システム開発
１９－Ｆ－１２

経済活性化のための
技術用日本語プラットフォームの開発
に関するフィージビリティスタディ

報　　告　　書

― 要　旨 ―

平成２０年３月

財団法人　機械システム振興協会

委託先　財団法人　日本特許情報機構

この事業は、競輪の補助金を受けて実施したものです。

http://ringring-keirin.jp/

序

わが国経済の安定成長への推進にあたり、機械情報産業をめぐる経済的、社会的諸条件は急速な変化を見せており、社会生活における環境、都市、防災、住宅、福祉、教育等、直面する問題の解決を図るためには技術開発力の強化に加えて、多様化、高度化する社会的ニーズに適応する機械情報システムの研究開発が必要であります。

このような社会情勢の変化に対応するため、財団法人機械システム振興協会では、財団法人日本自転車振興会から機械工業振興資金の交付を受けて、システム技術開発調査研究事業、システム開発事業、新機械システム普及促進事業を実施しております。

このうち、システム技術開発調査研究事業及びシステム開発事業については、当協会に総合システム調査開発委員会(委員長：東京大学名誉教授 藤正 巖氏)を設置し、同委員会のご指導のもとに推進しております。

本「経済活性化のための技術用日本語プラットフォームの開発に関するフィージビリティスタディ」は、上記事業の一環として、当協会が財団法人日本特許情報機構に委託し、実施した成果をまとめたもので、関係諸分野の皆様方のお役に立てれば幸いであります。

平成２０年３月

財団法人 機械システム振興協会

はじめに

1990 年代、インターネットが成長し、世界中の言語情報が流通するようになった。そして、2000 年代に入ると Web サービスが爆発的な普及を見せている。こうしたなか世界の情報は、英語をグローバルスタンダードとして動きつつあり、米国や欧州では、そのような状況に向けた多くの対応が見られる。例えば、英語圏においては、とりわけ産業用ドキュメントの作成に用いられる英語を、そのあるがままに使用するのではなく、国際言語として多民族(ノンネイティブ)が意思疎通できるように正確でわかり易くするための実践が行われてきた。例としては、キャタピラー社、IBM、ゼロックス、ジェネラルモーターズ等の多国籍企業が、英文マニュアルを多言語に翻訳しやすくするために独自の制限英語を作っている。また、欧州と米国の航空業界は、メンテナンスマニュアルのために Simplified Technical English（STE）を制定した。更に、機械翻訳システムに関しては、ＥＵ23 カ国の言語間のシステムが実用レベルで利用されている。

他方、わが国において、知的基盤となるのが日本語であることはいうまでもないが、日本語テクニカルライティングに関しては、実践的な規定集が関連団体や企業で集積されているものの、いわば口述伝承のようなスキルに留まり、業界横断的な本格的な規格作りの動きは見られない。従って、企業の現場レベルで潜在的に高度なモノ作りの匠を保有していたとしても、必須の英語によるコミュニケーション不足のために、とりわけ技術情報の発信において、国際化が進まない状況がある。日本語をこのままの状況で放置しておけば、英語圏との情報ギャップが拡大する一方である。その対応策として、まずは人間のコミュニケーション能力の向上が挙げられる。しかし、併せて機械翻訳の品質と効率を高めるため、原文に着目し、機械処理に適した産業用日本語をベースとしたドキュメント作成を、普及促進することが、わが国の将来にとって喫緊の課題である。この点に関し、わが国における日本語の解析、機械翻訳に関する自然言語処理技術は、世界的に最高の水準にある。これらを有効に活かし、真に情報発信の国際化を進めることが、わが国の発展のために欠かせない重要な取り組みである。

本スタディの報告書は、このような現状認識のもと、わが国経済の活性化、国際競争力の強化に資する、インターネット時代に相応しい仕様の「技術用日本語」の普及促進を推し進めるための技術用日本語プラットフォーム開発計画を、産業界に初めて示した画期的なものである。このプラットフォームを通して、日本産業界に、標準となる日本語（技術用日本語）が根付き、日本企業の国際競争力向上のための知的基盤強化の一助ともなれば幸いである。

本スタディを実施するにあたり、財団法人機械システム振興協会のご高配に深謝するとともに、本スタディにご協力いただいた関係各位に心から謝意を表する次第である。

平成２０年３月

財団法人　日本特許情報機構
専務理事 兼 特許情報研究所 所長
寺本　義憲

目　次

１　スタディの目的

英語圏においては、1978 年カータ米国大統領令によって Plain English が奨励され、それが法律、金融、公共の場面に広く浸透し、産業、商業用文書の作成に用いられる英語に関しても正確でわかり易くするための実践が行われてきている。

しかし、わが国において紛れもなく知識基盤となっている日本語については、業界横断的なそうした取り組みは見られず、技術情報の発信において、国際化が進まず、資源としての言語の活用も遅れ、英語圏との情報ギャップが拡大する一方である。

その一方で、わが国における日本語の解析、機械翻訳に関する自然言語処理技術は、世界最高水準にあり、こうした技術の活用は、英語圏におけるテクニカルライティングの水準を超える可能性を秘めている。

このような状況下、経済の活性化、国際競争力の強化のために、産業用ドキュメントの作成に用いる日本語をインターネット時代に適応した仕様の日本語、すなわち「技術用日本語」に変革することが一つの解決策になると考えている。

しかし、そうした技術用日本語の必要性を提唱するだけでは、状況の改善がなされない。

そこで、技術用日本語導入を実践的かつ具体的に進めるため、実際に技術用日本語プラットフォームを開発し、そのプラットフォームを浸透させる活動を通じて、日本産業界に技術用日本語を根付かせ、わが国の知識基盤の強化を図ることを最終目標とする。

「技術用日本語プラットフォーム」とは、技術用日本語を扱うための基盤システムであり、①技術用日本語プラットフォームシステム及び②技術用日本語アプリケーションシステムからなる。まず、①技術用日本語プラットフォームシステムとしては、技術用日本語の辞書データベースからなる「技術用日本語文書作成集合知サーバ」を基本モジュールとして、その上に、技術用日本語を作成支援するためのワードプロセッサ的役割を果たす「技術用日本語オーサリングシステム」が載る。そして、②技術用日本語アプリケーションシステムとして、「技術用日本語日英機械翻訳システム」と、概念表現レベルでの構造検索を可能にする「技術用日本語文書検索システム」が加わる。

こうした「技術用日本語プラットフォーム」をベースとして、将来的には、中小ベンチャー企業用海外ビジネスドキュメント作成支援システム、TLO 用外国出願作成支援システムなどの個別の技術用日本語アプリケーション/サービスの開発・提供も想定される。

本年度のスタディは、上記①及び②の両システムの最も基本となる「技術用日本語共通基盤」を含む、これらのシステム自体を対象とする基本的な「開発計画」を提案することを目的とする。

また、インターネット時代の技術用日本語の必要性を確認しつつ、技術用日本語を用いた文書処理の高度化、高精度化、高能率化によりもたらされる経済活動の活性化の形態を、その波及効果及び事業展開の具体的イメージを示すことで明らかにし、更に、技術用日本語導入による効果を、実験を通じて検証することも目的としている。

２　スタディの実施体制

本スタディの実施体制として、図１に示すとおり、（財）機械システム振興協会内に総合システム調査開発委員会を、また（財）日本特許情報機構内に、技術用日本語プラットフォーム委員会を設置した。そして各作業は、技術用日本語プラットフォーム委員会での審議を経て着手し、その結果を委員会で審議した。業務分担として、開発課題の考察については、技術用日本語プラットフォーム委員会の主導で実施し、一部のサブシステム（概念構造検索系、辞書サーバ、技術用日本語共通基盤仕様、日本語処理系）の検討は再委託した。

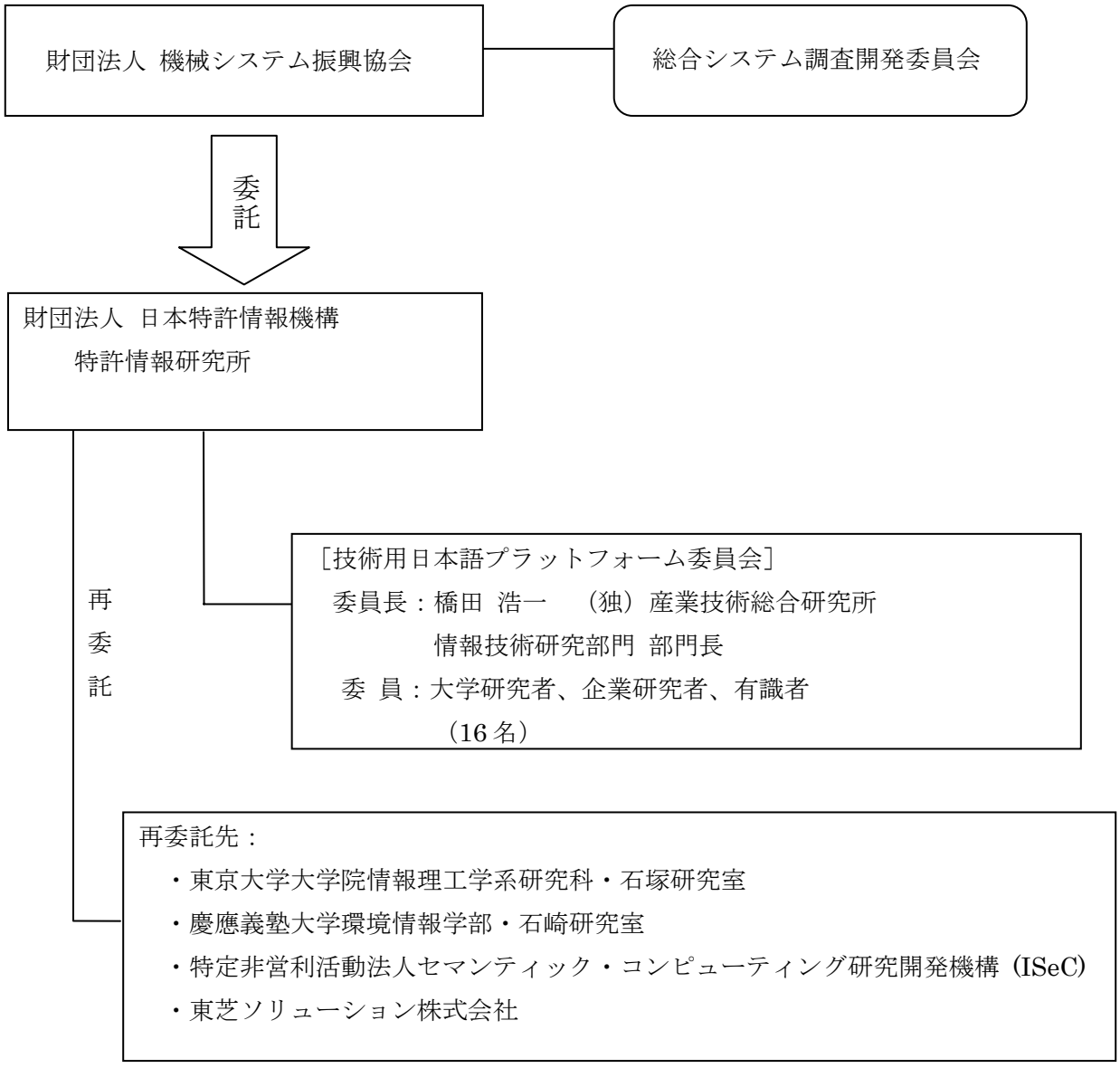


図１：委託事業実施体制

総合システム調査開発委員会委員名簿

（順不同・敬称略）

委員長	東京大学 名誉教授	藤　正　　巖
委　員	埼玉大学 総合研究機構 地域共同研究センター 教授	太　田　公　廣
委　員	独立行政法人産業技術総合研究所 エレクトロニクス研究部門 副研究部門長	金　丸　正　剛
委　員	独立行政法人産業技術総合研究所 産学官連携推進部門 産学官連携コーディネータ	志　村　洋　文
委　員	東北大学大学院 工学研究科　教授 （未来科学技術共同研究センター長）	中　島　一　郎
委　員	東京工業大学大学院 総合理工学研究科 教授	廣　田　　薫
委　員	東京大学大学院 工学系研究科 准教授	藤　岡　健　彦
委　員	東京大学大学院 新領域創成科学研究科 教授（副研究科長）	大　和　裕　幸

