機械翻訳向けプリエディットのための情報明示化方略の体系化

島田紗裕華 [1]　山口大地 [1]　宮田玲 [2]　藤田篤 [3]　佐藤理史 [2]
[1] 名古屋大学工学部　[2] 名古屋大学大学院工学研究科　[3] 情報通信研究機構
shimada.sayuka@b.mbox.nagoya-u.ac.jp

概要

機械翻訳向けに入力文を編集するプリエディットの作業を支援する枠組みと技術の開発が求められている．我々は，翻訳しやすく原文を編集するための指針として，原文の情報を明示化する書き換え方略に注目し，人手プリエディット事例のボトムアップな分析により，6 カテゴリ・26 サブカテゴリからなる明示化方略体系を構築した．また，検証事例を用いて，構築した体系の網羅性を確認した．さらに，段階的に定義したプリエディットのプロセスにおける要素技術の開発の見通しを立てた．

1　はじめに

機械翻訳（MT）を活用する方法の 1 つとして，翻訳対象である起点テキストを事前に編集するプリエディットがある．プリエディットの有効性は様々な研究・実践において示されてきたが [1, 2, 3]，翻訳結果を修正するポストエディットに比べて，実務翻訳での活用は十分進んでいない．入力文に対するプリエディット結果を直接出力する手法も提案されているが [4]，実用場面においてプリエディットは主に人手でなされており，そのプロセスを支援する枠組みと技術の開発が求められている．我々は，人間による段階的な意志決定を支援する視点から，プリエディットのプロセスを，(1) 翻訳しにくい表現の検出，(2) 検出した表現の分類，(3) 翻訳しやすい表現に書き換えるために適用可能な方略の列挙，(4) 最適な方略の選択，(5) 方略の適用事例の生成，(6) 最適な方略適用結果の選択に分けた（図 1 参照）．そして各ステップについて，人間はどのような行為を行っているか，機械による自動化は可能であるか，ということを検討しながら人間と機械の適切な役割分担を見極めることを目指している．

このプロセスを具体的に検討するにあたり，MT にとって翻訳しにくい表現とはどのようなものか，翻訳しやすくするためにどのような書き換え方法が

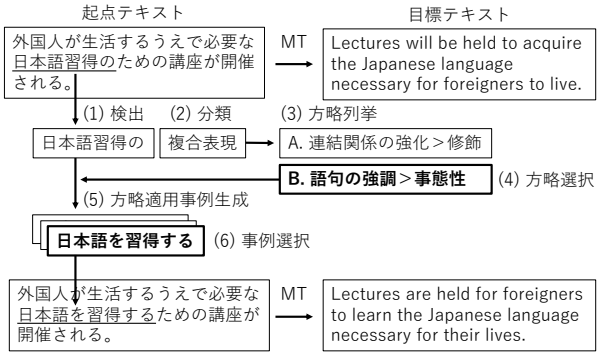

図 1　プリエディットのプロセスと機械翻訳の例

有効なのか，に関する言語化された指針が必要である．このような指針は制限言語などの形で定義・適用・評価されてきたが [5, 6, 7]，それらは主にルールベース MT や統計的 MT を対象としており，現在主流の方式であるニューラル MT を対象とした研究は限られている [3, 8]．そもそも挙動が予測できない MT を後段に据えるプリエディットでは，個々の書き換えにおいて翻訳品質の向上を保証することが原理的に難しい．また，書き換えによる出力の変化は，使用する MT システムにも依存する [7, 8]．以上を前提としつつも，我々は，ある程度一般に有効な書き換え方略を定めることは有効だと考える．

宮田・藤田 [3] は，複数の翻訳方向・ニューラル MT システム・テキストドメインを対象とした人手プリエディット事例の観察を通じて，情報を明示化するような書き換えが翻訳品質向上において重要であると指摘した．さらに，明示化方略として，情報の追加，関係の明示，意味の限定，正規化の 4 つを提示しているが，それ以上細かい方略は定めていない．そこで我々は，プリエディット支援技術の開発に向けた第一ステップとして，明示化方略に関する詳細な体系を構築した．

本稿では，以下，人手プリエディット事例のボトムアップな分析による明示化方略の体系構築の方法と結果について報告する（2 節）．また，体系の構築時に使用していない事例を用いて，体系の網羅性を

