analyser un autre, cliquez sur l'hyperlien *Corpus* dans le coin supérieur gauche de la page des résultats. Vous reviendrez à la page d'accueil de TermoStat, où il sera possible d'aller chercher un autre fichier. Sur cette même page, dans l'encadré *Mes corpus* au bas de l'écran, vous pouvez aussi accéder à une liste de fichiers déjà analysés (si vous en avez). Vous pourrez ainsi revoir les résultats d'un dépouillement sans devoir resoumettre votre fichier, simplement en cherchant dans la liste le nom du fichier que vous avez soumis à l'analyse. Cela dit, les fichiers peuvent disparaître de façon imprévisible, donc si vous voulez garder une copie, le plus sûr est d'en faire une vous-même.

Notez que vous avez également accès aux fichiers déjà analysés en format texte brut à partir des liens qui apparaissent sous le titre *Mes documents* sur la page TermoStat.

Questions de réflexion

1. À la lumière de ces résultats, pouvez-vous déterminer des points forts et faibles de cette approche à l'extraction dans différents contextes?
2. Trouvez-vous que TermoStat est plus ou moins facile à utiliser ou plus ou moins convivial que d'autres extracteurs que vous connaissez? Pourquoi?
3. Dans cet essai, les *Variantes orthographiques* identifiées par TermoStat incluent les formes singulières et plurielles de noms ainsi que les formes accordées des adjectifs. Quelles autres sortes de variations orthographiques ou autres pourraient se trouver dans un texte? Quels effets pourraient-elles avoir sur les résultats d'une extraction par un outil comme TermoStat, à votre avis? (Vous pouvez tester votre hypothèse en soumettant un autre fichier qui contient des variations à l'analyse et en regardant les résultats.)
4. Croyez-vous que le traitement automatique des langues peut poser d'autres défis (par exemple, parmi ceux décrits dans L'Homme 2008, chapitre 2), reliés à la forme des candidats ou à leur environnement, et qui seraient pertinents pour TermoStat? Lesquels et pourquoi?

Tutoriel mis à jour par Cheryl McBride / Joanne Desroches