技術用日本語プラットフォーム委員会委員名簿

（順不同・敬称略）

委員長	独立行政法人産業技術総合研究所 情報技術研究部門　部門長	橋　田　浩　一
副委員長	東京工科大学 メディア学部　教授	横　井　俊　夫
委　員	東京大学大学院 情報理工学系研究科　教授	石　塚　　満
委　員	慶應義塾大学 環境情報学部　教授	石　崎　　俊
委　員	名古屋大学大学院文学研究科 研究科長	町　田　　健
委　員	独立行政法人国立国語研究所 研究開発部門言語資源グループ グループ長	山　崎　　誠
委　員	IRD 国際特許事務所 所長・弁理士	谷　川　英　和
委　員	トヨタ自動車株式会社 知的財産部　企画統括室　主幹　弁理士	佐　野　夏　茂
委　員	キヤノン株式会社 知的財産法務本部　副本部長	大　野　　茂
委　員	(株)富士通研究所 ソフトウェア＆ソリューション研究所　主管研究員	潮　田　　明
委　員	(株)日立製作所 システム開発研究所 uValue イノベーションセンタ　主任研究員	間　瀬　久　雄
委　員	(株)三菱総合研究所 情報技術研究センター　主席研究員	白　井　康　之
委　員	(株)ジャストシステム イノベーションテクノロジー研究開発部	荒　川　直　哉
委　員	(株)日本システムアプリケーション 顧問	荻　野　孝　野
委　員	(株)東芝 研究開発センター知識メディアラボラトリー 主任研究員	熊　野　　明
事務局	(財)日本特許情報機構 専務理事 兼 特許情報研究所 所長	寺　本　義　憲
事務局	(財)日本特許情報機構 特許情報研究所　調査研究部長	奥　　直　也
事務局	(財)日本特許情報機構 特許情報研究所　企画調査課長	大　塩　只　明

３　スタディ成果の要約

本スタディは、平成１９年１０月から平成２０年２月にかけて、以下の４つのステップで実施した。

３－１　技術動向調査
３－２　開発課題の考察
３－３　技術用日本語の検証実験
３－４　市場調査、事業シミュレーション

以下に、各作業における成果の概要を述べる。

３－１　技術動向調査

技術用日本語の必要性及び有用性を検証し、技術用日本語プラットフォームの開発に役立つ、日本語への取り組みや有用技術を積極的に把握するため、関係者へのヒアリングや文献調査を行った。

国内については、関連する技術の現状を調査した結果、技術用プラットフォームの開発に資する技術及び活動が多数存在することがわかり、また、それぞれのシステムが抱える課題を技術用日本語が解決する可能性がきわめて高いこともわかった。併せて、技術用日本語プラットフォームシステムの基幹技術であるＣＤＬ（概念記述言語）及びセマンティックオーサリングについて、その現状をまとめた。

海外の技術動向調査においては、技術用日本語の規則知識にも関連するオントロジーを記述するには、グラフィカルなアプローチが優位であることが明らかになった。この点は、訪問調査においても確認している。また、制限英語に関しては、テクニカルライティングのために、人間用の制限英語と、機械処理での利用を前提とした制限英語が存在していることがわかったが、両方を同時に満足する制限言語は、国内においても海外においても存在しないことがわかった。技術用日本語は、人間用でもあり、機械用でもある新しい言語となる。

以下に主要点を示す。

３－１－１　日本語の規格化の技術動向

（1）日本語に関する技術動向

技術用日本語は、正確で分かりやすく簡潔な日本語のサブセットを如何に設定するかの問題である。そのために必要となる技術の動向を調査して、本スタディの**開発計画**の妥当性を得ることであった。

① 日本語ワープロの機能としての校正・推敲支援機能

構文的な曖昧さをもたらすような修飾関係（例：「白いカゴの中の小鳥」）あるいは並列関係（例：「太郎及び花子または次郎」）を指摘する機能があるが、今後は、構造化文書の編集や、修辞支援の機能といった革新が期待されている。即ち、技術用日本語プラットフォームが現状の市販ワープロの校正支援機能を超えるには、構文構造、意味構造への取組みが必要であることがわかった。

② 機械翻訳システム

機械翻訳の精度を上げるためには前編集作業を行うことになるが、技術用日本語で書かれた文書そのものは、前編集済みの文となり、技術用日本語の規則がユーザに開示された機械翻訳規則となる。技術用日本語という規則を人間と機械翻訳が共有することで、文の作成基準が明確となると共に、入力文の解析が正確になり飛躍的に機械翻訳精度が高まる事が期待できる。技術用日本語プラットフォームのプロトタイプシステムでは、ヒューマンインタフェースをはじめとして多くの機能を共有して開発コストの削減につなげることができる。

③ 文書検索

文書(コンテンツ)は検索精度に影響を与える特性を持っている。(表 3-1-1)

表 3-1-1 文書検索で扱われる文書とその特性

#	文書種別	構造化度	文書執筆の制約	一文書の長さ	語調	検索目的
01	特許	明細書 タグあり	請求項の構成に 若干の制約あり	長い ばらつきあり	書き 言葉	公知例調査、 技術動向分析
02	論文	論理構造 あり	特になし	長い	書き 言葉	技術動向分析
03	ニュース文・ 新聞記事	５Ｗ１Ｈ で構成	語彙の統一	短い	書き 言葉	トピックス情報 の収集
04	法律文	定型構文	語彙・構文の 統一	長い ばらつきあり	書き 言葉	法律の内容理解 旧法律との差分
05	設計書	記載項目 あり	特になし	長い ばらつきあり	書き 言葉	設計仕様の理解
06	議事録	記載項目 あり	特になし	短い	書き 言葉	結論と根拠理解 発言者の特定
07	マニュアル	なし	誤解を与えない 語彙の使用	長い ばらつきあり	書き 言葉	正しい操作方法 の理解
08	辞典・事典	見出しと 内容	特になし	短い	書き 言葉	言葉の意味の 調査
09	チャット	なし	特になし	短い	話し 言葉	情報収集 マーケティング
10	宣伝・広告	なし	特になし	短い	混在	製品・サービス 情報の収集
11	日記(ブログ)	なし	特になし	短い	混在	マーケティング トレンド分析

12	クレーム・ アンケート	なし	特になし	短い	混在	マーケティング トレンド分析
13	手紙(メール)	なし	特になし	短い	混在	やりとり内容と その時系列変化
14	問合せ （Q＆A）	質問と その回答	特になし	短い	混在	質問に対する 適切な回答取得

検索したい内容を自然言語文で入力して検索する「概念検索（連想検索)」という検索方式も普及してきている。しかし、文字列を活用したレベルでの検索にとどまっているため、検索精度に限界が生じている。技術用日本語は、今まで計算機にとって不得手だった、単語間あるいは文間の構文的・意味的関係を解析して、人間の直感により近い概念のレベルに踏み込んでいく起爆剤的な役割を担っている。技術用日本語によって書かれた文書が蓄積されるようになれば、人間の直感により近い文書検索を実現できる可能性は高い。このように文書検索は技術用日本語の重要な応用システムになるが、逆に、このような検索機能は技術用日本語の規則適用のためのパターンマッチングにも必要な技術である。

④ 辞書サーバ関連プロジェクトと言語資源

これまで言語資源は書籍や記録メディアとして公開されることが多かった（表 3-1-2)。新たな取り組みではシステムやデータをサーバが提供することにより、ユーザが資源の構築に参加していくことなどが期待されている。本技術は、技術用日本語の集合知サーバの機能語辞書等、既存コンテンツとして一部利用できるものがあるが、技術用日本語の規則知識の管理に必要な新機能は新規開発が必要である。

表 3-1-2　言語資源データ一覧表

データ名称	形態	開発者等	発売･公開	価格	規模
青空文庫	電子図書館(デジタルアーカイブ)	数名の有志	1997 年 7 月	無償(登録も不要)	6000 作品 (2007 年 1 月現在)
京都テキストコーパス	タグ付きコーパス (形態素･構文情報)	京都大学コーパス作成プロジェクト	2000 年ごろ	無償(コーパス変換は毎日新聞 1995 年版が必要)	4 万文の新聞記事
毎日新聞	新聞記事	毎日新聞社	1991 年 ～現在	各年 12 万円 本社版のみ	約 11 万件の記事
朝日新聞	新聞記事	朝日新聞社	1984 年 ～現在	各年 18 万円	約 17 万件の記事
読売新聞	新聞記事	読売新聞社	1987 年 ～現在	各年 12 万円 ～27 万円	約 38 万件 (2003 年度版)
THE DAILY YOMIURI	新聞記事	読売新聞社	1990 年 ～現在	各年 11 万円と 17 万円	毎年約 1 万記事

日本語話し言葉コーパス	タグ付きコーパス（音声・言語）	国立国語研究所・NICT・東京工業大学	2005 年 6 月 1 日	研究機関: 個人研究者 5 万 企業 25 万円	音声データ 661 時間分、テキストデータ 752 万語
日本語会話データベース	音声・テキストコーパス	北九州市立大学上村研究室	不明	無償（使用許諾が必要）	120 名分の会話データ
Wikipedia	オンライン百科事典	ウィキメディア財団	2001 年 5 月 20 日（日本語版）	無償	462313 件（2008 年 1 月 31 日現在）
KY コーパス	会話能力テストデータ（第二言語習得）	研究代表者:カッケンブッシュ寛子	1996 年～ 1998 年	無償（使用許諾が必要）	90 人分の会話テープの言語資料
CHILDES（日本版）	幼児言語習得言語データ	JCHAT プロジェクト	1993 年 5 月	無償（使用許諾が必要）	約 23MB（日本語）
日本語動詞の結合価	動詞概念別の結合価データ集	荻野孝野・小林正博・井佐原均	2003 年 12 月	48,000 円 B5 版4 分冊（本文編 1 冊、データ編 3 冊）、ビューアつき CD-ROM1 枚 データ本体は、書籍版所有の上で別途相談可。	11000 概念
日本語語彙体系	シソーラス	NTTコミュニケーション研究所	1997 年 9 月（書籍版）、1999 年 9 月（CD-ROM 版）	80000 円（書籍版）、63000 円（CD-ROM 版）	30 万語の日本語辞書と 14000 件の構文辞書

⑤ 日本語に対する多数の取り組み

日本語に関しては、日本語の規範・規格に係る国語審議会における当用漢字、常用漢字、新常用漢字（答申予定）がある。技術用日本語で漢字をどう取り込むかは議論の対象だが、読める漢字を基準にするという考え方もある。

工学的な観点から、「醍醐プロジェクト」（円滑な情報伝達を支援する言語規格と言語変

換技術、2003～2005 年）の研究がある。「日本語の規格化とは、日本語のサブセットを設定すること。」として、目的は、3 つあり、正確に情報が伝達できること（＝Readability)、情報を十分伝達できる規格であること、及び規格を満たしているかの判定が機械にできること(テキストの（平易度）診断)と定義されている。技術用日本語が必要とする規格は、構文構造以上のものが多く、語彙レベルを除き、直接的な流用はない。

また、外国人に対する日本語教育の分野において、日本語能力試験で制限日本語が有効に利用されており、日本語に対する制限が有効に機能している例をみることができる。日本語教育においては、言語学の観点ではなく、言語教育の観点から文法等の体系化が行われている。この日本語教育からの文法という考え方も、技術用日本語設計のために、おおいに参考にすべきであることがわかった。

⑥ 法令日本語と数学日本語

厳密な情報表現のツールとして日本語を使ってきた代表的な分野に、法曹の分野と数学の分野があり、特に法曹の分野では、すべての公式文書を原則、日本語だけで表現し、法令日本語という曖昧さを徹底的に排除する日本語の用法が確立されている。

他方、数学日本語は、数学教育の観点から、定義、公理、定理、そして、証明を論理的に明晰に表現する、すなわち、分かりやすく誤りの無いように表現するための日本語である。

これらは、技術用日本語の規則を策定し、技術用日本語を評価するときに重要なコンテンツと成りえる。

（2）技術文書を代表する特許文書に関する技術動向

① 特許文書の構造と特性

特許文書は、その技術内容がすべての産業技術分野に広く行きわたり、かつ国際的に通用する分類に基づき整理され、その利用性の向上が図られている。このように特許文書はきわめて体系的に整理され、その利用が多岐にわたる技術文書であり、技術情報を開示するという目的に叶うように、個々の文献において技術説明が完結しており、広く産業技術文書を代表する。この点から、特許文書は、広範な産業技術分野に対する技術用日本語活用の代表となり得る。

特に、知的財産推進計画２００６及び２００７では、外国での特許権獲得を促進するために必要な、特許出願明細書の文章の平易化・明瞭化が謳われ、特許庁においては、翻訳作業や権利行使を踏まえた、望ましい明細書の事例を作成する「望ましい明細書の作成に関する調査研究」が実施されている。

なお、技術内容を公開するための文献としての利用する場合については、出願済みの特許請求の範囲の記載を技術用日本語に書き換えることによって、機械翻訳の精度向上や、検索対象としての効率向上に結びつけることも可能である。しかし、権利範囲を示す記載

として重要な役割を果たす「特許請求の範囲」については、その位置づけを正しく認識し、その取扱いについては細心の注意を払う必要があることがわかった。

② 特許文書における機械翻訳

特許文書は通常の技術文書と異なった特徴をもった文がある。そのため、その機械翻訳にはその特徴を活かした方法が有効であることがわかった。例えば、特許請求項は慣習的に１項１文であり、構成要素をすべて含んだ上に名詞で終わる体言止めで書かれることが多い。長文に対しては、短文化することが有効である。英語での標準的な特許文書の表現を再現しやすくするために、前処理による書き換えを行うことで、機械翻訳の精度を高めることができる。

特に、特許庁では、機械翻訳の精度を向上させることによって、海外特許庁の審査官が、わが国の審査結果等の情報を利用しやすくするため、拒絶理由通知書について分析を行い、定型化可能な表現パターンの抽出等を行う「特許審査経過情報の英語による海外特許庁への発信に係る調査」を実施している。技術用日本語プラットフォームでは、こういったユーザサイドのニーズを取り入れる必要がある。

