# ターミノロジー理論に基づくオントロジーのマッピング：認知的翻訳アプローチ

加納史子　（Fumiko Kano Glückstad）
Copenhagen Business School, International Language Studies and Computational Linguistics
Dalgas Have 15, DK-2000 Frederiksberg, Denmark
E-mail: fkg.isv @ cbs.dk

概要：バイリンガルを認知的に研究する記憶表象の研究分野で De Groot (1997) は、概念表象 (conceptual representation) と語彙表象 (lexical representation) の関係を表す 6 つモデルの 1 つである概念特徴 (distributed conceptual feature) モデルについて、2 言語の語彙表象がいくつかの概念特徴を共有する場合、翻訳過程がスムーズであることを論じている。一方 Madsen (2004)らはターミノロジー手法に基づく概念モデリングを提案し、概念の特徴を属性とするオントロジーを構築する手法を提唱している。この手法を利用して、ソース言語と対象言語それぞれについてドメインオントロジーを構築し、語彙の類似度、グラフパターンと特徴属性の組合せをもとに、オントロジーマッチングを行うことで、言語資源の希少な組合せ言語間の翻訳を行うアイデアを提案する。

## 1. はじめに

デンマーク語と日本語などの希少な組み合わせの言語間において翻訳を行う場合、ソース言語と対象言語を直接結ぶ辞書や対訳言語資源に限りがあるため、英語等の資源豊富な言語を介して翻訳を行うことが多い。人間が辞書を使用して翻訳する場合でも、統計的機械翻訳システムが対訳言語資源を活用して翻訳を行う場合でも、ソース言語にて本来意味する用語の概念を、英語等の中間言語を介して、対象言語を使用する読者に伝えることは、困難な課題だといえる。本提案は、ソース言語が持つ用語の概念をより的確にかつ効率的に対象言語を使用する読者に伝える手法に関するものである。本提案に際しては、まず、単語の翻訳における認知プロセスに関する研究で De Groot (1997) が提唱する概念特徴モデルと、ターミノロジー手法の一つとして Madsen (2004) らが提唱する知識ベースアプローチとの間に共通点を見出した。

本稿では、第 2 章と第 3 章でそれぞれ、De Groot (1997) の概念特徴モデルと Madsen (2004) のターミノロジー手法に基づくオントロジーを説明し共通点を整理する。さらに、第 4 章で、ターミノロジー手法を活用して構築したオントロジーをマッピングすることで、概念特徴モデルにて説明される、用語のもつ概念を認知的に翻訳する方法を提案する。最後に、第 5 章で今後の課題、第 6 章でまとめについて述べる。

## 2. 概念特徴モデル

独立した単語 (名詞) の翻訳に関する認知プロセスの実験において、De Groot (1997) は、単語の具象性と使用頻度が、的確な訳語特定に要する反応時間に影響を及ぼすことを証明している。そして、特に単語の具象性が反応時間に及ぼす影響について、バイリンガル記憶表象の 6 つのモデルのひとつである概念特徴モデルを使って説明している。

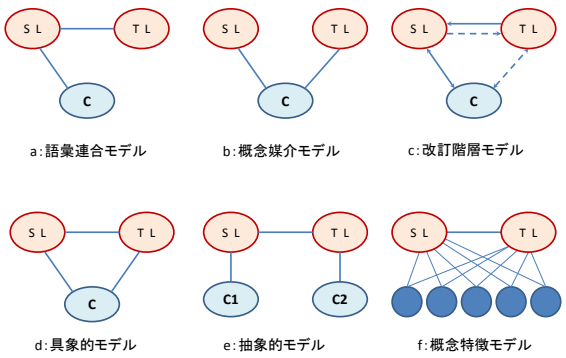


図 1　バイリンガル記憶表象モデル

記憶表象の階層モデルは、もともと Potter ら (1984) により提唱された。Potter らのモデルには、図 1-a のように、2 言語の語彙表象間およびソース言語の語彙表象と概念表象の間が結びつく語彙連合モデルと、図 1-b のように、2 言語の語彙表象の間には結びつきは存在せず 2 つの言語の語彙表象がそれぞれ概念表象と結びつく概念媒介モデルがある。これらのモデルに加えて、Kroll ら (1994) は、図 1-c のように、対象

