De l'intérêt de lister les collocations [nom de grandeur , argument (de la grandeur) , unité de mesure] en japonais

Raoul Blin

De l'intérêt de lister les collocations [nom de grandeur , argument (de la grandeur) , unité de mesure] en japonais

2015/07/13

Raoul Blin

CNRS-CRLAO

Nous présentons succintement la catégorie des noms de grandeur (NG) et expliquons lesquelles de leurs caractéristiques il serait intéressant de faire figurer dans un lexique. Nous proposons aussi quelques pistes méthodologique pour collecter automatiquement ces données.

1 Les noms de grandeur (présentation générale)

Les noms de grandeurs sont une classe particulière de noms communs, qui prennent un argument et renvoient une valeur numérique. Par exemple, la « hauteur » s'applique nécessairement à un objet (l'argument) et renvoie en français une valeur numérique:

Ex.1 La hauteur de la Tour Eiffel est de 300 mètres.

Ces noms apparaissent dans différentes langues, dont le français, japonais et anglais:

Ex.2 *eferu tou no takasa ha 300 meetoru.*
Eiffel tour DET hauteur TH 300 m
« La hauteur de la Tour Eiffel est de 300 m. »

Au moins en japonais et en français, ces noms se distinguent des autres noms communs par un ensemble spécifique de propriétés distributives (voir pour le japonais: (Blin 2013b)). Discriminer ces noms des autres noms communs est donc nécessaire pour les analyses syntaxiques (automatiques). La distinction est aussi importante pour l'interprétation. Prenons le cas des noms de grandeurs composés japonais de la forme[1] <compteur individuel+*suu/zuu*> ("nombre de X"). Par exemple *nin-zuu* ("nombre de personnes"). Traiter ces noms comme des noms communs « ordinaires » entraîne une erreur évidente d'interprétation:

(contexte: *nyuugakushiki ni sankasuru gakusei ga atsumatta.*
« Les étudiants qui participent à la cérémonie de début d'année se sont rassemblés. »

Ex.3 *gakusei ha 15 nin desu.*
étudiant TH 15 personnes cop
« Les étudiants sont [au nombre de] 15. »

Ex.4 *(gakusei no) ninzuu ha 15 nin desu.*
(étudiant DET) personne-nombre 15 cop
« Le nombre de personnes (d'étudiants) est 15. »

Dans le premier exemple, le groupe numéral *15 nin* quantifie le groupe nominal *gakusei*. Un groupe verbal distributif (ex : *au* « rencontrer ») peut être appliqué à chaque individu de l'ensemble désigné par ce groupe nominal impliquerait un nombre d'itérations de l'événement égal au nombre d'individus impliqués.

Ex.5 *gakusei ni, hitori zutu atta.*
étudiant NI , un par un rencontré
« J'ai rencontré les étudiants un par un. »

1 "compteur individuel" ou "classificateur (numéral)"

=> 15 fois une rencontre (15 rencontres)

Il n'en va pas de même avec le nom de grandeur. Contrairement au cas précédent, il n'y a pas de distribution possible comme en témoigne l'impossibilité d'associer un quantificateur distributif:

Ex.6 *nin - zuu wo (*hitorizutu /*hitotuzutu) nooto ni kakimashita.*
personnes -nombre O (*un par un / *un par un) cahier sur écrit
* « J'ai écrit un par un *le* nombre de personnes. »

Ce n'est pas un empêchement syntaxique mais bien sémantique. En effet, de (6) , on infère qu'il n'y a eu qu'un seul acte d'écriture, ce dont on rendra compte en traduction en français avec un singulier, comme souligné dans la traduction de l'exemple (6). Une traduction avec un pluriel aurait été erronée:

=/=> j'ai écrit les 15 nombres de personnes sur le/s cahier/s.

En accord avec cette observation, toute interprétation ou traduction qui irait dans le sens d'une identification du nom au groupe numéral échoue pour le nom commun "ordinaire" *gakusei*, et est correct pour le nom de grandeur. Ainsi, à partir de (4) on ne peut pas identifier *gakusei* à *15 nin* (« 15 personnes ») alors que l'identification se déduit naturellement pour le le nom de grandeur :

(3) * "l'étudiant est/vaut 15."