③ 特許文書における分類と検索

検索精度を向上させるためには、クエリーから内容を特徴付けるキーワードをいかに正しく特定するか、及び、特定したキーワードを検索対象文献中のキーワードといかに正しく照合させるか、の２点が技術課題となる。キーワードを正しく特定し、正しく照合させるためには、そのキーワードが文献の中でどのような位置付け（役割）で使われているかを、語彙的、構文的、文脈的な観点から把握することが必要である。技術用日本語では、文章を構成する語彙や構文をある程度統一し、発明の構造や内容がより明確になることにより、そのキーワードの文献内における位置付け（役割）をより正確に解析できるようになる。また、キーワード単位の解析・照合から一歩深く踏み込んで、概念レベルでの発明内容の特定・照合といった人間に近い検索方式の実現により、検索精度の一層の向上が期待できることがわかった。

④ 特許工学及び特許請求項を対象とした言語処理

特許工学とは、発明の創造、保護、活用からなる、いわゆる知的創造サイクルに係る活動を定型化するものであるが、そこで扱われる特許工学支援ツール（ＣＡＰＥ(Computer Aided Patent Engineering)ツール）のうち、例えば、特許明細書作成の業務の効率を大幅に向上させるツールは見当たらない。特許明細書をさらに構造化、標準化できれば、作業効率を大幅に上げる特許明細書作成支援ツールの登場の可能性があることがわかった。
また、分析評価系の特許明細書分析ツールや特許価値評価ツールを実現するためには、特許明細書の言語処理技術が必要となる。特許明細書のうち、特に特許請求項については、

可読性向上や検索支援の観点での言語処理研究が進められてきた。技術用日本語によって特許請求項の可読性が向上することで、一般の企業人や研究者にとって、特許がより身近なものになることが期待でき、それによって新しい特許出願が促進されることにつながると考えられる。

⑤ 特許日本語ライティング

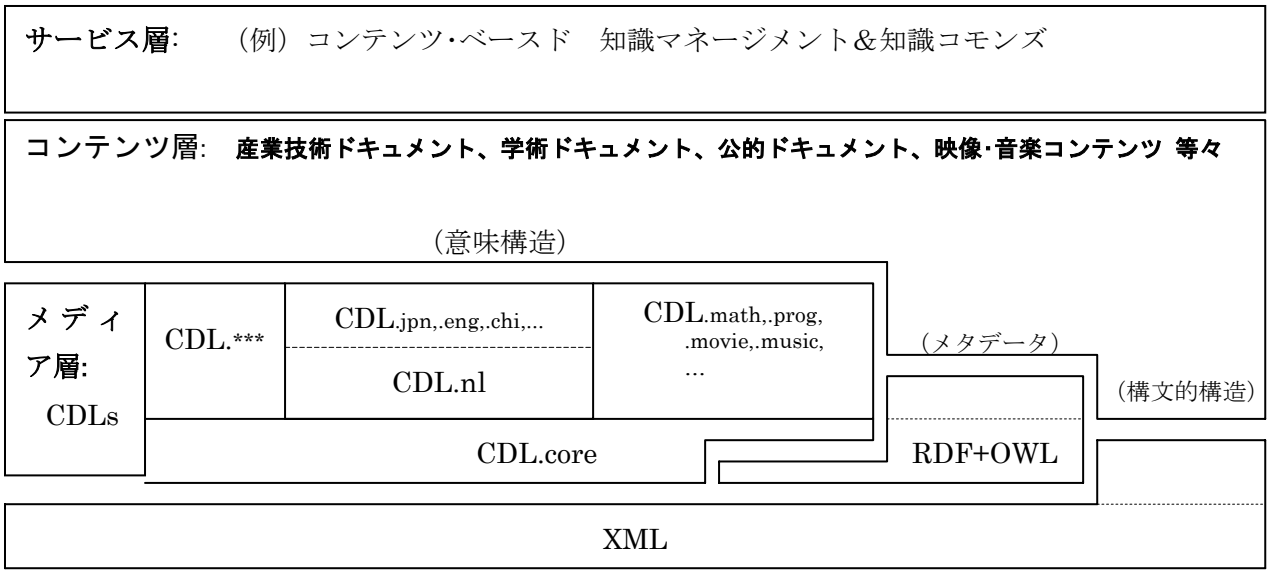
外国に出願をするにあたっては、日本語からその現地語への翻訳は必須であり、その経済的かつ人的負担は非常に大きく、かなり以前から、機械翻訳をより有効に機能させるための取り組みがなされてきたことがわかった。特に、昭和 61 年度にまとめられた特許抄録作成のガイドラインに関しては、その報告のなかで、「特許抄録作成ガイドラインは自動翻訳のために設定されたものではない。しかし、人間にわかりやすい文章を作成することは機械処理にも影響を持つことが予想される。」ということで、人間にも読みやすい文章を書くことは機械にもわかりやすい文章になることが説かれている。

また、ガイドラインに従って書き換えを行った方が、翻訳に関し、品質、時間ともに明らかに前処理や後処理を行うよりも効果的であることが報告されている。技術用日本語プラットフォームはこのガイドラインを言い換え規則の中に規則化することが必要であると考えている。

（3）文書処理高度化のためのＣＤＬ（概念記述言語）に関する技術動向

ＣＤＬは、技術用日本語プラットフォームの基幹をなす技術で、様々な表現メディアや多様なコンテンツ、これらの意味の概念化を共通的な形式で記述するための概念記述言語（CDL: Concept Description Language）である。

図 3-1-1　CDL の言語階層



ＣＤＬは、セマンティックコンピューティングの研究開発を推進するための共通言語であり、セマンティックコンピューティングが目指す、知的システム・知的環境の設計原理をまとめた。

技術用日本語におけるＣＤＬの役割は、次の４点である。

ａ．技術用日本語の文書の意味内容をコンピュータが理解できるよう概念記述に使用し、高度な文書処理を実現する。
ｂ．人に対する明晰性とコンピュータに対する明晰性とを両立させるための仲介となる。
ｃ．「コンピュータが CDL へ（ほぼ）自動変換できる日本語」を技術用日本語と定める。
ｄ．非明晰文書から明晰文書への変換支援を行う明晰ワープロにおいて、コンピュータにより文書構造を記述する。

（４）文書制作高度化のためのセマンティックオーサリングに関する技術動向

技術用日本語も一般の日本語も、意味的にも構文的にも構造化された言語的コンテンツである。構造化とは、単語や句が修飾したり参照することで形成されるもので、これはグラフとして可視化できる。言語現象を正確に表現するために文法に従って記述される。両者の違いは、文法の違いになる。技術用日本語は、一般日本語文法よりも制約の数が多い言語である。

セマンティックオーサリング(semantic authoring) では、既存の文章を事後的に構造化するというより、典型的には文章を書く代わりにこのようなグラフを作る。その対象は１文というよりも１文章であり、1 段落である。文章中の意味的、論理的なまとまりが相互に結ばれることに基づいてコンテンツが作成される。まとまりは、文章、段落、文、句など自由に選択できる。結ばれる関係名の集合はあらかじめ定められていて、その名前を明記することで文意が明確になる。文間の制約が支援されること、及びグラフ表示によって作成中のコンテンツが理解しやすくかつ編集しやすくなることで、従来の文章作成よりも質が高いコンテンツを簡単に作ることができることを示した。

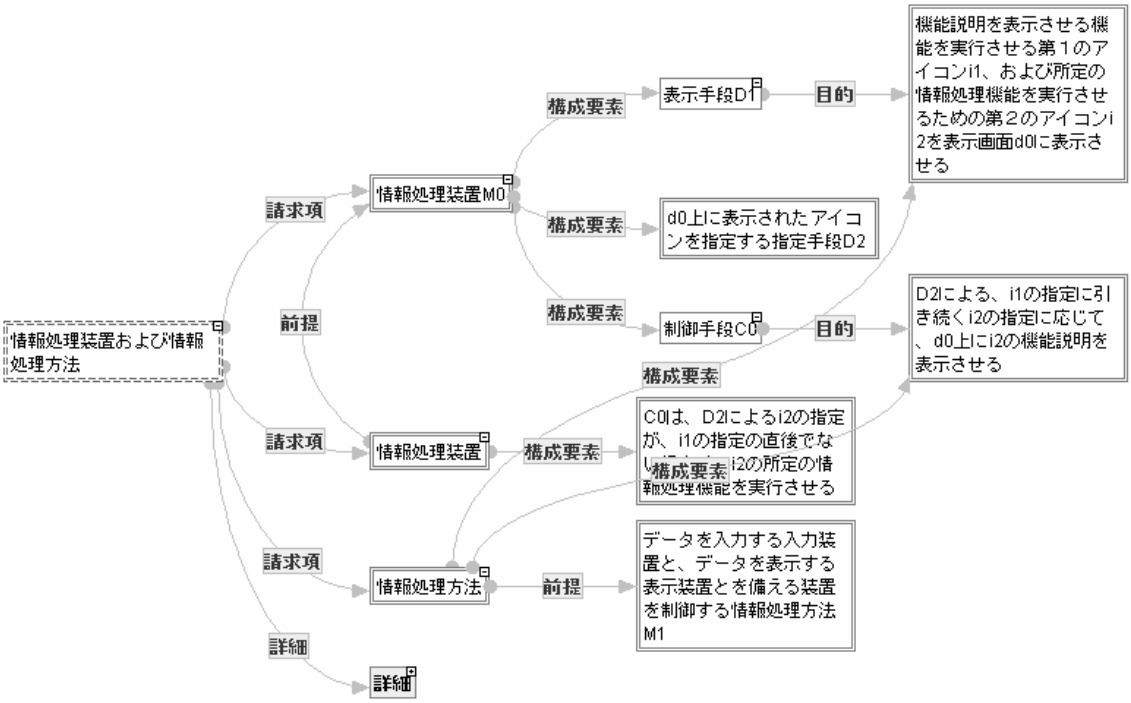


図 3-1-2　オントロジーによる特許明細書の構造化

３－１－２　海外技術調査

技術用日本語が手本とすべき新技術が海外にあるかどうかを調査するため、文献及び訪問調査を実施した。調査対象は大きく分けて二つで、一つは制限英語と呼ばれるテキスト形式のもの、もう一つは、知識を表現するために、グラフィカルな表現形式を用いる方式である。

（1）制限英語

英語を語彙及び文法で制限して、正確で分かりやすい英語を流通させようとする制限英語の用途は大きく分けると二つになり、一つは、人間用の技術英語であり、もう一つは機械用の Controlled English と呼ばれるものである。前者はテクニカルライティングで、後者は機械翻訳や知識の獲得等への応用を目指している。前者では、Simplified Technical English のように既に商用サービスが提供されている。これらのシステムは制限語彙及びライティング規則からなっており、構文や意味に関する自然言語処理技術を前提としておらず、技術用日本語が目指す目標にとっては、不十分であることが分かった。

他方、Controlled English は、コンピュータで制限英語を構造解析し、その解析結果から知識（オントロジー）を獲得する応用や、多言語機械翻訳へ応用する研究がある。知識体系（オントロジー）記述の研究は、技術テキストの意味解析を行うボーイング社の CPL(Computer Processable Language)、車のアセンブリーラインの指示書を記述するためのフォード社の Standard Language、EU プロジェクト関連で 2 件：SEKT (Semantically

Enabled Knowledge Technologies)及びACE (Attempto Controlled English)を調査した。CPLを除き、それらの英語は制限が厳しく、余りに単純化されたものであり、自然な英語とはいえない。また、制限を緩めると自然言語特有の曖昧性を生じさせることになる。多言語機械翻訳では、米国CMU(カーネギーメロン大学)のKANT(KANT(Knowledge-based Accurate Natural-language Translation) を調査した。システム側の制限英語が厳しいとユーザが拒否反応を示すとの報告がある。

（2）グラフィック形式

コンセプトマップを利用したオントロジー記述の研究を調査するためにカタルーニャ工科大学及び高等科学研究院を訪問した。コンセプトマップは、意味ネットワークの一種で、リンクにラベルが付けられる。オントロジー記述言語の標準であるOWL (Web Ontology Language) への相互変換ツールを開発している。セマンティックオーサリング技術との比較では、コンセプトマップはラベルが標準化されていなくて、自由に記述できるため、変換時に曖昧性が生じる問題がある。また同大学は、2008年2月からEUのFP7の一環としてALIVEプロジェクトに参加する。グラフィカルな形式を組織科学に適用して社会サービスへ応用する研究がある。

グラフィカルな形式には、会議などの議事録、発想記述で最近盛んに使われているマインドマップがある。中心となるテーマから放射状に課題を広げる方法はシンプルでわかりやすい。オントロジーとの連携が期待される。マインドマップは、ある単一のテーマに対して自分が何を考えているかを反映しているが、コンセプトマップには、マップとしてもシステムを俯瞰する図としても使える柔軟性がある。

以上の調査から、テキスト形式の文章作成には新しい知見を見つけることはなかった。グラフィカルな情報の利用がライティング技術における効率性を高める先進性を有していることが確認できた。技術用日本語の基盤システムはグラフィカルな表現であり、かつテキスト形式も可能なため、両方の利点を活用することで今後の発展が期待できることが分かった。

