Gradients de prototypicalité conceptuelle et lexicale

Xavier Aimé* ***, Frédéric Fürst**, Pascale Kuntz*, Francky Trichet*

*LINA - Laboratoire d'Informatique de Nantes Atlantique (CNRS-FRE 2729)
Équipe COD - Connaissances & Décision
2 rue de la Houssinière BP 92208 - 44322 Nantes Cedex 03
pascale.kuntz@univ-nantes.fr, francky.trichet@univ-nantes.fr
**LARIA - Laboratoire de Recherche en Informatique d'Amiens (CNRS-FRE 2733)
UPJV, 33 rue Saint Leu - 80039 Amiens Cedex 01
frederic.furst@u-picardie.fr
***Société TENNAXIA
19 rue de Réaumur - 75003 Paris
xaime@tennaxia.com

Résumé. Longtemps les ontologies ont été limitées à des domaines scientifiques et techniques, favorisant au passage l'essor du concept de « *connaissances universelles et objectives* ». Avec l'émergence et l'engouement actuel pour les sciences cognitives, couplés à l'application des ontologies à des domaines relatifs aux Sciences Humaines et Sociales (SHS), la subjectivité des connaissances devient une dimension incontournable qui se doit d'être intégrée et prise en compte dans le processus d'ingénierie ontologique (IO). L'objectif de nos travaux est de développer la notion d'Ontologie Pragmatisée Vernaculaire de Domaine (OPVD). Le principe sous-jacent à de telles ressources consiste à considérer que chaque ontologie est non seulement propre à un domaine, mais également à un endogroupe donné, doté d'une pragmatique qui est fonction tant de la culture que de l'apprentissage et de l'état émotionnel du dit endogroupe. Cette pragmatique, qui traduit un processus d'appropriation et de personnalisation de l'ontologie considérée, est qualifiée à l'aide de deux mesures : un gradient de prototypicalité conceptuelle et un gradient de prototypicalité lexicale.

Mots clés : Ontologie Pragmatisée Vernaculaire du Domaine, Prototypicalité Conceptuelle, Prototypicalité Lexicale, Catégorisation, Émotions, Culture, Apprentissage, Gradient, Personnalisation, Adaptation, Pragmatique.

1 Introduction

D'un point de vue linguistique, la pragmatique s'intéresse aux éléments du langage dont la signification ne peut être comprise qu'en fonction d'un contexte d'interprétation donné. Dans le cadre des ontologie de domaine (qui sont des spécifications formelles de conceptualisations partagées Gruber (1993)), il s'agit d'enrichir la sémantique formelle intrinsèque à une ontologie de domaine (OD) à l'aide d'éléments caractéristiques d'un contexte de création ou d'usage

comme la culture, le mode d'apprentissage, ou encore l'état émotionnel. Pour ce faire, nous prenons comme point de départ les ontologies telles quelles sont définies aujourd'hui au moyen du langage OWL. Nous y ajoutons deux gradients, dont le but est de modéliser une différence intuitive de degré de vérité dans le processus de catégorisation : (1) une pondération des liens « *is-a* » dans la hiérarchie des concepts, les **gradients de prototypicalité conceptuelle**, (2) une pondération des synonymes du label utilisé pour dénoter un concept donné, le **gradient de prototypicalité lexicale**. Ces gradients permettent de prendre en considération différents points de vue pour une même ontologie, et par la-même d'enrichir la sémantique formelle initiale intrinsèque à l'ontologie d'une pragmatique inhérente au contexte d'usage considéré (fonction de la culture, des modes d'apprentissage ou encore du contexte émotionnel). Après avoir exposé notre approche[1] dans les sections 2.1 et 2.2, les gradients de prototypicalité conceptuelle et de prototypicalité lexicale, fondés sur les processus cognitifs de catégorisation, sont explicités dans les sections 2.3 et 2.4.

2 Gradients de prototypicalité

Nos travaux reposent sur l'idée fondamentale que tous les concepts ne sont pas constitués de membres *équidistants* par rapport à la catégorie qui les subsume, mais qu'ils comportent des membres qui sont de meilleurs représentants que d'autres (Kleiber, 2004). Ce phénomène est également applicable pour l'ensemble des termes[2] désignant un concept. C'est sous cette hypothèse, validée par des travaux de psychologie cognitive tels que Cordier (1985), que nous proposons deux mesures : (1) le **gradient de prototypicalité conceptuelle**, qui correspond à une pondération des liens « *is-a* » dans la hiérarchie des concepts, et qui permet de mesurer les différentes représentativités des sous-concepts ; (2) le **gradient de prototypicalité lexicale**, qui, pour un concept donné, correspond à une pondération des synonymes du terme saillant.