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

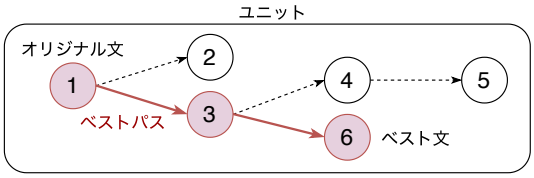

図 2　プリエディット事例におけるユニットとベストパス

予備的に検証した結果を示す（3 節）．最後に，上記プロセスを対象に，方略体系をベースとした各種の技術開発の見通しについて議論する（4 節）．

2　情報明示化方略の体系構築

2.1　構築方法

方略体系の構築には，宮田・藤田 [3] が収集した人手プリエディット事例データを用いた．この事例データは，MT 訳が一定の翻訳品質[1)]に達するまで，作業者が試行錯誤的に起点テキストに対する最小単位の書き換え [9] を繰り返すことで構築したものである．書き換え前後の起点テキストの各ペアを事例と呼ぶ．最初に与えられる起点テキスト（**オリジナル文**）とそれに対する書き換えの履歴をまとめて**ユニット**と呼び，作業終了時に最も高い MT 品質を達成した起点テキストを**ベスト文**と呼ぶ（図 2 参照）．1 つのユニットにおけるプリエディットの履歴は，各事例をノードとし，書き換え前後の関係にあるノード間をエッジで結ぶことで，オリジナル文を根とする木構造として表現できる．この木構造において，オリジナル文からベスト文までの経路を**ベストパス**と呼ぶ．

宮田・藤田 [3] は，100 のオリジナル文を対象に，3 つの翻訳方向（日英，日中，日韓），2 つの MT システム（Google 翻訳[2)]と TexTra[3)]）の組み合わせの各々（計 600 ユニット）について，事例を収集している．我々はまず，30 のオリジナル文に対応した計 180 ユニットにおけるベストパス中の事例 757 件を抽出した．各事例は，宮田・藤田 [3] により，情報方略の観点から明示化，情報維持，暗示化のいずれかに分類されている．我々は，明示化方略と分類された事例を改めて精査し，最終的に 269 件の明示化方略事例を獲得し，体系構築の開発データとした．

各事例について，どのような表現に対して，どのような書き換えがなされ，どのような情報が明示されたかの 3 つの観点で分析し，それぞれ，対象，操作，明示内容としてラベルを付与した．分析の過程で，操作の観点を第 1 階層，明示内容の観点を第 2 階層のカテゴリ基準とする分類体系の構造を定めた．全事例に対して，筆頭著者がラベル付けと体系化を行い，共著者が結果を確認する，というサイクルを複数回繰り返すことで，各分類カテゴリの名称・範疇および分類基準を精緻化した．

2.2　構築した情報明示化方略体系

構築した明示化方略体系の全体像を表 1 に示す．明示化方略は，操作の観点から 6 カテゴリに分類され，さらに明示内容の観点から 26 サブカテゴリに細分類される．「開発」列には開発データ中の事例の頻度分布を，「検証」列には後述する検証データにおける頻度分布を示す．以下，各操作カテゴリの定義と具体例を述べる（具体例の一覧は付録を参照）．

(I) 暗示的情報の表出：前後の文脈や外部知識を用いて，テキスト中に明言されていない情報を追加する操作である．例えば，「テルモ」から「医療機器メーカーのテルモ」への書き換え（(I1) **カテゴリ**）や「１２日は台湾の休日のため休場」から「１２日は台湾の休日のため株式市場は休場」への書き換え（(I2) **主題**）が含まれる．

(C) 連結関係の強化：助詞や接続詞などの語句を連結する表現や照応表現を用いて，文間や語句間の関係を明示する操作である．例えば，(C6) **展開**は，節間の順接または逆接関係について事態の展開を明示する操作を指し，「河川や湖沼が氾濫し、住宅やハイウェーが浸水しました」から「河川や湖沼が氾濫し、その結果住宅やハイウェーが浸水しました」への書き換え事例が該当する．

(B) 境界の強調：語句や節，文の区切り目を約物を使用して記号的に明示する操作である．(B2) **語句の境界**に分類される例としては，「台湾国際貿易局」を引用符で囲み「“台湾国際貿易局”」へと書き換える例が挙げられる．