（3）英語教育政策

機械翻訳で技術用日本語から正確な英語が出力されたとしても、カナ漢字変換結果をユーザが選択するのと同じように最終確認は人間が行うことになる。そのため技術用日本語プラットフォームの使用者の英語に対するリテラシーを高めることが望ましい。英語の能力を高めるために各国が採っている英語教育政策を調査した。欧州では2001年新教育プログラムで母国語プラス欧州言語2語の1+2が提唱され、アジアでは英語教育が小学校で必修化している。我が国は遅れていたが平成20年1月に中教審から答申が出され、平成21年度から小学校での必修に向けた移行期に入り、平成23年度から必修になる。

３－２　開発課題の考察

（1）技術用日本語プラットフォームの課題設定の先進性と妥当性

客観的に事実を表現する技術的な文書において、日本語を明晰に使用するための工夫や提言は、無数といわれるほどに行われてきた。しかしながら、いまだ、確実な技術に裏付けられた方式とシステムには行き着いてはいない。このいまだ無かった方式とシステムに達するというのが技術用日本語プラットフォーム開発の目標である。なぜ、現在、そのような目標設定が可能であるのか、以下にその技術的な観点からの要点をあげる。

知見や経験の十分な蓄積

言語学における知見の蓄積、言語処理における要素技術の蓄積、日本語文章論や日本語ライティングにおける様々な経験の蓄積、特許文書などをはじめとする多様な文書分野におけるノウハウの蓄積など、技術用日本語に向けての十分な蓄積が得られている。しかし、今のところこれらの蓄積は個別でばらばらの試みの累積である。技術用日本語によって、これらの蓄積を体系的な方式に、手順だった作業の方式にまとめ上げる。いずれにしても、技術用日本語に関する十分な蓄積がなされていること、すなわち、技術開発に進めるための日本語に関する成熟した状況が存在すること、これが第一義の要点である。

適切なシステム実現技術

技術用日本語のオーサリングシステムの実現には、非明晰な日本語から明晰な日本語への機械翻訳技術が核となるシステム実現技術となる。日本におけるここ 20 年ほどの機械翻訳システムの研究開発、製品化、利活用などにおける努力は、十分な実現技術を提供してくれるまでに進展した。また、広く利用されている日本語ワープロや校正･推敲支援システムは、システム実現への第一歩と位置づけることができる。そして、システム実現には、相当規模の相当精度の規則や辞書の開発が不可欠である。各種電子化辞書や言語資源の利用環境が整備され、更に、Web 環境における集合知の仕組が整備され、加えて、統計的な学習方式などの知的な開発方式も実用化され、必要な規模や精度を容易に達することができる環境が整った。

広範な応用への連携技術

技術用日本語プラットフォームは、多様な技術用日本語仕様に対応できなければならない。技術用日本語プラットフォームは、翻訳・検索・要約等々の多様な文書処理、多様な方式の知識処理や推論処理、映像･画像･音像や図解・数式・プログラム言語などとの多様なマルチメディア処理、多彩な産業技術ドキュメントの多様な書式処理、これらの処理と効果的で効率的な連携処理が行えることが重要である。このような連携処理こそが、技術用日本語の利用価値を格段に高めることになる。この連携処理を実現するのが CDL（Concept Description Language：

概念記述言語）の役割である。CDL は、[課題設定の先進性と妥当性]「知見や経験の十分な蓄積」における日本語の蓄積を[課題設定の先進性と妥当性]「適切なシステム実現技術」における実現技術に結びつけるためにも非常に効果的な役割を果たす。すなわち、日本語の概念構造を表層レベルから深層レベルにいたって連続的に記述し、適切な処理に結びつける役割を果たす。

技術用言語に関しては、英語は、日本語に比べかなり先行している。日本語がおかれた社会的、言語的、技術的な状況が英語とのキャップを大きくした。それでは、技術用日本語は、技術用英語（制限英語）の現状に追い付くということが目標なのであろうか。否である。追い付くのは当然として、更にその先へと追い越すのが目標である。すでに、10 件以上の制限英語が実用化され、各方面で利用されている。これらの制限英語は、それぞれに評価すべき点を持ち合わせてはいるが、先行しただけに程ほどの技術を使った程ほどの機能のものばかりである。例えば、制限英語のライティング環境の実現は、形態素解析技術と簡単な構文解析技術の利用にとどまるレベルのものである。一方、技術用日本語のオーサリング環境の実現には、機械翻訳をはじめとした最新の文書処理技術、CDL という最新の概念表現・処理技術、Web 集合知という最新の協調技術が動員されることになる。これは、丁度、英文タイプライタと日本語ワープロとの経緯を思い起こさせるものである。

（2）開発課題とフィージビリティスタディの概要

技術用日本語プラットフォームの開発課題は、大きく、技術用日本語共通基盤仕様、技術用日本語プラットフォームシステム、技術用日本語アプリケーションシステムの 3 つからなる。技術用日本語共通基盤仕様は、多様な技術用日本語仕様に対して明晰性を体系的に統一的に定義するための共通の枠組みと仕様を記述するための共通の形式とを仕様としてまとめたものである。すなわち、技術用日本語に対するメタ仕様、あるいは、プラットフォーム仕様である。

技術用日本語プラットフォームシステムは、技術用日本語オーサリングシステムと技術用日本語言語知識集合知サーバからなる。技術用日本語オーサリングシステムは、明晰な日本語文書の作成をインタラクティブに支援するシステムである。規則ベースや辞書ベースの内容を入れ換えることによって、様々な技術用日本語仕様に対応することができる。集合知によって、この規則ベースや辞書ベースの内容を漸進的に整備していく共通環境が技術用日本語言語知識集合知サーバである。

技術用日本語アプリケーションシステムにおいては、文書処理の代表として翻訳と検索を取り上げる。すなわち、技術用日本語日英機械翻訳システムと技術用日本語文書検索システムである。技術用日本語によって、機械翻訳や文書検索の高度化が効果的に達成されることを実証する課題である。

本フィージビリティスタディにおいては、技術用日本語共通基盤仕様に基づいて実験用

暫定規則をまとめ実験を行い、技術用日本語プラットフォームシステムと技術用日本語アプリケーションシステムに関しては、基本検討を行った。

３－２－１　技術用日本語共通基盤仕様（第０版）

技術用日本語共通基盤仕様（第 0 版）は、多様な技術用日本語仕様に対して明晰性を体系的に統一的に定義するための共通の枠組みと仕様を記述するための共通の形式とを仕様としてまとめたものである。すなわち、技術用日本語に対するメタ仕様、あるいは、プラットフォーム仕様である。本仕様は、（財）日本特許情報機構特許情報研究所の特許版･明晰日本語策定委員会において策定された明晰日本語基本仕様（第 0 版）に準拠する。したがって、詳細については、下記の報告書を参照されたい。

特許版・明晰日本語策定委員会：『明晰日本語』、特許版・明晰日本語策定委員会報告書、（財）日本特許情報機構　特許情報研究所、2008 年 3 月

この明晰日本語基本仕様（第 0 版）に基づいて実験用暫定規則をまとめ、フィージビリティスタディを行った。

（1）設計方針

① 明晰な日本語であること

技術用日本語は、明晰な産業技術ドキュメントを作成するための日本語である。産業技術ドキュメントとは、事柄を客観的に論理的に記述するドキュメントであり、事柄を情緒的・感覚的・主観的に記述する部分はない、あるいは、あってもごく限定された部分となるドキュメントである。ドキュメントが明晰であるとは、ドキュメントが対象とする読み手が内容を正確に容易に読み取ることができるということである。また、技術用日本語は、ドキュメントの書き手にとっても明晰である。すなわち、書きたい内容を容易に適切に表現できる日本語である。

② 様々な目的に対応できる多様な日本語仕様を包含できること

技術用日本語は、それぞれの目的に沿って定められる様々な日本語仕様を包含する。一つの日本語仕様によって、技術用日本語が定まるのではない。技術用日本語の共通基盤仕様とは、すべての技術用日本語仕様に対応できる、あるいは、すべての技術用日本語仕様を生成できる基盤となる仕様である。

③ 人にもコンピュータにも明晰であること

技術用日本語は、人にとってもコンピュータにとっても明晰である。コンピュータにとって明晰であるとは、コンピュータが処理・理解できる表現形式に自動的に変換できる、あるいは、変換を効果的に支援する環境を用意することができるということである。表現形式は、処理システムごとに異なる。そのためにも、技術用日本語には、

仕様の多様さが必要となる。人に対する明晰性とコンピュータに対する明晰性は、必ず両立するというわけではない。言語の明晰性は、共通の常識を前提にして定められる。人が共通に持つ常識の能力は、コンピュータが持つ能力に比べ格段のものである。コンピュータにとって明晰となるために、技術用日本語には、人には不必要な詳細な表層表現が求められる。不必要な詳細さは、人にとっての明晰さを妨げることになる。このような明晰性のトレードオフを解消するのが CDL の役割である。すなわち、コンピュータにとって明晰な技術用日本語テキストは、CDL による適切な内部表現テキストに変換される。そして、その CDL テキストを人にとって明晰な技術用日本語テキストに変換する。CDL が二つの明晰性を仲介するということである。

④ 他の表現メディアとの明晰な連携が行えること

産業技術ドキュメントは、日本語だけで表現されているわけではない。図・表・図解・写真・数式・化学式・仕様記述言語、そして、書式、実に多彩な表現メディアが用いられている。技術用日本語は、これらの多彩な表現メディアとの連携表現にも明晰性を保証する。

⑤ 支援環境と一体となること

技術用日本語を適切に使用し明晰な文書を作成するための支援環境を前提とした共通基盤仕様である。支援環境の設計と一体となる共通基盤仕様である。技術用日本語プラットフォームシステムがその支援環境である。

⑥ 漸進的に精度向上が図れること

個々の技術用日本語仕様に対しても、最初から完成された仕様を求めることはできないし、また、求める必要もない。あるレベルに達した仕様から始め、実際に使用･運用しながら精度を漸進的に高めていくことになる。そのためにも、共通基盤仕様という設定が必要になる。技術用日本語言語知識集合知サーバを用いて斬新的な精度向上を支援する環境を構築する。

（2）表現機構モデル

日本語文章論や日本語ライティングの知見を整理すると、技術用日本語に求められる明晰性は、以下のようになる。

[求められる明晰性]

① 語彙の意味が共通の了解の範囲内であり、多義語に関しては、どの語義に対応するのかが明確に判断できる。

② 語彙間の共起関係が、標準的な日本語の用法に沿う。

③ 新しい概念を表現する新規の複合語に関しては、辞書登録の簡便な仕組が設けられる。臨時複合語に関しては、構成要素間の関係が明解に読み取れる、すなわち、文への簡明な言い換えが可能である。

④ 名詞句（「X1 の X2 の・・・の Xn」）は、構成要素間の関係が明解に読み取れる、

すなわち、文への簡明な言い換えが可能である。

⑤ 文（節）は、構成成分（最上位文節）の範囲と役割が明確である。構成成分の配列や読点の用法は、標準的な規則にしたがう。係り受け関係の入れ子構造の複雑度は、閾値以下である。係り受け関係は、非交差条件を満たす。

⑥ 論理的関係（並列句や並列節）においては、構成要素、及び、スコープなどが明確である。

⑦ 否定は、作用するスコープや否定によって表現される概念範囲が明確である。

⑧ 名詞は、概念の新規導入なのか照応詞として機能するのかが明確に判断できる。照応詞として機能する場合は、対応する先行詞を簡明に同定できる。

⑨ 文章、あるいは、文内の情報構造が明解である。

⑩ 文章、あるいは、文内の談話構造が明解である。

求められる明晰性を技術用日本語仕様へと具体化し、支援環境の設計･実装に結びつけるために、共通基盤仕様として日本語の表現機構モデルを設定する。このモデルは、日本語がどのようにして情報を表現し伝達するのかの仕組をモデル化したものである。ただし、実際の日本語の仕組をそのまま反映するものではない。実際の仕組をある程度単純化し正規化したものである。しかしながら、技術用日本語が目的とする情報の表現・伝達機能としては、この表現機構モデルで十分である。

表現機構モデルは、2 段階のモデルからなる。第 1 段階のモデルは、情報を言語としてどう捉えるのかをモデル化するグラフモデルである。第 2 段階のモデルは、捉えた情報をどのように表出するのかをモデル化する線状化モデルである。線状化モデルは、言語の外部表現である線状構造と内部表現であるグラフ構造との対応付けをモデル化する。明晰性と負荷の少ない対応付けとは一致するとみなす。

グラフモデルは、日本語国文法の詞と辞という考え方を一般化し、情報の構造化を入れ子構造によって一般化したモデルである。このモデルによって、構造表現として、文章、段、文、述語成分、格成分、述語と基本成分の共起、複合語をモデル化し、構造にまたがる表現として、照応表現、情報構造、論理的表現（並立表現と否定表現）をモデル化し、要素表現として、複合語と複合辞、多義語、難解語をモデル化する。

グラフモデルは、CDL の土台であるハイパーE-R モデルに直接的に対応する。グラフモデルは、日本語の表層レベル（構文レベル）の概念記述を行う CDL.jpn-surface のグラフ表記そのものである。