2.1 Approche objective, fondée sur les propriétés

La composante objective de notre gradient vise à prendre en compte l'aspect *intensionnel* d'une conceptualisation aux travers des propriétés de concepts. Le rôle des attributs d'une catégorie en vue d'un rattachement à une autre catégorie a été développé dans Smith et al. (1974). Nous appelons cette pondération *objective*, car elle est issue du courant *objectiviste* selon lequel la catégorisation s'opère sur la base de propriétés communes (Kleiber, 2004). Il s'agit d'un courant classique qui se fonde sur une vue catégorielle de l'univers tel qu'a pu le définir Aristote. Cette objectivité provient uniquement de la conceptualisation élaborée par l'endogroupe, conceptualisation capturée dans l'ontologie via la hiérarchie des concepts et leurs propriétés.

L'idée directrice est que plus un concept va ajouter d'attributs à ceux hérités de son père, plus il sera sur la voie de la spécialisation et moins il sera prototypique, *i.e.* représentatif de sa catégorie. Nous considérons cette valeur comme étant le rapport entre le nombre d'attributs du

[1]Ce travail de recherche est réalisé dans le cadre d'un projet mené en collaboration et financé par la société Tennaxia (http ://www.tennaxia.com), société de service et de conseils en veille juridique et réglementaire dans le domaine Hygiène, Sécurité, Environnement et Développement Durable (HSE-DD).

[2]Cet ensemble est composé d'un terme saillant (Sowa, 1984), considéré comme le plus fort lexicalement pour désigner le concept, et d'une liste de synonymes.

fils et le nombre d'attributs du père. De manière formelle, la fonction $objectif : \mathcal{C} \times \mathcal{C} \rightarrow [0, 1]$ est définie comme suit :

$$objectif(c_f, c_p) = \left(\frac{attributs(c_p)}{attributs(c_f)}\right)^n$$

Où (1) $attributs(c_p)$ est le nombre d'attributs du concept parent, (2) $attributs(c_f)$ le nombre d'attributs du concept fils, et (3) n le nombre de concepts fils du concept c_p.

Nous élevons le rapport entre le nombre d'attributs des deux concepts à la puissance n de manière à pouvoir tenir compte de la structure de l'héritage et ainsi favoriser davantage les éléments dont la valeur *objectif* est forte. Le concept le plus prototypique d'une décomposition se trouve ainsi renforcé proportionnellement au nombre d'éléments de cette décomposition.

2.2 Approche subjective, fondée sur les fréquences

La composante subjective de notre gradient vise à prendre en compte l'aspect *extensionnel* d'une conceptualisation à travers les termes utilisés pour dénoter les concepts. Cette approche est fondée sur la fréquence d'apparition d'un concept appartenant à un domaine D dans l'univers d'un endogroupe G. De la sorte, plus un élément sera fréquent dans cet univers, plus il sera *jugé* comme *représentatif / typique* de sa catégorie. Cette notion de typicalité fait l'œuvre des travaux d'Eleanor Rosch (Rosch, 1973). Dans notre contexte, l'univers de l'endogroupe va être constitué par l'ensemble des documents contenus dans $\Omega_{(D,G)}$.

De manière formelle, la fonction $subjectif_{G,D}(c_f, c_p) : \mathcal{C} \times \mathcal{C} \rightarrow [0, 1]$ est définie comme suit[3] :

$$subjectif_{G,D}(c_f, c_p) = \frac{information(c_f)}{information(c_p)}$$

Où $information(c) = \sum_{terme \in monde(c)} (frequence(terme) * \frac{count(document, terme)}{count(document))}$
Avec :

- $frequence(terme)$ retourne la fréquence d'apparition du terme dans les documents appartenant à $\Omega_{(D,G)}$;
- $count(document, terme)$ retourne le nombre de documents appartenant à $\Omega_{(D,G)}$ où le terme apparaît ;
- $count(document)$ retourne le nombre de documents appartenant à $\Omega_{(D,G)}$;
- $monde(c)$ retourne tous les termes (*i.e.* terme saillant et synonymes) relatifs au concept c et à sa descendance.

Intuitivement, la fonction $information(c)$ permet d'évaluer le taux d'utilisation d'un concept dans un corpus, et ce à l'aide de l'ensemble des termes associés (*i.e.* terme saillant et synonymes) non seulement à ce concept mais aussi à sa descendance - du fait que tous les sous-concepts participent à la connaissance d'un concept. Nous pondérons chaque fréquence par

[3] Sachant que cette fonction n'est applicable que s'il existe :
- un lien de type « *is-a* » entre le concept père c_p et le concept fils c_f (avec une relation d'ordre $c_f \leq c_p$),
- ou une chaîne entre le concept *ascendant* c_p et le concept *descendant* c_f (avec une relation d'ordre $c_f \leq ... \leq c_p$).

le rapport entre le nombre de documents où le terme est présent et le nombre total de document. Une idée présentée fréquemment dans peu de documents est moins influente en terme de connaissance qu'une idée défendue et citée peu de fois dans chacun des documents mais de façon uniforme dans tout le corpus[4]

2.3 Gradients de prototypicalité conceptuelle

Les gradients de prototypicalité conceptuelle prennent en compte, dans le processus de catégorisation et de classement, deux aspects des concepts : intensionnel d'un côté par leur structure, extensionnel de l'autre par leur fréquence d'évocation et d'utilisation dans l'univers de l'endogroupe.