(P) 語句の強調：意味や情報量が変化しない範囲で語形の変更や語句の追加を行い，語句の特定の内容を強調するような操作である．(P2) **事態性**には，「政府調達協定への加盟」から「政府調達協定に加盟すること」への書き換えや，「署長」から「署長を務めていた人物」への書き換えが含まれる．

1）「語句の選択がやや不自然であるが、原文の情報が完全に翻訳され、訳文は文法誤りを含まない」レベル以上を基準とした．ただし，この基準に達しないと判断した場合は，途中で作業を終了する．
2）https://translate.google.com
3）https://mt-auto-minhon-mlt.ucri.jgn-x.jp

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

表 1　明示化方略体系と事例の頻度分布

分類カテゴリ	事例数 開発	検証
(I) 暗示的情報の表出		
(I1) カテゴリ	17	13
(I2) 主題	15	11
(I3) 性・数	4	2
(I4) メタ的情報	3	0
(I5) 対象	2	4
(I6) 語意	4	5
(I7) 補足情報	7	7
(C) 連結関係の強化		
(C1) 時	11	1
(C2) 範囲	7	2
(C3) 引用・発言	6	0
(C4) 参照	4	5
(C5) 並列	5	3
(C6) 展開	17	12
(C7) 修飾	28	18
(C8) 背景	13	6
(B) 境界の強調		
(B1) 引用・発言内容の範囲	4	0
(B2) 語句の境界	16	24
(P) 語句の強調		
(P1) 発音	7	8
(P2) 事態性	22	42
(P3) ニュアンス	22	43
(L) 語義の絞り込み		
(L1) 語義	16	3
(L2) 意味	10	18
(N) 正規化		
(N1) 漢字の使用	8	7
(N2) 省略回避	8	1
(N3) 体言止め回避	12	4
(N4) 句点挿入	1	0
合計	269	239

(L) 語義の絞り込み：語の多義性を解消する操作である．(L1) **語義**は語義曖昧性の解消に関するもので，「はんらん」を「氾濫」と書き換えるような事例が該当する．(L2) **意味**は語句をより狭義の表現へ書き換えて語句の意味を明示するカテゴリである．「１４１バーツとなった」から「１４１バーツを記録した」へ書き換える事例が該当する．

(N) 正規化：より一般的・標準的・辞書的な表記・文体に書き換える操作である．他のカテゴリと異なり，明示内容ではなく操作の観点から下位分類を定めた．例えば，(N1) **漢字の使用**は，「ひきこもり」から「引きこもり」への書き換えを含む．

3　明示化方略体系の予備的検証

2.1 節で説明したプリエディット事例データから，検証データを用意し，構築した明示化方略体系の網羅性を検証した．開発データに含まれないオリジナル文 15 文に対する計 90 ユニットのベストパスに含まれる書き換え事例 363 件から，明示化であると判断した事例 247 件を抽出し，検証データとした．

これらの全事例を，事前に定めた方略の定義に基づき，筆頭著者が方略体系の各カテゴリに分類し，判断に迷った事例について共著者が確認した．表 1 の「検証」列に分類結果を示す．247 件のうち 239 件をいずれかのカテゴリに分類することができた．分類できなかった事例が 8 件（約 4 %）に留まったことから，構築した方略体系は明示化方略の種類を広くカバーできていると言える．

分類できなかった事例の一部を表 2 に示す．1 つ目の例では，「アジアと欧州市場でドルは」という主題とそれに対する述部「すでに上昇していた」の位置を近づけることで主述関係が明確化されており，(C) **連結関係の強化**に分類できるものの，その下位分類を定めることができなかった．2 つ目の例は，文中で列挙された要素が「3 つ」であるという情報を明示していることから，(I) **暗示的情報の表出**に含まれる．しかし，下位分類はいずれも語句の内容に関する明示化を対象としており，表現自体の位置づけに関わる本事例は分類できなかった．これらの事例を網羅できるよう，新しいカテゴリの追加や既存カテゴリの定義・範疇の見直しが必要である．