（3）仕様の記述形式

① 言い換え規則

技術用日本語の仕様をどのような形式にまとめあげるのか、仕様の枠組みをどのようなものにするのかである。仕様の基本形式としては、次の二つが考えられる。

ａ．明晰文書を作成するために守るべき書き方（ライティング）規則の形式にまとめる。
ｂ．非明晰文書を明晰文書へと書換えていく言い換え（パラフレージング）規則の形式にまとめる。

従来、この種の仕様は、すべてａの形式にまとめられている。各種文書に対する作成規則やガイドライン、日本語テクニカルライティングにおける規則、制限英語における規則など、すべてがこの形式である。

技術用日本語については、ｂの形式、すなわち、言い換え規則の形式にまとめる。以下にあげるのが、その理由である。

自然な作成工程：

産業技術ドキュメントの作成者は、一般的な日本語という枠組みやそれぞれの分野特有の日本語用法の枠組み、これらの枠組みの中で思考しつつ文書を作成する。これらの枠組みに馴染まないような規則にしたがって文書作成することを強いても、実際に効果的な作成作業には結びつかない。まず、これらの枠組みに沿ってすなおに文書を作成し、その文書の部分々々を書き換え、求められる明晰性を達成していくという手順の方が、本来の自然な作成工程である。

適用できる規則：

遵守規則としてまとめられた書き方規則の多くは、実際の適用に際してあまりにも幅がありすぎるものとなる。したがって、適用の判断が大きくふれ、実際には努力目標程度に扱われることになる。一方、言い換え規則においては、どう適用するのか、適用した結果がどうなるのか、すべてが具体的で明解である。日本語による言い換えであるから、適用結果が適切であるかどうかを誰もが一様な基準で判断することができる。ただし、どの規則をどのような順番で適用するかは、作成者の選択に委ねられる。

支援環境の存在：

オーサリングシステムでは、基本的な言い換えは、確度の順にしたがってシステムが行ってくれる、あるいは、システムが提示した言い換えの候補から選択する、このような支援機能によって、効率よく明晰な文書が作成できるようになる。言い換えは、人とシステムがインタラクションする最適なコミュニケーション手段である。

既存文書への対応：

従来の手法にしたがって文書を作成し、それを書き換えて明晰性を達成するという作成工程は、蓄積されている既存の産業技術ドキュメントを明晰化する作業にも適用できる。

書き方規則としての利用：

言い換え規則は、書き方規則として用いることもできる。オーサリングシステムの一文ごとの入力モードでは、言い換え規則が書き方規則として機能する。

なお、言い換え規則の設計には、言語学的な厳密さは求めない。対象とする産業技術ドキュメントの特性に基づく許容範囲に納まる厳密さで良いとする。

② 言い換え規則と言い換え辞書

技術用日本語仕様は、大きく言い換え規則に関する仕様と言い換え辞書に関する仕様に分かれるとする。すなわち、

言い換え規則:(2)で述べた[求められる明晰性]における④から⑩に対応するための言い換え規則である。文章、文、節、句に対する構造的な言い換えの規則である。この規則は、技術用日本語の分野の違いによる影響が少なく、すべての技術用日本語に適用できる仕様を多く含む。

言い換え辞書:(2)で述べた[求められる明晰性]における①から③に対応するための言い換え辞書である。辞書は大きく一般語彙辞書と専門語彙辞書に分かれる。専門語彙辞書は、ほとんどが名詞であり、技術用日本語の分野の違いを最も反映する。複合語に対しては、句や文への言い換え情報が付加される。また、特殊な専門語彙には、理解しやすい表現への言い換え情報が付加される。一般語彙辞書の内容は、大きく詞-部と辞-部に分かれる。詞と辞の区分は、言語学における区分と異なる。辞には、通常の助詞、接続詞、空辞、そして、複合辞が含まれる。多義語については、語義を区別できる簡明な言い換え情報が付加される。これは、辞-部に対しても同様である。また、動詞の結合価パターンなどの標準的な共起情報も言い換え辞書に付加される。

③ 言い換え規則の記述形式

言い換え規則の記述形式を以下のように定める。この記述形式は、人間用であり、そのままコンピュータに実装できる形式ではない。最終的には、CDL による言い換え規則の記述形式を定め、コンピュータへの実装にも対応できるようにする。

―――――――

<規則識別子><タイトル>:<文章記述による説明><明晰化効果>

<形式的な記述>

<例>

<説明>

―――――――

<規則識別子>:=(<カテゴリ番号>－<規則番号>)

言い換え規則を識別するための番号記号である。<カテゴリ番号>は規則のカテゴリを識別する番号で、<規則番号>は、カテゴリ内で規則を識別する一連番号である。仮の規則識別子として、‘(#-#)’を用いる。

<タイトル>

人が規則を判別するためのタイトルである。

<文章記述による説明>

規則を技術用日本語（規則記述用）による文章で説明したものである。

<明晰化効果>：＝[<分野>:<効果>. . .]

言い換え規則の適用が、明晰化の効果をもたらすかどうかの目安を表す。<分野>は、どの分野向きの技術用日本語であるかを表す。<効果>は、明晰化の効果の度合いを示す整数値（5～1）である。

<形式的な記述>：＝<言い換え前パターン>→<言い換え後パターン>　[<条件>]

ある程度形式的な言い換え規則の記述である。ただし、形式化はある程度であり、厳密な意味での形式化ではない。<条件>は、パターン内のパラメータなどに関する条件の記述である。

<例>：＝（例）<例示>

言い換えの例である。

<説明>：＝（注）<注釈>

他の規則との関連や適応に当たっての注意事項のなどの説明である。

④ 言い換えのイメージ

（原文）　：翻訳モデルＰ（Ｊ｜Ｅ）を実現するにあたって、フランス語と英語、及びドイツ語と英語など、互いに近い関係にある言語間の翻訳では、語アライメント方式と呼ばれる統計的翻訳がよい成績を収めてきた。

↓

（原文３分割）：翻訳モデルＰ（Ｊ｜Ｅ）を実現する。その実現にあたって、言語間の翻訳では、語アライメント方式と呼ばれる統計的翻訳がよい成績を収めてきた。**その言語は、**フランス語と英語、及びドイツ語と英語など、互いに近い関係にある。

↓

（再結合）：翻訳モデルＰ（Ｊ｜Ｅ）を実現するにあたって、フランス語と英語、及びドイツ語と英語など**のように**、互いに近い関係にある二言語間の翻訳では、語アライメント方式と呼ばれる統計的翻訳がよい成績を収めてきた。

３－２－２　技術用日本語プラットフォームシステム

（1）技術用日本語オーサリングシステム

① システム機能

技術用日本語によって明晰な日本語テキストをオーサリングするプロセスを効果的に支援する。オーサリングのスタイルには、以下の３つが設けられる。

ａ．非明晰なテキストを明晰なテキストへと変換（推敲・校正）するプロセスをインタラクティブに支援する。これは、蓄積されている既存テキストを明晰化するというオーサリングスタイルである。

ｂ．逐次的に入力される一文ごとに明晰性をチェックし、非明晰な箇所を指摘し明晰化への言い換えを支援する。新しく明晰なテキストを作成するときのオーサリングスタイルである。

ｃ．基本的な明晰性を強要するグラフ形式での入力を支援するセマンティックオーサリングに基づくオーサリングスタイルである。グラフ形式からテキスト形式への変換機能を利用すれば、明晰な通常の日本語テキストを得ることができる。

オーサリングシステムは、それぞれの目的に応じて様々な技術用日本語仕様に対応することができる。仕様は、言い換え規則ベースと言い換え辞書ベースの内容によって定められる。規則ベースや辞書ベースの内容を入れ換えたり、一部を改変したり、追加したりするための簡便な手立てが用意される。

② システム構成

オーサリングシステムは、非明晰な日本語テキストを明晰な日本語テキストに変換する日本語-日本語間の機械翻訳システムとみなすことができる。機械翻訳方式としては、できるだけ表層に近いトランスファ方式が有利である。日本語同士の間での翻訳であるため、概念や意味のレベルに渡る内部表現形式の利用は、適切ではない。

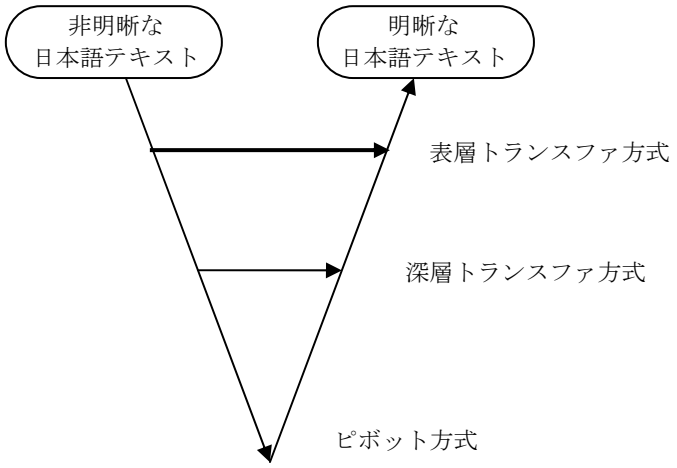


図 3-2-1　非明晰 - 明晰トランスファ

ただし、機械翻訳として次のような新しい機能の実現が必要である。

ａ．構造の変換：一文を複数の文に変換したり（言い換えたり）、逆に、複数の文を

一文に変換したりする（言い換えたりする)。複合語、句、節、文、これらの間の相互変換（相互言い換え）を行う。

b．インタラクティブな翻訳：言い換え規則適用の適切さの判断や他の規則選択などは、ユーザがインタラクティブに介入し、明晰化をという翻訳作業を進める。

上記のような新しい機能を実現するためには、オーサリングシステム内部における日本語テキストの適切な表現形式が重要となる。この内部表現形式を記述するコンピュータ用言語として CDL（Concept Description Language）の構文概念レベル CDL.jpn-surface を用いる。

技術用日本語オーサリングシステムのシステム構成図を以下に示す。システム機能で説明した 3 つのオーサリングスタイルにしたがって、ユーザのディスプレイ画面上の非明晰テキストや明晰テキストの表示形式は、異なってくる。いずれのスタイルにしても入力された日本語テキストは、日本語→CDL.jpn-surface 変換によって CDL.jpn-surface テキストに変換される。この CDL.jpn-surface テキストに対して、言い換えエンジンは、言い換え規則ベースと言い換え辞書ベースから適応可能な規則を選び出し明晰化に向けた言い換えを行う。言い換えられた CDL.jpn-surface テキストは、CDL.jpn-surface→日本語変換によってユーザフレンドリーな表示形式に変換されユーザに提示される。ユーザは提示内容から言い換えの適切さを判断する。不適切な場合は、他の言い換え規則の選択を言い換えエンジンに指示する。このインタラクションの仕組は、日本語ワープロのインタラクションの仕組に準ずることになる。原則として、個々のユーザが規則ベースや辞書ベースを細かく管理することはしない。ユーザは、最初にどの技術用日本語仕様を用いるのかを決めればよい。仕様を決めることは、どの集合知サーバにリンクするかを決めることである。

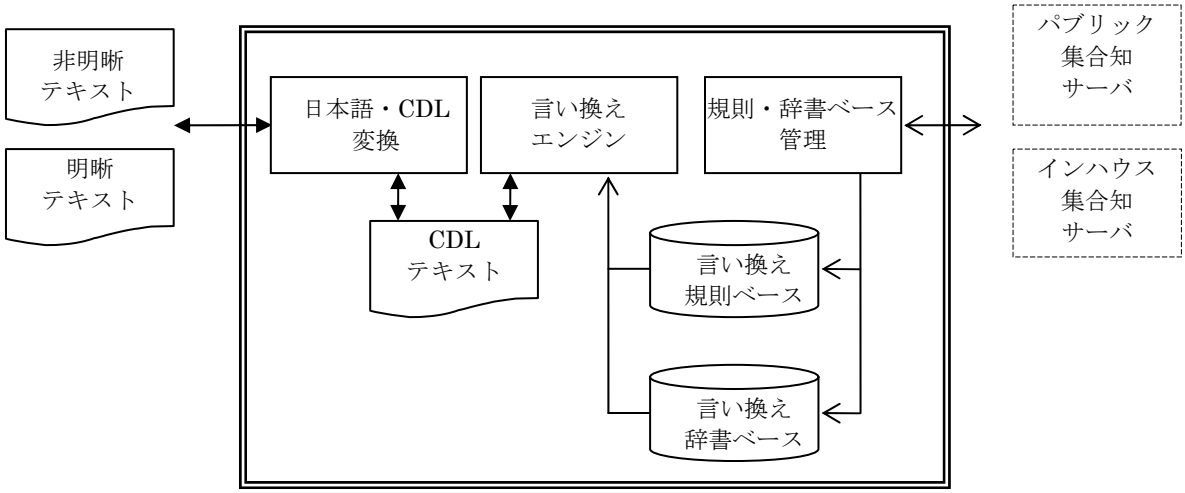


図 3-2-2　技術用日本語オーサリングシステムの構成

（2）技術用日本語言語知識集合知サーバ

Web 上で多くのユーザが参加して情報や知識を出し合うなどの共同作業を行い、知識体系を構築していくことにより、最終的に価値のある知識になることを一般的に集合知と呼ぶ。この集合知は、各個人あるいは各専門家の個々の決定や知識が必ずしも正しい訳ではなく、多人数の協調によって価値のあるものが適切に構築されていくという考え方に基づくものである。技術用日本語言語知識集合知サーバは、言い換え規則や言い換え辞書などの言語知識を集合知化していくためのサーバである。