言語の習熟度に応じて、各表象間の結びつき強度が変化するという改訂階層 (revised hierarchical) モデルを提唱している。

De Groot (1997) によると、翻訳の認知プロセスにおける単語の具象性の影響は、2つの異なる理論にて説明できる。まず第一に考えられる理論として、階層モデルには、具象的モデルと抽象的モデルがある。すなわち、図1-dは、具象性の高い単語を示す典型的なパターンで、2言語を結びつけるルートが語彙連合および概念媒介の両方を介するため、翻訳がスムーズになると説明している。これに対し、図1-eは抽象的な単語を示す典型的なパターンで、2言語の語彙表象がそれぞれ特有の概念を有しており、語彙表象でのみ結びついているため、翻訳過程を困難にしている。第2の理論は、具象効果を概念特徴モデルを用いて説明するもので、図1-fのように、具象的な単語は抽象的な単語に比べ、2言語間においてより多くの概念特徴を共有するため翻訳過程がよりスムーズであるとしている。

さらにDe Groot (1995) の別の研究によると、母国語に加えて2種類の第二外国語を習得している被験者について、より流暢な第二外国語と母国語間の翻訳においては、概念媒介による結びつきが強く、苦手な第二外国語と母国語間の翻訳においては、語彙連合の傾向が高いことが証明されている。

これらの説明から、2言語間の語彙表象が、概念特徴をより多く共有していれば、翻訳がスムーズになり、特に、距離のある言語間にてその効果を発揮できると想定される。つまり、共通の概念特徴を2言語間において組織的に抽出するシステムが存在すれば、的確かつ効率的に翻訳を行える可能性があると考えられる。

## 3. ターミノロジカルオントロジー

ターミノロジー手法に基づくオントロジー (以下「オントロジー」) は、Madsen (1999) が提唱するアプローチで、ターミノロジー理論に基づく概念モデリング (以下「概念モデル」) により構築されるオントロジーである。オントロジー構築は、半自動オントロジー構築システムの開発を目的とした CAOS (Computer-Aided Ontology Structuring) プロジェクトにおいて Madsen ら (2004, 2005) が提唱したもので、本研究で利用するオントロジー半自動構築システム「i-term」の雛形となっている。CAOS の概念モデルは、属性値の形式的記述によりモデル化された特徴で構成される。これらの特徴記述は、いくつかの原則とルールに基づき構成される。中でも重要な原則として、概念の持つ特徴は、属性値として記録され、下位概念は上位概念に記録されたすべての特徴を継承する。これは、ISO 704: 5.4.2.2 において定義されているターミノロジー手法の原則でもある。さらに、属性と値は 1 対のペアとなる。図 2 の例では、学位という属性において「学士」と「修士」の両方を値として持つことはできず、一つの属性に対し一つの値のみを持つ。次に、概念を区分する区分要因は、一つの概念を区分する要因として、原則としてオントロジー内で一度限り使用される。つまり、区分要因により生じた異なる属性値 (特徴) は、その下位概念レベルにおいて兄弟関係を構成する。一方で、一つの概念に対して、一つ以上の区分要因を有することができる。つまり、概念を区分する要因として「年齢」「年数」「種類」といった複数の区分要因を定義して、区分基準として最も重要な要因を設定することができる。このような原則とルールに従い構築したオントロジーでは、近隣に位置する親子、兄弟関係の概念は、特徴という点で互いに区別されることになる。

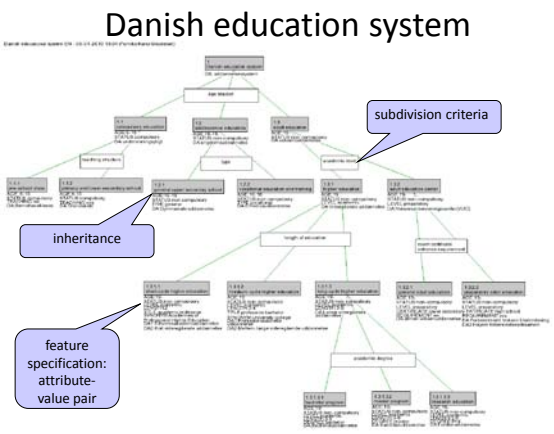


図 2　ターミノロジカル・オントロジー

最終的に、図 2 の全体像が示すように、各用語の語彙表象下に、組織的に概念特徴が記述され蓄積されて、オントロジーが構築されていく。すなわち、CAOS の原則に基づきオントロジーを構築することで、第 2 章で De Groot が説明する概念特徴を、組織的に抽出することが可能になると考えられる。