また，分類可能ではあったが，体系構築の際にあまり想定していなかった特徴を持つ事例もあった．「増加」を「増産」に書き換える事例は，増加させる対象が生産量であるという情報を追加したものであり，(I) **暗示的情報の表出**＞ (I5) **対象**に分類できたものの，このカテゴリでは元々，「売る」を「持株を売る」に書き換えるような要素の単純な追加のみを想定していた．

最後に，事例数が比較的多い分類カテゴリとして，(P) **語句の強調**における (P2) **事態性**と (P3) **ニュアンス**について述べる．(P2) **事態性**は，語句の事態性を明示する書き換え操作である．これに分類した事例の特徴は大きく 2 つに分けることができる．一方は「政府調達協定への加盟」から「政府調達協定に加盟すること」への書き換えのように動作性・能動性を強調するもの，他方は「パニック」を「パニック状態」へ書き換えるように状態性・受動性を強調するものである．(P3) **ニュアンス**は，情報量があまり変化しない語形の変更や語句の追加を想定しており，他の分類カテゴリのいずれにも分類しにく

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

表 2 構築した明示化方略体系で分類できなかった事例

	書き換え前	書き換え後
1	アジアと欧州市場でドルは、９日発表の経済指標で、独失業が史上最悪であることが示されるであろうとの見方で、すでに上昇していた。	９日発表の経済指標で、独失業が史上最悪であることが示されるであろうとの見方で、アジアと欧州市場でドルは、すでに上昇していた。
2	普通科、専門学科 (工業科、商業科、農業科など)、総合学科に分かれます。	普通科、専門学科 (工業科、商業科、農業科など)、総合学科の 3 つに分かれます。

い事例を消極的に含めることができるように定義していた．例えば「低位株」から「超低位株」へ，程度に関するニュアンスを強める書き換えは，このカテゴリに含まれる．プリエディットの操作をなるべく具体的に言語化するという目的をふまえると，(P2) 事態性と (P3) ニュアンスについては，より詳細に下位分類を展開したり特定の特徴を認定し新たなカテゴリを立てたりすることが必要だろう．ただし，体系の利用の観点からは，このような詳細化・拡張により，体系が不必要に複雑になる可能性がある点に注意しなければならない．

4 プリエディット技術の開発方針

構築した明示化方略体系に基づいて，1 節で提示したプリエディットのプロセスの各ステップに関する技術開発の見通しを述べる（図 1 参照）．

(1) 翻訳しにくい表現（本研究では，明示化可能表現）の検出と (2) 検出した表現の分類には，明示化可能表現の体系が必要である．2，3 節におけるプリエディット事例の分析過程で抽出・列挙された明示化対象の表現を適度な抽象度でまとめ上げながら，体系を構築する．このとき例えば，(C) 連結関係の強化＞ (C6) 展開の事例において観察された「因果関係をつなぐ連用中止接続」という明示化可能表現を，単に「連用中止接続」と抽象化すると，明示化方略が適用できない多くの表現も該当してしまう点に注意が必要である．

明示化可能表現の自動検出・分類には，各種の既存手法が適用可能である．特定の言語表現パターンで定義できる場合は，ルールベースの方法 [10] が適している．また，機械学習を用いた系列ラベリング問題として定式化することもできる．機械学習に必要なアノテーション済みの大規模なプリエディットデータは整備されていないが，BERT [11] などの事前学習済み言語モデルを活用することで，比較的小規模な教師データでも一定の精度が得られる可能性がある．また，MT 結果を利用できる設定であれば，翻訳誤り箇所を推定しそれに対応した原文箇所を同定する手法 [12] や折り返し翻訳を用いて原文中の翻訳しにくいテキストスパンを検出する手法 [13] を組み合わせることで，特に書き換えが必要な表現への絞り込みが期待できる．

(3) 適用可能な方略の列挙のためには，明示化可能表現と明示化方略の対応表を事前に定義する．ただし，1 つの表現に対して，複数の方略が適用可能な場合もあるため，(4) 最適な方略（あるいはその組み合わせ）を選択する必要がある．これに対して，対応表にあらかじめ方略の優先度を付与しておく方法や個々の明示化可能表現に対して動的に方略の優先度を定める方法がある．後者の方法については，複数の方略でプリエディットを最後まで行い，各々に対する MT 結果も考慮しながら品質推定の手法を用いて，優先度を求める方法も考えられる．