① システム機能

ａ．言語知識（言い換え規則、言い換え辞書、文書範例）の協調作成支援機能

言語知識集合知サーバは、技術用日本語文書の作成・編集・検索を行うための言語知識データを協調作成するための支援機能を持つ。集合知サーバが保持している言語知識データの内容は文章、文、節、句に対する構造的な言い換えの規則と言い換え辞書として分野に関係なく利用可能な一般語彙辞書と、分野別の専門用語辞書である。

言語知識集合知サーバでは、人手あるいは統計的手法に基づいて協調的に言語知識を収集・編集・改良する作業を支援する機能を支援する。人手による収集・編集・改良機能としては、未登録の単語や専門用語を登録する際、既に辞書に登録されている単語から、類義語、同義語、関連語などを提示する機能がある。単語間関係を新規に登録する場合も同様の支援機能が用意される。

統計的手法による自動作成・編集・改良機能として、未登録の単語や用語を、蓄積された産業技術ドキュメントコーパスから自動抽出したり、語間関係(同義語、類義語、関連語、オントロジー)を自動抽出したり、言い換え規則を自動抽出したりする機能が提供される。

ｂ．言語知識データベースの更新管理機能

言語データを編集・改良するに当たり、多人数が同時に変更を行う場合は、データの整合性を保つことが容易ではない。そこで、何らかのポリシーを設け、多重のチェック機構を設置した上でデータの変更・更新を定期的に行うとともに、サーバ間やデータベース間における整合性の保持を管理する機能を提供する。

統計的手法によって言語知識を抽出する協調作業の場合、全ての言語知識データベースを単純に共有するのではなく、データごとに参照・編集・改変に対しアクセス制限を設ける。これは言語知識の編集・改良における整合性の保持のためであるのと同時に、外部に向けて公開することによって利益を損なう可能性のある言語知識データを公開しないようにするためのアクセス管理もかねる。同様に、データのアクセス制限ごとに利用ユーザにも閲覧・編集に関する権限付与を管理する必要がある。

② システム構成

技術用日本語言語知識集合知サーバのシステム構成図を、以下に示す。技術用日本語に

必要な言語知識に関するデータを各知識データベースに格納し、人手や統計的手法に基づく協調作成の支援機能と更新管理機能により知識データベースの構築及び管理を行う。技術用日本語言語知識集合知サーバは、企業内の機密情報と一般に公開できる知識を区別し、インハウス集合知サーバとパブリック集合知サーバの二つのタイプに分けられる。また、技術用日本語オーサリングシステムにおける規則・辞書ベース管理と通信を行い、技術用日本語に必要な言語知識を収集したり、規則ベースや辞書ベースを更新したりする。このようにして、様々な言語知識体系を構築するための様々な集合知サーバシステムを構成することができる。

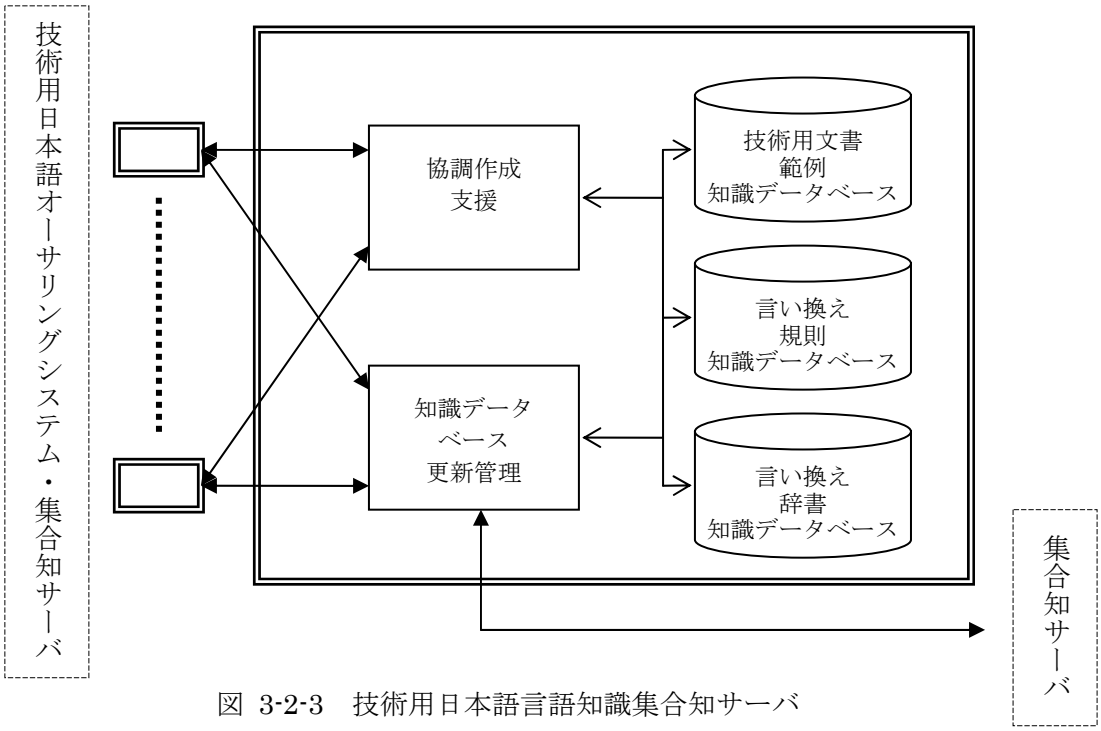


図 3-2-3　技術用日本語言語知識集合知サーバ

多数の技術用日本語言語知識集合知サーバから構成される集合知サーバシステムは、大きく 3 層のレイヤに分かれる。

ユーザレイヤ：集合知サーバと連携しながら文書の作成をおこない、時には、言語知識の改良にも参加する技術用日本語オーサリングシステムのレイヤ

インハウスレイヤ：企業レベル・部署レベルのドメインごとにユーザをまとめ、ドメイン固有の言語知識を集約するインハウス集合知サーバのレイヤ

パブリックレイヤ：各コミュニティの全てのユーザが共有する言語知識を格納するパブリックサ集合知サーバのレイヤ

ユーザレイヤにおいては、技術用日本語文書オーサリングに付随して言語知識データの作成・編集に関するインターフェイスが提供される。インハウスレイヤでは、パブリック

レイヤと連携しながらそれぞれのドメインにおける言語知識データの作成・更新の管理が行われる。パブリックレイヤでは、コミュニティ全体に共通する言語知識データが保持され、インハウスレイヤと定期的に同期しながら言語知識データの変更の衝突解消が行われ、集合知サーバシステムの整合性の維持管理が行われる。

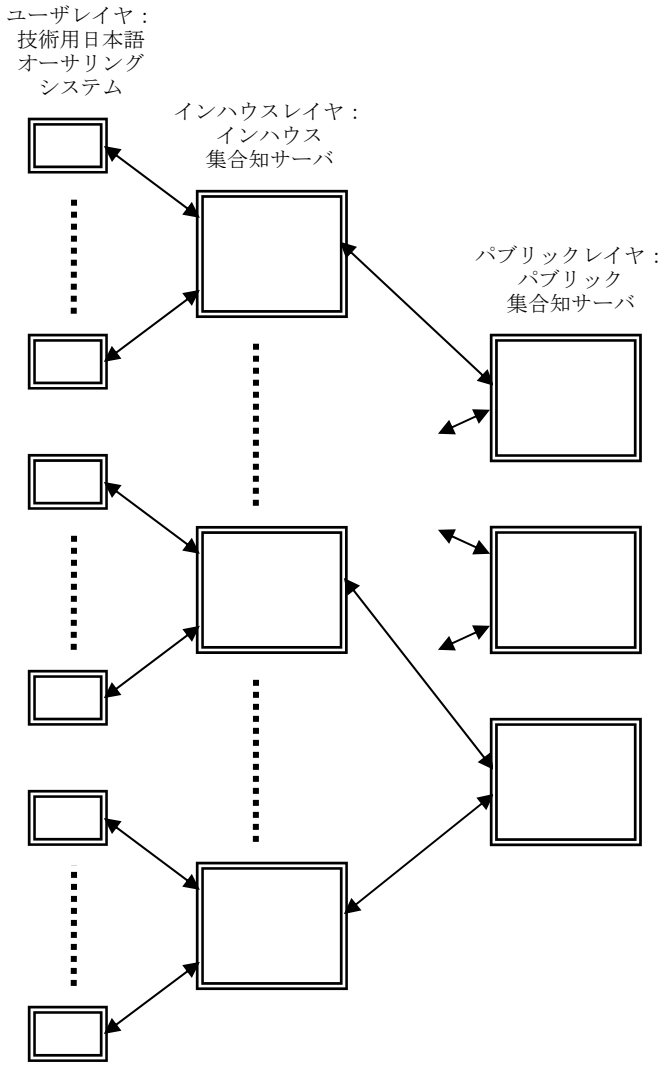


図 3-2-4　集合知サーバシステムの階層

３－２－３　技術用日本語アプリケーションシステム

（1）技術用日本語日英機械翻訳システム

技術用日本語は日本語を理解する人間にとって明晰であるだけでなく、自然言語処理ソフトウェアにとっても明晰でなければならない。

自然言語処理シフトウェアの一例である日英機械翻訳は、実用化の研究開発が始まって

20 年以上になるが、人間が普通に書き普通に理解できる日本語でも、正しく翻訳できない例はまだまだ多くある。この理由は、自然言語がもつ曖昧性に起因するものが大きい。そこで、翻訳する日本語の表現を明晰化して技術用日本語に書き換えれば、機械処理における曖昧性が減少し、正しい翻訳結果を出力することができるはずである。

ここでは、技術用日本語を作る明晰化と、英語への機械翻訳との関係について述べ、技術開発課題を整理する。

① システム機能

技術用日本語日英機械翻訳システムは、技術用日本語オーサリングシステムによって書き換えられた技術用日本語を英語に翻訳する。

技術用日本語アプリケーションシステムにおいて、技術用日本語オーサリングシステムは、その技術用日本語のテキストを CDL 表現内部形式に変換する。その CDL 表現内部形式のデータが、技術用日本語日英機械翻訳システムの入力となる。したがって、技術用日本語日英機械翻訳システムは、テキスト入力以外に、CDL 表現内部形式データを受付ける必要がある。

後述するトランスファ方式日英機械翻訳システムでは、日本語を解析した結果を、システム固有の形式の内部構造として保持している。一般的にその内部構造には、格フレーム構造や概念依存構造などがある。いずれの場合も、CDL 表現内部形式で入力された技術用日本語は、そのシステム固有の形式の内部構造に変換する処理が必要である。

システム固有の形式の内部構造に変換された後は、既存のシステムが持つ、トランスファ処理と言語生成処理を経て、英語訳文を出力することができる。

② システム構成

機械翻訳システムの方式には、規則ベースのトランスファ方式、ピボット方式、用例ベースの翻訳方式、統計ベースの翻訳方式などがあり、実際のシステムは、用途に応じていずれか、あるいは複数の方式が組み合わされて実現されている。

技術用日本語の主旨を考えると、機械翻訳で正しい訳文を出力するだけが目的ではない。機械翻訳以外でも効果を得るためには、訳文とは独立に日本語の構造を明確に解析する必要がある。これには日本語形態素解析・構文解析技術が必須であり、トランスファ方式の機械翻訳システムをベースにすべきである。また、CDL 表現内部形式を入力として英語への翻訳を行うためにも、日本語形態素解析・構文解析技術の出力としての内部構造が明確で、トランスファ以降の処理が独立している必要がある。そのためにも、トランスファ方式の機械翻訳システムが望ましい。