## 4. オントロジーマッピング

冒頭で説明したとおり、本提案は、デンマーク語と日本語など、直接的な言語資源に限りがある希少な組み合わせの言語間において、ソース言語が持つ用語の概念をより的確にかつ効率的に対象言語を使用する読者に伝えることに焦点をあてている。しかし、同じ手法を資源豊富な言語間 (例えば EU 諸国言語間) に適用するシナリオも想定できる。その場合は、以下に説明するバイリンガルオントロジーの代わりに、2 つの単一言語オントロジーをマッピングすればよい。

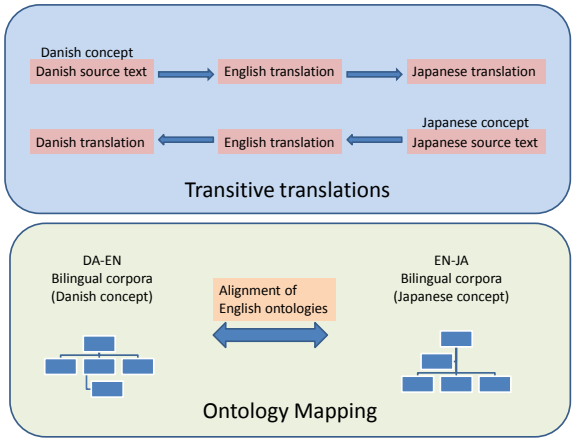


図 3　バイリンガルオントロジー

具体的なワークフローとして、デンマーク語と日本語のマッピングを想定した場合、ある特定の専門分野において、デンマーク語と英語および英語と日本語それぞれについて、対訳コーパスまたは内容的に整合されたバイリンガルコーパスを特定する。それぞれのバイリンガルコーパスから、オントロジー構築の対象となる用語ペアとその定義を抽出する。つまり、デンマーク語の用語とその英語訳、日本語の用語とその英語訳、および、これらの用語の英語による定義を抽出する。これらの用語と定義から、共通する区分要因と属性を見出し、3 章の原則とルールに従い英語表現によりバイリンガルオントロジーを構築していく。このバイリンガルオントロジーには、バイリンガルコーパスから抽出されたデンマーク語または日本語による表現が、属性値とともにメタ情報として語彙表象下に記録される。つまり、図 3 のように、ソース言語を英語を介して翻訳するという 1 ステップ翻訳ではなく、英語で表現された 2 つのバイリンガルオントロジーを整合することで、デンマーク語と日本語の対訳を特定することになる。現状では、第 3 章で紹介したオントロジー半自動構築システム「i-term」を利用して、手作業でオントロジーを構築している。

次に、2 つのバイリンガルオントロジーをマッピングする手法については、2 つのアプローチが考えられる。まず語彙表象自体の類似度によりマッピングを行う方法がある。図 4 のように、両オントロジーにて「higher education」が存在すれば、「高等教育」のデンマーク語訳は「videregående uddannelse」だと想定できる。しかし、「短期大学」については、このような手法を適用することはできない。そのため、語彙表象下に記録された属性値の組み合わせがもっとも類似する語彙表象を特定して、そこに記録されたデンマーク語メタ情報をデンマーク語訳だと想定する。これら 2 つの手法以外にも、全体的なオントロジーのグラフパターンを照らし合わせ、上位・下位概念の関係を推論する手法を組み合わせることも考えられる。

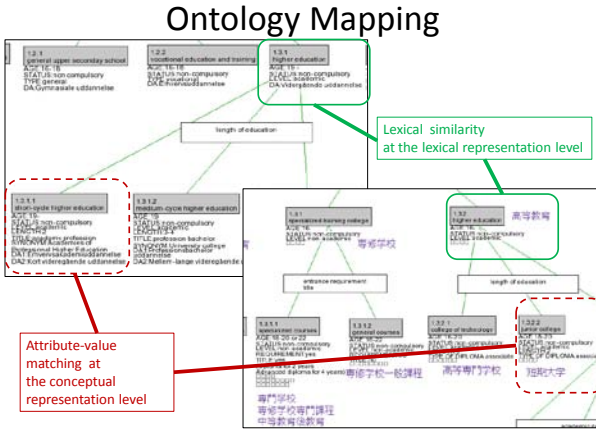


図 4　　オントロジーマッピング

## 5. 今後の課題

本稿で使用したデンマークと日本の教育制度に関する用語のマッピングについては、すべて手作業で、対訳データから用語を抽出してバイリンガルオントロジーを構築しマッピングを行った。今後の課題としては、まず、今回行った手作業をどこまで自動化できるかを検討する必要がある。対訳コーパスの整合と用語抽出については、既存の公開ツール等を活用することが可能だろう。ターミノロジーの原則に基づくオントロジーの自動構築については、Madsen ら、コペンハーゲンビジネススクールのターミノロジー研究グループにおいて検討中の課題である。