(4) « Le nombre (de personne) est de/vaut 15. »

2 Lister la catégorie

Nous n'avons pas trouvé de listage des noms de grandeurs. Les descriptions sont même rares. Il existe une description syntaxique et sémnantique pour le japonais (sauf (Blin 2013b)), et des discussions sur leur représentation sémantique (formelle, dans la lignée de (Montague 1973)). Un listage manuel est possible car le nombre est relativement faible[2]. Cependant, la catégorie est ouverte qui s'enrichit. En effet, les noms de grandeurs désignent souvent des concepts associés aux technologies. Leur nombre évolue avec les technologies. Un suivi manuel est peu raisonnable, notamment parce qu'il faut couvrir des domaines de spécialités. Reste la détection automatique.
Plusieurs stratégies sont possibles et peuvent être combinées:
1) Utiliser les ressources dictionnairiques existantes pour étendre aux hyper/hyponymes les données précédemment acquises. Cette approche ne permet pas de capturer les nouveaux termes.
2) Extraire les noms de grandeurs dans des distributions prédéfinies. Il n'existe pas en japonais de distribution propre aux noms de grandeurs (Blin 2013b) mais un *ensemble spécifique* de distributions. On devra donc faire la recherche des occurrences dans plusieurs distributions et recouper les résultats. Ceci nécessite un corpus d'une taille considérable. Une recherche exploratoire (voir section 5) sur la collocation [argument , nom de grandeur] sur un corpus japonais de 55 millions de phrases laisse penser que cette taille est très insuffisante. Il est à craindre (pour des raisons matérielles) qu'il faille viser des corpus plus proches de celui utilisé par (Kawahara and Kurohashi 2006b) dans un travail similaire sur les distribution <N no N> (“N de N). Ce corpus comprend plus de 160 millions de phrases et ne permet malgré tout que de traiter aux auteurs de ne fournir la structure argumentale que de 150 noms de grandeurs (sur les 1013 de notre liste). Une autre étude de la même équipe sur les structures argumentales verbales (Kawahara and Kurohashi 2006a) recourt quant à elle à un corpus de 470 millions de phrases et couvre une quantité significative de verbes[3]. La taille de corpus nécessaire pour détecter automatiquement des noms de grandeurs sur la base d'une analyse distributionnelle approcherait donc plutôt les 500 millions de phrases.

2 En japonais, un classement manuel grossier aboutit à 1013 cas
3 Par l'usage, nous estimons que la quasi totalité des verbes est couverte.

3 Extension de l'étude aux arguments et unités de mesure

Quitte à recourir à une procédure automatique, autant rassembler le maximum d'informations pertinentes relatives au nom de grandeur. Pour un bon usage (en interprétation et génération) des noms de grandeurs, plusieurs informations seront utiles, qui peuvent être collectées en même temps que les noms de grandeur eux-mêmes. On sera intéressé par connaître les échelles d'unités associées (voir (Blin 2009)) aux noms : les mètres (ou toute autre unité compatible) pour les hauteurs, les grammes pour les poids, etc. Les unités peuvent différer selon le domaine et il faudrait en tenir compte. Ainsi, en japonais comme en français, l'unité usuelle (non scientifique) pour exprimer la vitesse est le km/h . Mais en japonais, la vitesse du vent est usuellement exprimée en m/s. Dans le domaine maritime, il faudra utiliser le noeud (marin). Ceci suppose une étude par genre/style, et donc un corpus constitué en conséquence (Blin 2012) .
Il faut aussi déterminer la structure argumentale des noms de grandeurs, en particulier les contraintes sémantiques imposées aux arguments. Ceci est fondamental pour désambiguïser certaines phrases. Par exemple, *taijuu* (« poids corporel ») ne s'applique qu'à un animal (humain compris). Cette donnée permet de prédire que *kokuban no taijuu* (« tableau noir + DET + poids corporel ») ne s'interprète pas « poids (corporel) du tableau » . On devra chercher une autre interprétation, vraisemblablement « le poids corporel (de quelqu'un) *qui est inscrit* au tableau ».
Déterminer les arguments et les unités de mesure semble aisé une fois acquise la liste des noms de grandeurs. Il suffit d'observer les collocations dans quelques distributions prédéfinies.