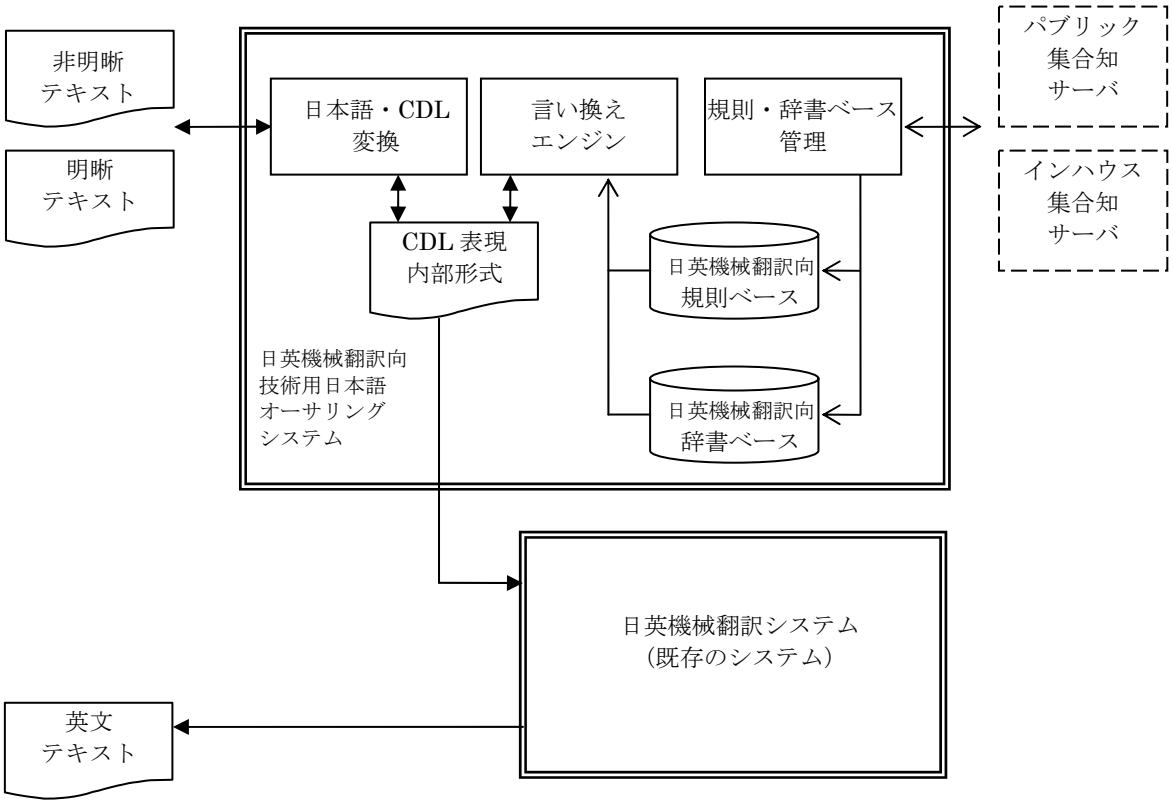


図 3-2-5　技術用日本語日英機械翻訳システムの構成

日英機械翻訳向け技術用日本語オーサリングシステムでは、非明晰テキストを明晰テキストに変換する。ここで実際には、非明晰テキストに対して、言い換えエンジンを利用して CDL 表現内部形式に変換する。言い換えエンジンは、日英機械翻訳向けの規則ベースと、辞書ベースを用いて変換を行う。

（2）技術用日本語文書検索システム

① システム機能

技術用日本語文書を CDL 化して文書データベース化することによって、単語間の係り受け関係の種別も記録され、更にフレーズ、センテンスといった大きな概念レベルでの階層化表現もデータベース化されることになる。これによって、単なる係り受け関係による検索や自然言語文によるいわゆる概念検索を超えた文書検索が実現されることになる。更に、より意味に踏み込んだ内容を把握しての検索が可能になり、検索性能の向上も可能になる。すなわち、

ａ．係り受け関係の有無だけでなく、関係の種別も考慮した検索が可能になり、より詳細で精度の高い検索が実現できる。

ｂ．属性で修飾された単語やフレーズなどという、単語より詳細あるいは大きな概念を単位とした検索も可能になる。
ｃ．自然言語検索文に出現する単語によるベクトル空間というような表層的特徴ではなく、検索文の単語間の関係も考慮した概念意味レベルの内容把握に基づいた検索（意味的検索）が可能になる。
ｄ．単語間類似性や文中での単語間関係の類似性を利用した意味に踏み込んだ類似検索も可能になる。（係り受け関係による検索や、自然言語検索文による概念検索と呼ばれる従来法でも類似検索は実現できるが、単語間類似性の利用にとどまる。）

② システム構成

技術用日本語文書は通常の日本語文書に比べて人間がその構造や意味を明晰に把握しやすくなると共に、コンピュータにとっても意味解析が容易になる点が大きな利点である。コンピュータによって捉えられた技術用日本語文書の概念意味（表層に近いレベルであり深層意味の理解は別途となる）は、自然言語テキスト概念意味の共通的（日本語や英語といった特定の言語に非依存）記述言語として、NPO 法人セマンティックコンピューティング研究開発機構(ISeC)で開発され、W3C(Web 技術の国際標準を定めるボストンに拠点を置くコンソーシアム)で国際標準化活動が進められている CDL(Concept Description Language)で記述され、原文書と共にデータとして蓄積、利用されることになる。

直接的効用は上述のように、キーワード検索の様な単純な検索でなく、意味に踏み込んだ検索が可能になることである。検索インターフェイスとしては、以下の３種類のインターフェイスによって利用されることになる。

ａ．自然言語文による問合わせ
ｂ．関係データベースに対する問合わせ言語 SQL 風の問合わせ言語
ｃ．CDL 表現に相当するグラフによる問合わせ

この a や c は b に変換されて検索が行われることになると想定される。したがって、b がこれらの中軸になるので、b を中心に問合わせ言語設計と、語彙間の関係も考慮するという点で CDL 表現に対する意味に踏み込んだ検索機能を実現する。

自然言語の概念意味の共通的記述言語に対応する CDL の仕様は、CDL.nl（CDL for natural language）である。技術用日本語オーサリングシステムにこの CDL.nl への変換機能を組みこむと、下図のような構成となる。

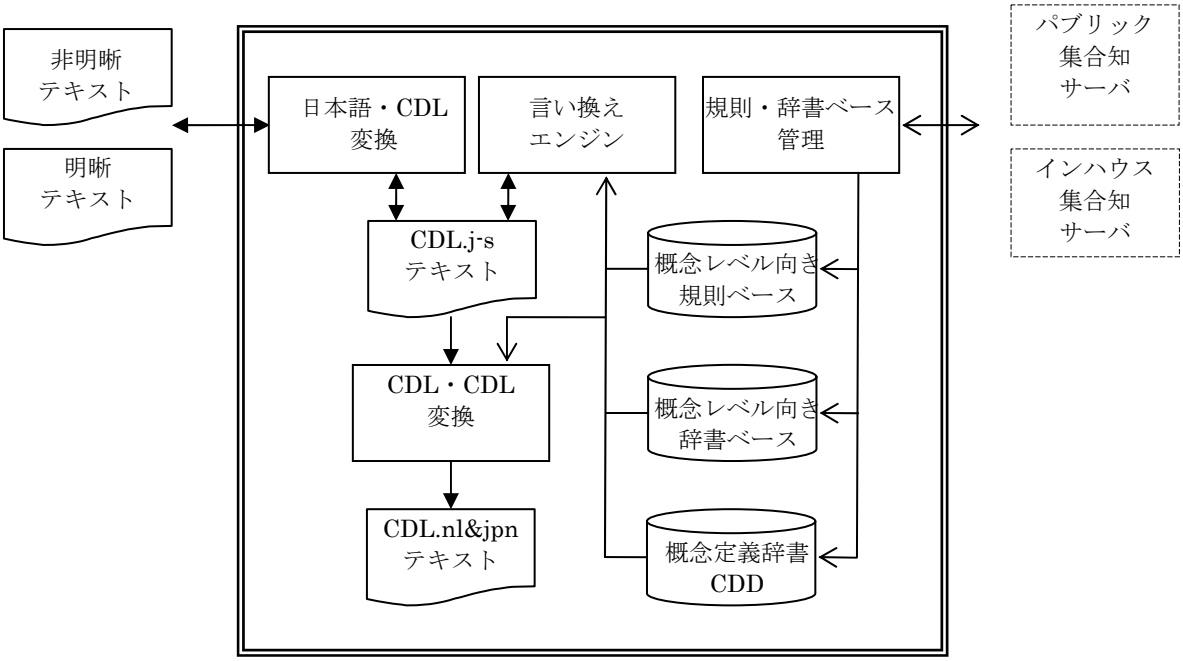


図 3-2-6　概念レベル向き技術用日本語オーサリングシステム

文書検索サーバシステムの方のシステム構成は、下図のようになる。

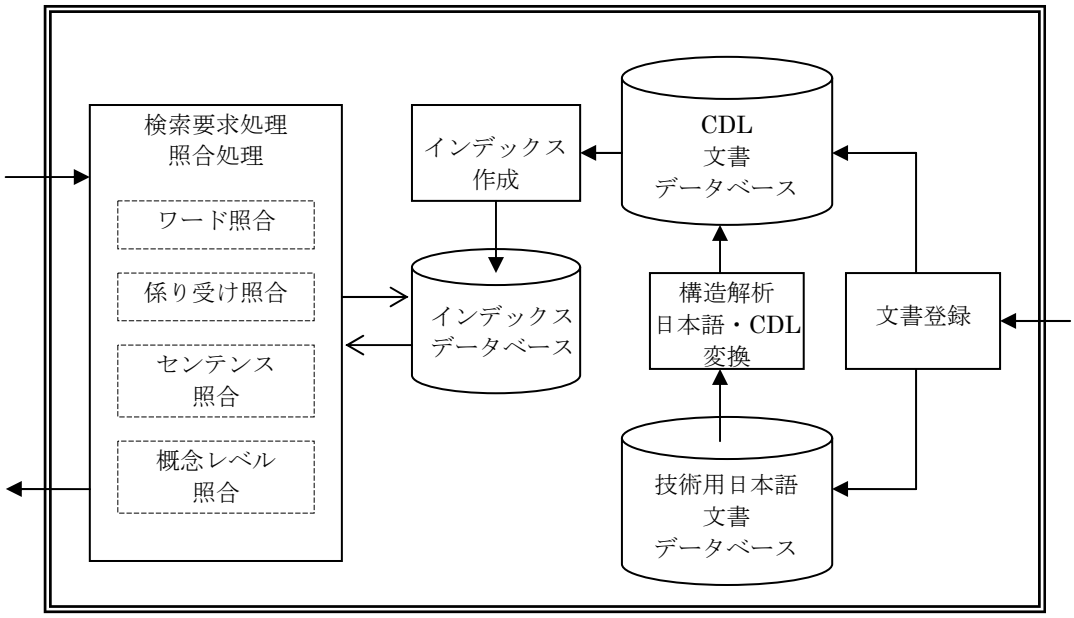


図 3-2-7　技術用日本語文書検索サーバシステム

３－３　技術用日本語の検証実験

本スタディでは３つの実験を行った。

（１）技術用日本語暫定仕様の作成

（２）機械翻訳用規則による言い換えと評価

（３）セマンティックオーサリングによる言い換え

以下、順に説明する。

３－３－１　技術用日本語暫定仕様の作成

技術用日本語暫定仕様は本スタディの知見に基づいて作成した「ＣＤＬ規則（２０規則)」と特許機械翻訳の知見に基づいて作成した「機械翻訳用規則（７規則)」を統合した暫定規則（２３規則）である。

暫定版の言い換え規則の基となったＣＤＬ規則は本スタディの日本語に関する技術動向で得た知見に基づいて長文特許をＣＤＬ変換して、その結果を演繹して抽出した。また、機械翻訳規則の７規則は長年にわたる特許機械翻訳の知見を基に作成した。 両者を比較、整理して２３規則にまとめ、暫定版の言い換え規則を作成した。

表 3-3-1　技術用日本語暫定規則『暫定規則（２３規則)』

項番	文法カテゴリ	規則ラベル	規則説明	備考
1	構文 複文	接続詞文分割	主文と従文の間が接続表現で結ばれているときは、分割することができる。その時主従の文の関係を明示する。 ※この規則は非常に個数が多い。談話関係と対応する。 Ex.「(従文）するが、(主文)」　→「(従文）する。しかし（主文)」に分割する。 Ex.「(従文）しておくと、(主文)」→「(従文）しておく。すると（主文)」に分割する	
2	構文 複文	引用句外出し	文中での引用文「～」は外に出して、後ろで、「そう」言ったとつなげる。 Ex(原文). 所長は得意気に、「これより美しい画面を見たことがありますか」と語りかけた。→ (技術用日本語文).「これより美しい画面を見たことがありますか」。所長は得意気にそう語りかけた。	
3	構文	補足説明処理	括弧などで、前文を節で補足説明をしているものは、括弧をとり、文中に埋め込むか、外に出す。 逆に、名詞句を括弧に入れて簡略化する変換は可能	機械3