(5) 方略の適用事例の生成，(6) 最適な方略適用結果の選択については，現時点で自動化に向けた見通しは十分立っていない．例えば，(N) 正規化＞ (N5) 体言止めの回避であれば，HaoriBricks3 [14] など既存の自然言語処理ツールで実現できる．また，(L) 語義の絞り込み＞ (L1) 語義は，語義曖昧性解消のテーマで長年取り組まれてきた研究成果 [15, 16] を生かせるだろう．しかし，(I) 暗示情報の表出をはじめ，明示化方略を適用するためには，対象文に書かれていない文脈情報や外界知識が必要となることが多く，完全な自動化は容易ではない．まずは，人間の役に立つ候補の提示を目指すのが現実的である．

5 おわりに

本研究では，プリエディットにおける明示化に焦点を当て，書き換えの操作と明示内容の観点から階層化した明示化方略体系を構築した．予備的な検証結果から，構築した体系の比較的高い網羅性を示せたが，さらに異なる条件での本検証結果をふまえた体系の見直しが必要である．

また，明示化方略体系をベースとしたプリエディット支援技術の開発の見通しを提示した．今後は要素技術の開発を進めると同時に，機械的にできること・できないことを見極め，翻訳の効率と品質の向上に効果的なワークフローの定義を目指す．

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

謝辞

本研究は科研費（課題番号：19K20628, 19H05660）および KDDI 財団調査研究助成（課題名：平易な文化財情報を執筆・翻訳する技術）の支援を受けた.

参考文献

[1] Yusuke Hiraoka and Masaru Yamada. Pre-editing plus neural machine translation for subtitling: Effective pre-editing rules for subtitling of TED talks. In **Proceedings of Machine Translation Summit XVII**, pp. 64–72, Dublin, Ireland, 2019.

[2] 土井惟成. プリエディット手法としての産業日本語に関する一考察. **Japio YEAR BOOK 2020**, pp. 324–330, 2020.

[3] Rei Miyata and Atsushi Fujita. Understanding pre-editing for black-box neural machine translation. In **Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)**, Online, 2021.

[4] Sneha Mehta, Bahareh Azarnoush, Boris Chen, Avneesh Saluja, Vinith Misra, Ballav Bihani, and Ritwik Kumar. Simplify-then-translate: Automatic preprocessing for black-box translation. In **Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI)**, New York, USA, 2020.

[5] Arendse Bernth and Claudia Gdaniec. MTranslatability. **Machine Translation**, Vol. 16, No. 3, pp. 175–218, 2001.

[6] Sharon O'Brien and Johann Roturier. How portable are controlled language rules? In **Proceedings of the Machine Translation Summit XI**, pp. 345–352, Copenhagen, DK, 2007.

[7] Rei Miyata, Anthony Hartley, Cécile Paris, Midori Tatsumi, and Kyo Kageura. Japanese controlled language rules to improve machine translatability of municipal documents. In **Proceedings of the Machine Translation Summit XV**, pp. 90–103, Miami, Florida, USA, 2015.

[8] Shaimaa Marzouk. An in-depth analysis of the individual impact of controlled language rules on machine translation output: A mixed-methods approach. **Machine Translation**, Vol. 35, pp. 167–203, 2021.

[9] 藤田篤, 柴田知秀, 松吉俊, 渡邉陽太郎, 梶原智之. 言い換え認識技術の評価に適した言い換えコーパスの構築指針. 言語処理学会第 21 回年次大会ワークショップ自然言語処理におけるエラー分析発表論文集, 2015.

[10] 白井諭, 池原悟, 河岡司, 中村行宏. 日英機械翻訳における原文自動書き替え型翻訳方式とその効果. 情報処理学会論文誌, Vol. 36, No. 1, pp. 12–21, 1995.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional Transformers for language understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)**, pp. 4171–4186, Minneapolis, Minnesota, USA, 2019.

[12] Lucia Specia, Frédéric Blain, Marina Fomicheva, Erick Fonseca, Vishrav Chaudhary, Francisco Guzmán, and André F. T. Martins. Findings of the WMT 2020 shared task on quality estimation. In **Proceedings of the 5th Conference on Machine Translation (WMT)**, pp. 743–764, Online, 2020.