Gradient de prototypicalité conceptuelle locale (GPCL). Soit $protconloc_{G,D} : \mathcal{C} \times \mathcal{C} \rightarrow [0, 1]$ la fonction qui, pour tout couple de concepts $c_f, c_p \in \mathcal{C}$ tel qu'il existe un lien de type « *is-a* » entre le concept *père* c_p et le concept *fils* c_f (avec une relation d'ordre $c_f \leq c_p$.), retourne une valeur réelle positive ou nulle représentant le gradient de prototypicalité conceptuelle de ce lien, et ce dans une ontologie pragmatisée vernaculaire relative au domaine D et à l'endogroupe G. Cette fonction se décompose de la manière suivante :

$$protconloc_{G,D}(c_p, c_f) = [\alpha * objectif(c_f, c_p) + \beta * subjectif_{G,D}(c_f, c_p)]^{\gamma}$$

avec (1) $\alpha \in [0, 0.5]$ une pondération de la composante objective, (2) $\beta \in [0, 0.5]$ une pondération de la composante subjective, et (3) $\gamma \geq 0$ une pondération de l'état mental de l'endogroupe G.

Les valeurs de α et de β peuvent varier suivant le domaine traité, la volonté des experts lors de la création de l'ontologie, le contexte d'usage de l'ontologie considérée, etc. Ils permettent ainsi d'accorder plus ou moins d'importance à l'aspect structurel de la conceptualisation par rapport à l'évocation des concepts dans les documents. Quant au facteur γ, il a pour objectif de tenir compte de l'état mental de l'endogroupe. Selon Mikulinger et al. (1990), un état mental négatif favorise la diminution de la valeur de représentation, et inversement pour un état mental positif. Nous caractériserons (1) un état mental *négatif* par une valeur de $\gamma \in]1, +\infty[$, (2) un état mental *positif* par une valeur de $\gamma \in]0, 1[$, et (3) un état mental *neutre* par une valeur de 1.

Ainsi, une très faible valeur de γ va avoir pour effet d'augmenter considérablement la valeur de ce gradient pour des concepts initialement placés comme peu représentatifs (un état positif facilite l'ouverture d'esprit, la valorisation etc.) A l'inverse, une très forte valeur de γ, pour un état mental fortement négatif, va avoir pour effet de *sélectionner* uniquement les concepts à forte typicalité, éliminant *de facto* les autres concepts. Une ontologie pragmatisée vernaculaire de domaine est par conséquent une ontologie vernaculaire de domaine placée dans un contexte particulier, défini par les trois paramètres α pour la composante objective, β pour la composante subjective et γ pour l'état mental.

[4]Par corpus, nous entendons un ensemble de ressources de natures potentiellement différentes (textuelles, sonores, visuelles, etc.) mais inhérentes à un même domaine.

2.4 Gradient de prototypicalité lexicale (GPL).

Ce gradient a pour but de valuer le fait que tous les synonymes d'un terme considéré comme le terme saillant, *i.e.* le *label préconisé* pour un concept, n'ont pas forcément la même représentativité au sein de l'endogroupe. La question est en effet « *pourquoi nommons-nous plus le concept x avec le vocable y plutôt que z ?* » Pour définir ces variations de représentativité lexicale, nous reprenons le gradient calculé précédemment, à la différence près que nous n'allons pas utiliser la composante objective liée aux propriétés (du fait que nous sommes sur la même catégorie). Nous allons une fois encore nous inspirer de la formule calculant le contenu en information d'un concept, en prenant pour évaluer ce gradient le rapport entre la fréquence d'utilisation de ce terme et la somme de fréquences de tous les termes du concept dans $\Omega_{(D,G)}$.