著者のプロジェクト課題としては、2 つのバイリンガルオントロジーのマッピングのルールを定義して、さらに自動化するアルゴリズムについて今後研究を進めていく予定である。

次に、オントロジーマッピングで得られたデンマーク語と日本語の対訳ペアについて、これらの訳語が、ソース言語の翻訳としてどのように評価できるかが問題となる。コペンハーゲンビジネススクールは、欧州議会等で活躍する修士レベルの翻訳者・通訳者を養成する国立大学でもあり、何を翻訳クオリティ評価の基準とするかについては、翻訳プロセス研究グループにおいて議論となっている課題でもある。まず、客観的な翻訳のクオリティに関する評価については、マッピング手法で得られたデンマーク語と日本語の対訳ペアを統計的機械翻訳システムに組み込み、統計的機械翻訳をベースラインとして翻訳結果を BLEU スコア等の客観的指標を使って比較する実験を計画中である。しかし、これらの機械翻訳システムを評価するために開発された客観的指標だけで、用語の翻訳クオリティを判断することはできないといえる。オントロジーマッピングで得られた訳語は、人間の記憶表象の階層モデルにおいて、2 言語の語彙表象が共通の概念特徴と結びつくことにより得られた訳語だと解釈できる。つまり、オントロジーマッピングで得られた用語の翻訳は、機械翻訳等の方法で得られた翻訳と比べて、対象言語を使用する読者に対して本来用語が持つべき意味 (すなわち概念) をより的確に伝えることができると仮説できる。この仮説に基づき、翻訳クオリティを認知的な観点から評価する方法を検討することも、今後の重要な課題であると考えている。

## 6. まとめ

本稿では、単語の翻訳における認知プロセスに関する研究で De Groot (1997) が提唱する概念特徴モデルと、ターミノロジー手法の一つとして Madsen (2004)らが提唱する知識ベースアプローチとの間に共通点を見出し、ターミノロジー手法を活用して構築したバイリンガルオントロジーをマッピングすることで、希少な組み合わせの言語間における用語の翻訳について、用語がもつ概念特徴を活用して、用語を認知的に翻訳する方法を提案した。

## 文献


De Groot, A. M. B., & Hoeks, C. J. (1995) *The development of bilingual memory: Evidence from word translation by trilinguals.* Language Learning, *45*, 683-724.

De Groot, A. M. B. (1997). *The cognitive study of translation and interpretation: Three approaches*. In J.H. Danks, G. M Shreve, S. B. Fountain, & M. K. McBeath (Eds), Cognitive processes in translation and interpretation, pp. 25-56. Thousand Oaks, CA: Sage Publications.

Kroll, J. F. & Stewart, E. (1994). *Category interference in translation and picture naming: Evidence for asymmetric connections between bilingual memory representations.* Journal of Memory and Language, 33, pp. 149-174.

Madsen, B. N. (1999). *Terminologi 1. Principper og metoder*. Copenhagen: Gads Forlag.

Madsen, B. N., Thomsen, H. E. & Vikner, C.. (2004a). *Principles of a system for terminological concept modelling*. In: Proceedings of the 4th International Conference on Language Resources and Evaluation, Vol. I. Lisbon, pp. 15-18.

Madsen, B. N., Thomsen, H. E. & Vikner, C.. (2004b). *Comparison of principles applying to domain specific versus general ontologies*. In: Alessandro Oltramari, Patrizia Paggio, Aldo Gangemi, Maria Teresa Pazienza, Nicoletta Calzolari, Bolette Sandford Pedersen, Kiril Simov (eds.): OntoLex 2004: Ontologies and Lexical Ressources in Distributed Environments. ELRA, pp. 90-95.

Madsen, B. N., Thomsen, H. E. & Vikner, C.. (2005). *Multidimensionality in terminological concept modelling*.
In: Bodil Nistrup Madsen, Hanne Erdman Thomsen (eds.): *Terminology and Content Development*, TKE 2005, 7th International Conference on Terminology and Knowledge Engineering, Copenhagen, pp. 161-173.

Potter, M.C., So, K.F., von Eckardt, B., & Feldman, L.B. (1984). *Lexical and conceptual representation in beginning and proficient bilinguals*. Journal of Verbal Learning and Verbal Behavior, 23, pp. 23-38