4 Lexicaliser ou non les collocations

Doit-on lexicaliser les collocations, et si oui ? Selon nous, l'association d'une grandeur à une échelle d'unités relève des connaissances générales. Une inscription au niveau du lexique pourrait se justifier partiellement pour rendre compte de « coutumes linguistiques », qui n'ont pas de motivations sémantiques. Ainsi l'expression de la vitesse du vent en m/s au Japon relève d'une coutume linguistique, elle n'obéit pas à une contrainte syntaxique.

5 Etudes exploratoires

Nous avons mené une étude exploratoire pour estimer la taille minimum de corpus nécessaire à l'étude. Pour une entreprise de grande envergure d'extraction des chaînes <nom *no* nom-argument> (N1 de N2 où N2 est argument de N1), (Kawahara and Kurohashi 2006b)utilise un corpus généraliste de 160 millions de phrases. Pourtant, seuls 150 des noms de grandeurs de notre liste ont peu être analysés dans ce corpus. Soit le corpus est trop petit, soit il doit être plus ciblé vers des domaines de spécialités. Une étude détaillée des résultats serait nécessaire.

Les chiffres présentés ci-dessous sont tous obtenus avec Sagace v4.2 (Blin 2014b; Blin 2014a). Le lexique JLFS.1.2.2015-07-02_v150709 est issu du lexique jalexGram-0.010 (Blin 2015b) . Il comprend 1013 noms de grandeurs, rassemblés selon des critères assez lâches. Pour certains noms, l'appartenance à la catégorie reste à discuter. Le corpus corefjp.0.004.150702 (Blin 2015a) compte 55 millions de phrases. Les données sont données brutes , sans post-analyse ni nettoyage. Il faut s'attendre à ce que les chiffre réels soient inférieurs. Une estimation informelle est que les chiffres réels valent 80 % de ceux affichés.

Expé 1 : Dans une distribution très polyvalentes valable pour le noms en général (voir en section 8 la description de la distribution telle que soumise à Sagace), nous avons relevé 683 noms de grandeurs (pour un total de 1 344 249 occurrences). Le corpus de 55 millions de phrases ne permet de capturer que 683 noms, soit 70% de la catégorie.

Expé 2 : Dans la distribution < nom-argument *no* NG> (ex. <$tatemono_{arg}$ *no* $takasa_{NG}$> « $hauteur_{NG}$ du $bâtiment_{arg}$ »), nous relevons 26827 collocations [nom argument, nom de grandeur] (pour un total de 144 329 occurrences) pour 403 noms de grandeur. A titre de comparaison, les noms de grandeur relevés par (Kawahara and Kurohashi 2006b) était de 150 entrées seulement. Notre corpus, bien que plus petit, permet de couvrir un sous-ensemble plus large du lexique. Par contre, un examen rapide des résultats laisse penser que le nombre de noms-arguments pour chaque nom de grandeur est plus élevé chez Kawahara et al.. La taille du corpus n'a donc pas un effet clairement déterminé sur la quantité des résultats.

Expé 3 : Dans la distribution <gnum *no* NG> où gnum est un groupe numéral à base d'unité de mesure, nous relevons les occurrences de la collocation [unité de mesure, nom de grandeur]. Nous relevons 344 collocations (pour un total de 27 34 occurrences) pour 122 noms de grandeur. Le nombre de collocations pourrait être augmenté en prenant aussi en compte la distribution avec valeur post-posée (*takasa 30 meetoru*, « hauteur 30 m », « une hauteur de 30 mètre ») mais le gain risque d'être faible. Le relevé ne peut être effectué de manière optimale avec Sagace, sans effectuer un prétraitement du corpus, consistant surtout à isoler et marquer le groupe numéral.