D'un point de vue formel, soit $protlex_{G,D} : L_C \times \mathcal{C} \times \Omega_{(D,G)} \rightarrow [0,1]$ la fonction qui pour tout concept $c \in \mathcal{C}$ et terme $t \in L_C$ tel que $t \in f_{labelC}(c)$ retourne une valeur positive ou nulle représentant le gradient de prototypicalité lexicale de ce terme, et ce pour un domaine D et un groupe d'individus G. Cette fonction se calcule de la manière suivante :

$$protlex_{G,D}(t,c) = \frac{\sum count(t)}{\sum count(f_{labelC}(c))}.$$

3 Conclusion

La finalité de nos travaux, focalisés sur la notion d'ontologie pragmatisée vernaculaire de domaine, est de prendre en compte la subjectivité de la connaissance via sa spécificité à un endogroupe et un domaine, son aspect écologique, et l'importance de son contexte émotionnel. Cet objectif nous conduit à étudier la dimension pragmatique d'une ontologie. En nous inspirant des travaux d'E. Rosch sur la prototypicalité, nous avons développé deux mesures identifiant deux gradients de prototypicalité, l'un conceptuel et l'autre lexical, de manière à pouvoir - entres autres - pondérer les liens « *is-a* » des hiérarchies catégorielles. Ces pondérations permettent, dans un premier temps et au niveau local, d'ordonner les sous-concepts d'une catégorie donnée pour le premier gradient, la liste des synonymes d'un terme saillant pour le second. Bien évidemment, ces gradients ne modifient en rien la sémantique formelle inhérente à l'ontologie considérée ; les liens de subsomption restent valides. Ces gradients ne sont que le reflet de la prise en compte de la pragmatique, à savoir une sur-couche pragmatique sur la sémantique.

A partir de ces informations, plusieurs applications peuvent être envisagées :

- *évaluation / validation d'ontologies*. Les gradients de prototypicalité peuvent représenter des indicateurs de qualité de catégorisation, et par là même de qualité de la modélisation inhérente à une ontologie de domaine formalisée (au moyen d'OWL par exemple).Par exemple, il peut se poser le problème des concepts jugés très peu typiques. Que faut-il en faire ? Sont-ils à la bonne place dans la hiérarchie catégorielle ?
- *recherche d'information*. Les gradients de prototypicalité peuvent permettre de classer les résultats d'une requête (sous-entendue d'une requête dite *étendue*) suivant un ordre de pertinence, en plaçant au premier plan les éléments les plus représentatifs d'une catégorie ou d'un terme donné.

– *analyse de corpus spécialisé pour un endogroupe en liaison avec sa connaissance*. Les gradients de prototypicalité, sur un plan macroscopique, permettent d'extraire les concepts les plus représentatifs des catégories premières au sens aristotélicien.

Nos travaux se poursuivent actuellement sur le perfectionnement de ces deux gradients, en se référant à de nombreux travaux de psychologie cognitive et sociale. Enfin, une application de ces résultats à la recherche d'information est actuellement en cours.

Références

Cordier, F. (1985). Formal and locative categories. are there typical instance ? *Psychologica Belgica XXV*(2), 115–125.

Gruber, T. (1993). Toward principles for the design of ontologies used for knowledge sharing. In N. Guarino et R. Poli (Eds.), *Formal Ontology in Conceptual Analysis and Knowledge Representation*, Deventer, The Netherlands. Kluwer Academic Publishers.

Kleiber, G. (2004). *La sémantique du prototype*. Presses Universitaire de France - coll. Linguistique Nouvelle. ISBN 2 13 042837 1, 2e édition.

Mikulinger, M., P. Kedem, et D. Paz (1990). Anxiety and categorization-1, the structure and boundaries of mental categories. *Personnality and individual differences 11*(11), 805–814.

Rosch, E. (1973). Natural categories. *Cognitive Psychology* (4), 328–350.

Smith, E., E. Shoben, et L. Rips (1974). Structure and process in semantic memory : a featural model for semantic decisions. *Psychological Review* (81), 214–241.

Sowa, J. F. (1984). *Conceptual structures : Information Processing in Mind and Machine*. Addition-Wisley Publishing Company. ISBN 0-201-14472-7.

Summary

Since a long time, Domain Ontologies have been limited to scientific and technical domains, advantaging the concept rise of "unbiased and universal knowledge". With the current emergence of cognitive sciences and the application of ontologies to social and human sciences, subjective knowledge becomes an unavoidable subject, which must be integrated and developped in Ontological Engineering. The aim of our work is to develop the notion of Domanial Pragmatised Vernacular Ontology (DPVO). The principle underlying these DPVO consists in considering that each ontology is not only characteristic of a domain, but is also peculiar to a precise group which owns a pragmatics based on several parameters, and more precisely: culture, learning process and mental states. Such a pragmatics, which is a translation of an ontology customization process, is qualified by two measurements: a conceptual prototypicality gradient and a lexical prototypicality gradient.

Keywords : Conceptual prototypicality, Lexical prototypicality, Computational ontology, Environment, Typicality, Information Retrieval, Emotions, Culture, Learning, Gradient, Adjustment, Pragmatics.