[13] Kiyotaka Uchimoto, Naoko Hayashida, Toru Ishida, and Hitoshi Isahara. Automatic detection and semi-automatic revision of non-machine-translatable parts of a sentence. In **Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)**, pp. 703–708, Genoa, Italy, 2006.

[14] 佐藤理史. HaoriBricks3: 日本語文を合成するためのドメイン特化言語. 自然言語処理, Vol. 27, No. 2, pp. 411–444, 2020.

[15] Roberto Navigli. Word sense disambiguation: A survey. **ACM Computing Surveys**, Vol. 41, No. 2, pp. 1–69, 2009.

[16] Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. Recent trends in word sense disambiguation: A survey. In **Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI), Survey Track**, pp. 4330–4338, Online, 2021.

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).

付録

表 3 に情報明示化方略体系の各カテゴリの事例を載せる。実際の書き換え事例からテキストを部分的に抜粋した。

表 3　情報明示化方略の例

分類	書き換え前	書き換え後
（I）暗示的情報の表出		
（I1）カテゴリ	テルモから発売された。	医療機器メーカーのテルモから発売された。
（I2）主題	１２日は台湾の休日のため休場。	１２日は台湾の休日のため株式市場は休場。
（I3）性・数	ひきこもりの若者の就労	ひきこもりの若者たちの就労
（I4）メタ的情報	英ナショナル・ウェストミンスター・バンク（ナットウェスト）	英ナショナル・ウェストミンスター・バンク（通称ナットウェスト）
（I5）対象	米大統領選の民主党指名争い	米大統領選の民主党候補者指名争い
（I6）語意	ノルトライン・ウエストファーレン州	ノルト（北）ライン・ウエストファーレン州
（I7）補足情報	整理券は紛失したり折ったりすると無効になります。	お配りした整理券は紛失したり折ったりすると無効になります。
（C）連結関係の強化		
（C1）時	０１年４月刊行の「続続続筆とエンピツ」	０１年４月に刊行の「続続続筆とエンピツ」
（C2）範囲	この一、二年で急速に整った。	この一、二年の間に急速に整った。
（C3）引用・発言	最新の情報では、	最新の情報によると、
（C4）参照	連銀	同連銀
（C5）並列	独ノルトライン・ウエストファーレン州とバーデン・ビュルテンベルク州	独ノルトライン・ウエストファーレン州及びバーデン・ビュルテンベルク州
（C6）展開	河川や湖沼が氾濫し、住宅やハイウェーが浸水しました。	河川や湖沼が氾濫し、その結果住宅やハイウェーが浸水しました。
（C7）修飾	古文字（こもじ）の研究・論文などの著作	古文字（こもじ）についての研究・論文などの著作
（C8）背景	ドイツ連邦銀行の副総裁ガッデム	ドイツ連邦銀行の副総裁であるガッデム
（B）境界の強調		
（B1）引用・発言内容の範囲	来週１２日の理事会で、連銀準備再評価の政府改定案について協議する、と述べた。	「来週１２日の理事会で、連銀準備再評価の政府改定案について協議する」、と述べた。
（B2）語句の境界	台湾国際貿易局の広報担当者	“台湾国際貿易局” の広報担当者
（P）語句の強調		
（P1）発音	“職親制度”	“職親制度（しょくおやせいど）”
（P2）事態性	ＷＴＯの政府調達協定への加盟には合意する	ＷＴＯの政府調達協定に加盟することには合意する
（P3）ニュアンス	医療機器メーカーのテルモから発売された血糖測定器	医療機器メーカーのテルモから新発売された血糖測定器
（L）語義の絞り込み		
（L1）語義	河川や湖沼のはんらん	河川や湖沼の氾濫
（L2）意味	１４１バーツとなった。	１４１バーツを記録した。
（N）正規化		
（N1）漢字の使用	ひきこもり	引きこもり
（N2）省略回避	連銀	連邦銀行
（N3）体言止め回避	単位は億円。	単位は億円です。
（N4）句点挿入	業者名、申込方法、料金などを問い合わせましょう	業者名、申込方法、料金などを問い合わせましょう。

This work is licensed by the author(s) under CC BY 4.0
(https://creativecommons.org/licenses/by/4.0/).