Expé 4 : Dans la distribution <NG *ga* verbe valuant> qui est particulièrement intéressante pour l'analyse des noms de grandeurs (Blin 2013b), nous relevons 43 noms de grandeur (pour 108 occurrences). Le nombre d'occurrences pourrait être augmenté en prenant en compte la distribution avec le groupe numéral (NG ga groupe_numéral (ni/$VIDE) vvaluant)[4] avec la présence éventuelle d'adverbe. Sagace est peu adapté pour une telle étude et nous ne sommes pas en mesure d'évaluer le taux de réussite des analyseurs mrophosyntaxiques pour une telle structure. Avec cette distribution supplémentaire, la quantité de nom de grandeurs extraits restera malgré tout peu élevée.

Nous avons compté le nombre de noms de grandeurs connus qui apparaissaient dans les quatre analyses en même temps. Ce nombre s'élève à 26 [5], soit 2,5 % du nombre total de noms de grandeurs connus. C'est une valeur très faible. Pour couvrir l'ensemble des noms de grandeur, il faudrait très significativement augmenter la taille du corpus pour pouvoir couvrir l'ensemble des noms de grandeurs. Sur la base de ces résultats, à moins de disposer d'un corpus d'une taille considérable, on peut penser qu'il sera très difficile de déterminer l'appartenance d'un nom commun « inconnu » à la catégorie des noms de grandeurs, par simple analyse de fréquence de cette chaîne dans les quatre distributions retenues.

Cette étude exploratoire ne prend pas en compte tous les noms de grandeur composés comme *seityouritu* 成長率, « taux de croissance », construits par application de règles morphosémantiques *productives*. Une réflexion doit être menée pour décider si il faut lister ou pas les noms composés. D'un point de vue linguistique, on privilégiera l'hypothèse selon laquelle ces formes dérivées sont compositionnelles. Dans cette approche, nous supposerons que l'argument est donné par le radical (*seichou*) et l'unité par le « suffixe de grandeur » (*ritsu* « taux ») (Blin 2013b). La réflexion sur l'inclusion ou non des noms de grandeur composés nécessitera de se positionner aussi par rapport au débat sur les unités lexicales (unités longues, courtes etc, cf par exemple (Maekawa 2009)).

6 Conclusion

Déterminer automatiquement l'appartenance d'un nom commun à la catégorie des noms de grandeurs semble très difficilement réalisable par une simple observation de la fréquence de la chaîne dans un ensemble prédéterminé de distribution. Il faudrait en effet un corpus d'une taille extrêmement grande, dépassant certainement les plus grand corpus utilisés dans le monde académique ((Kawahara and Kurohashi 2010)?). Par contre, cette méthode semble tout à fait exploitable pour relever les collocations [nom de grandeur , argument du NG , unité associée] au NG et à l'argument. La question de la taille du corpus se pose aussi mais ne semble pas insurmontable. Nos observations permettent de dire que 55 millions de phrase est insuffisant, et que même 160 millions risque d'être insuffisant. Il devrait être possible d'augmenter le nombre de cas en ciblant les corpus. En effet, les noms de grandeurs usuels apparaissent en nombre suffisant pour être capturés dans des textes usuels. Par contre, il faudra se doter de corpus spécialisés pour le reste.

---

4 温度が１０度に上がった。
ondo ga <10 do>$_{gnum}$ ni <agatta>$_{vvaluant}$
temp. S 10 devrés à augmenté
« La température est monté à 10 degrés. »

5 ウエート, 仕事, 仕事量, 体重, 価格, 力, 収量, 圧力, 売り上げ, 売上高, 弾性率, 径, 放射能, 残量, 流量, 濃度, 直径, 積載量, 粘度, 総額, 自重, 輸送量, 重さ, 重量, 金額, 面積 ; parmi ces noms, certains apparaissent certainement dans le corpus dans un autre emploi que celui de nom de grandeur : 仕事 (*shigoto* , « travail », 直径 (*chokkei ,* « diamètre »).

7 Bibliographie

Blin, Raoul. 2009. “Le Groupe Numéral En Japonais.”

———. 2012. “Automatic Addition of Genre Information in a Japanese Dictionary.” *Acta Linguistica Asiatica* 2 (2): 83–96.

———. 2013. “Measure Nouns in Japanese : Their Semantic and Syntactic Characteristics.” *Lexicon Forum*, no. 6 (January): 173–202.

———. 2014a. “Comparaison de deux outils d’analyse de corpus japonais pour l’aide au linguiste, Sagace et MeCab.” In *Actes TALN-RECITAL 2014*, 497. Marseilles, France. http://hal.archives-ouvertes.fr/hal-01054370.

———. 2014b. “Manuel, Sagace 4.2.0.” http://crlao.ehess.fr/japonais-coreen/corpus/sagace/manuel/Manuel.pdf.

———. 2015a. “Corefjp-0.003.150528, (Another) Corpus for Written Contemporary Japanese.” http://goo.gl/p0Tx7h.

———. 2015b. “Metadonnées Du Lexique-Grammaire Du Japonais jalexGram-0.010.” http://goo.gl/3BvzGU.

Kawahara, Daisuke, and Sadao Kurohashi. 2006a. “Case Frame Compilation from the Web Using High-Performance Computing.” In *Proc. of LREC’2006*, 1344–47.

———. 2006b. “A Fully-Lexicalized Probabilistic Model for Japanese Syntactic and Case Structure Analysis.” In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, 176–83. New York City, USA: Association for Computational Linguistics. http://www.aclweb.org/anthology/N/N06/N06-1023.

———. 2010. “Acquiring Reliable Predicate-Argument Structures from Raw Corpora for Case Frame Compilation.” In *Acquiring Reliable Predicate-Argument Structures from Raw Corpora for Case Frame Compilation*, 1389–93. Valetta, Malta.

Maekawa, Kikuo. 2009. “Daiyousei Wo Yû Suru Daikibo Nihongo Kakikotoba Kôpasu (<tokushyû> Nihongo Kôpasu) [Compilation D’un Corpus Équilibré de Textes Contemporains En Japonais.” *Journal of Japanese Society for Artificial Intelligence* 24 (5): 612–22.

Montague, Richard. 1973. “The Proper Treatment of Quantification in Ordinary English.” In *Philosophy, Language, and Artificial Intelligence*, edited by Jack Kulas, James H. Fetzer, and Terry L. Rankin, 2:141–62. Dordrecht: Springer Netherlands. http://www.springerlink.com/index/10.1007/978-94-009-2727-8_7.

8 Annexe

Requête pour dénombre le nombre de noms de grandeurs, toutes distributions confondues :

```
>0 cat:particule  | ponctuation | debsegment | markDeb
=0 cat:nomgrandeur /-affich:trait:lemme /-compte
=0 cat:particule | ponctuation | copule | arab
```

Requête pour extraire et dénombrer les collocations [nom de grandeur , argument] dans la distribution <nom commun + *no* + nom de grandeur> :

```
>0 cat:particule  | ponctuation | debsegment | markDeb
=0 cat:nomcommun /-affich:trait:lemme
=0 の /-affich:" "
=0 cat:nomgrandeur /-affich:trait:lemme /-compte
=0 cat:particule | ponctuation | copule | arab
```

Requête pour extraire et dénombrer les collocations [unité de mesure , nom de grandeur] dans la distribution < groupe numéral + nom de grandeur>.

Requête 1 :

```
>0 cat:particule  | ponctuation | debsegment | markDeb | chiffre
=0 cat:chiffre
=0 cat:unite /-affich:trait:lemme
=0 の /-affich:" "
=0 cat:nomgrandeur /-affich:trait:lemme /-compte
=0 cat:(particule & -の) | ponctuation | copule
```

Nous excluons no des particules de sorte à ce qu'il n'y ait pas d'ambiguité sur la portée du groupe numéral. En effet, dans une chaîne <gnum no NG no + X> , il n'est pas exclu que le gnum porte sur X.