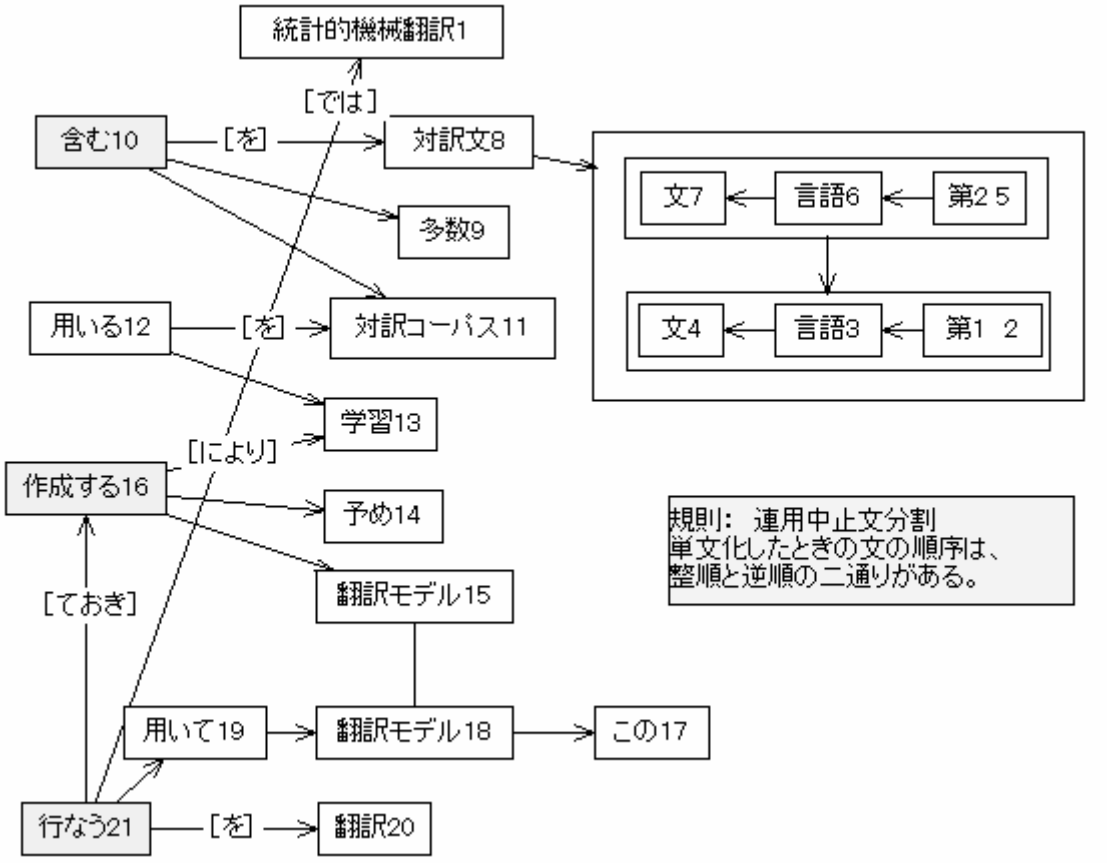
4	構文 複文	非制限用法文分割	連体修飾で非制限的用法は分割する。	機械 1
5	構文 重文	連用中止文分割	「(S1) し、(S2) した。」のような連用中止は、短ければ認める。長ければ分割する。	機械 1
6	構文	必須格補填	必須格は省略しない。特に主格に注意。	
7	構文	並列句文分割	n 個の並列句(節) を伴うときは、その主文を先頭に置き、並列句は文分割する。主文では、「以下の」を付ける。	機械 2
8	構文	並列句構文的均質性	並列句は、構文的に揃えることが望ましい。	
9	構文	主文先頭文分割	分割したときは、正順と逆順のどちらかを選択する。	
10	構文	並列句構成要素数明示	リストに分けたときは、個数を宣言したり、リスト要素に番号付けしたり、「先ず」と宣言する。	
11	構文	重い修飾句外出し	主格や目的格の格成分の修飾句・修飾節の語数が一定以上のときは、「以下のもの」や「以下のこと」で代用し、次の文でそれを展開する。 あるいは、体言止めで独立させ「を考える」、「とする」を補う。本文では「その」「それ」「前記」等で利用する。 Ex. …尤度変更手段は、出力チャンク作成手段により作成された出力チャンクと、当該出力チャンクに対応する入力文のチャンクとの対が、チャンク対記憶手段に記憶された第１のチャンク対と一致していることを検出して、…。(上で傍点を付与した部分は重いガ格であるので) → 出力チャンク作成手段により作成された出力チャンクと、当該出力チャンクに対応する入力文のチャンクとのチャンク対**を考える**。(これを踏まえて、後ろで「前記チャンク対」として代用する。)	
12	構文	大深度連体修飾	連体修飾は 2 段まで。3 段以上は警告する。機械的な解消策は、受け側の体言を元に戻し、文を切り、「それ」あるいは「その体言」で受けることである。	
13	構文	SVO 型変換	S+O の語数が一定以上であれば、S+V+O あるいは、O+V+S にする。 Ex(原文). 図３において、ソース言語のチャンク分けＥ＊の番号３，２，１のチャンクが、ターゲット言語のチャンク分けＪ＊の番号１，２，３のチャンクに対応している。→ (技術用日本語文). 図３において、ソース言語のチャンク分けＥ＊の番号３，２，１のチャンクが対応しているのは、ターゲット言語のチャンク分けＪ＊の番号１，２，３のチャンクである。	
14	構文	並列句語彙均等	並列句の用言、格助詞等では、同じ語彙パターンを繰り返す。(省略しない)	機械 4

15	形態	動詞機能表現排除	助詞相当句に対しては、本動詞は避ける。（平仮名標記として認める手もあるが）；これは英語でも過去分詞で名詞を修飾する場合は見られるが、不用意な言い方は良くない。 Ex. 「を用いた」: XML を用いた Web サービス→XML による Web サービス、或いは、XML の Web サービス（英語では by） Ex. 「で示す」: 右上に示すマーク→右上のマーク（英語では on） Ex. 「という」: 半導体というデバイス→半導体デバイス（英語でもヌル） ・・・等々	
16	形態	名詞多連続	名詞連続は 3 個以内とする。	
17	形態	修飾語「の」連続使用回避	「の」で連続して修飾するのは 3 個以内とする。	
18	語彙	冗長語句	冗長な(同じ言葉の繰り返しなど）表現は使わない Ex(原文). 主観的な評価でも５９．２％から６５．１％への評価の向上が見られた。→ (技術用日本語文). 主観的な評価でも５９．２％から６５．１％への向上が見られた。 Ex(原文).以下の説明では、～について説明する。→ (技術用日本語文).以下では、～について説明する。	
19	語彙	未定義専門用語	専門用語は必ず説明してから使う。	
20	語彙	非省略語彙	専門用語は省略しない。 Ex(原文)～チャンクと～チャンクとの対→ (技術用日本語文).～チャンクと～チャンクとのチャンク対	
21	構文	曖昧表現回避（機械翻訳用規則から追加）	日本語で多用される不必要に感覚的な表現を避ける。 ① 語句の関係の明確化 Ex (原文)…単語の**各々について**、…によって翻訳を行なう。→ (技術用日本語文)…各単語を、…によって翻訳する。	機械 5
22	構文	曖昧表現回避（機械翻訳用規則から追加）	日本語で多用される不必要に感覚的な表現を避ける。 ② 動作の結果を表す「もの」の具体化 Ex (原文)…を全て加算した**もの**が、…を意味する→ (技術用日本語)…を全て加算した値が、…を意味する	機械 6

23	構文	曖昧表現回避 (機械翻訳用規則から追加)	日本語で多用される不必要に感覚的な表現を避ける。 ③ 意味の多様な「する」「なる」を含む表現 Ex (原文)…との関係を逆にしたものとなっている→ (技術用日本語) …との関係が逆である	機械 7

CDL 規則による言い換え例を以下に示す。例文 4 が 4’ 或いは 4” に変換される。

4. 統計的機械翻訳では、第１の言語の文と第２の言語の文との対訳文を
多数含む対訳コーパスを用いた学習により予め翻訳モデルを作成しておき、
この翻訳モデルを用いて翻訳を行なう。



4’（逆順）. 統計的機械翻訳では、翻訳モデルを用いて翻訳を行なう。
その翻訳モデルは対訳コーパスを用いた学習により予め作成しておく。
その対訳コーパスは第１の言語の文と第２の言語の文との対訳文を多数含む。

4”（整順）. 対訳コーパスは、第１の言語の文と第２の言語の文との対訳文を多数含む。
その対訳コーパスを用いた学習により予め翻訳モデルを作成しておく。
統計的機械翻訳では、この翻訳モデルを用いて翻訳を行なう。

３－３－２　機械翻訳用規則による言い換えと評価

機械翻訳用規則による言い換えと評価は特許原文を人手により技術用日本語文に言い換え、「人にとっての読みやすさ」、「機械にとっての理解のしやすさ」が向上したかどうかを評価した。この評価は本スタディの最も重要な実験である。

適用する言い換え規則は本来なら「3-3-1 技術用日本語暫定仕様の作成 で作成した技術用日本語暫定仕様を用いるべきであるが、機械翻訳用規則（７規則）を利用した。理由は下記の２点である。

① 技術用日本語暫定仕様、ＣＤＬ規則に基づいて言い換えた技術用日本語文を理解できるアプリケーションプログラムが未だ存在しないのに対し、機械翻訳用規則に基づいて言い換えた技術用日本語文は機械翻訳システムが処理可能なこと。つまりコンピュータによるわかりやすさの評価に機械翻訳システムが使えること。

② 技術用日本語暫定規則（２３規則）は機械翻訳用規則（７規則）を包含しているため、少ない７規則の機械翻訳用規則で測った値以上の効果を上げる事が期待できる。

ここでは、実験の方法と結果を簡単に整理する。

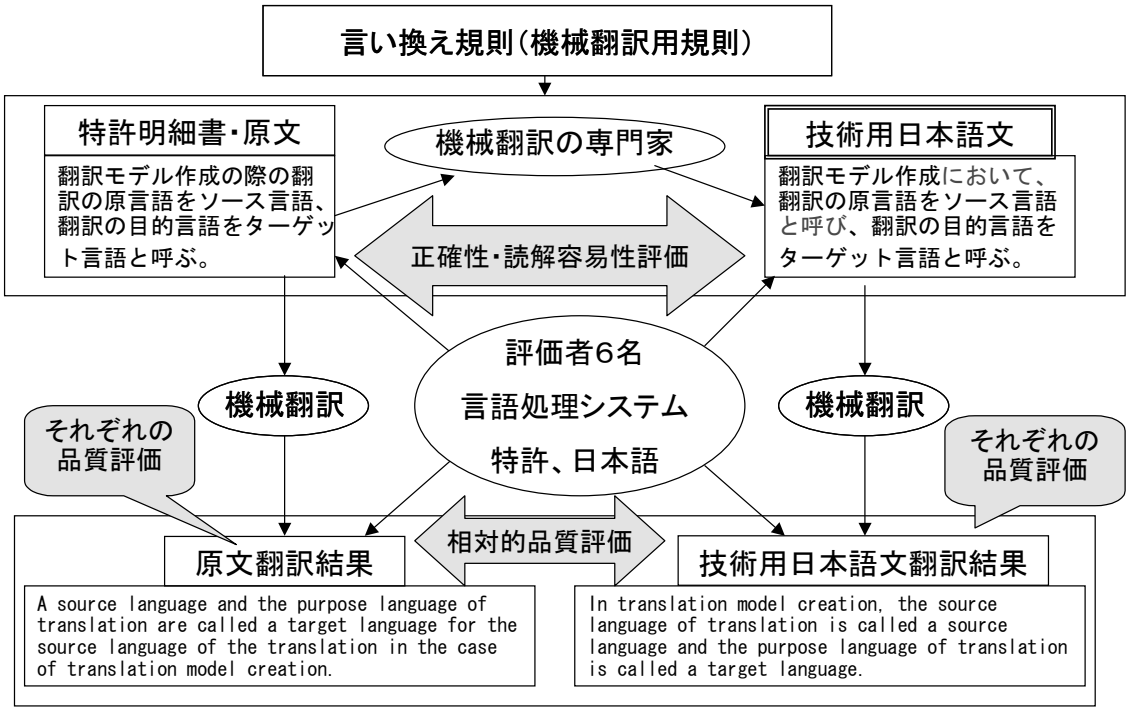


図 3-3-1 評価概要図

図 3-3-1 評価概要図 にあるように特許明細書・原文を機械翻訳の専門家が機械翻訳用規則を適用して技術用日本語文に言い換える。この原文と技術用日本語文を評価者が正確性と読解容易性の観点で評価して「正確性」と「人にとっての読みやすさ」を得た。次に原文と技術用日本語文を機械翻訳してそれぞれの機械翻訳結果を評価者が本文と技術用日本語文それぞれの機械翻訳結果を評価、次に相対的に良くなったかどうかを評価して「コン

ピュータにとっての読みやすさ」を得た。３つの特許文献の原文、338 文を評価した結果を表 3-3-2 評価結果 に示す。

表 3-3-2 評価結果

人にとっての読解容易性	**+47%向上**
コンピュータにとってのわかりやすさ	**+53%～69%向上**
正確性	**良し 80%**
・不的確（規則が不足）	**12%**

効果：機械翻訳用の言い換え規則（７規則）を適用して技術用日本語に言い換える事で、人にとっての読解容易性とコンピュータにとってのわかりやすさが共に向上することがわかった。

課題：正確性に関しては８０％が原文の技術内容を正確に言い換え出来ている事がわかったが、１２％は技術内容を正確に言い換えていない事が分かった。このことから正確な言い換えのための規則が不足している事がわかる。規則を追加した場合、追加した規則を検証することと、規則の増加に伴い言い換えを人手作業で行うことが難しくなることが予想される。そのため、何らかの機械的な支援が必要になる。

以下、実験の流れとここの評価結果に至るまでを要約する。

（1） 実験の流れ

実験の流れを図 3-3-3 に示す。 技術用日本語の書き換え実験の対象としたのは特許明細書 9 文献であり、専門家が評価したのは ＊を付けた 3 文献である。

表 3-3-3 実験データ一覧

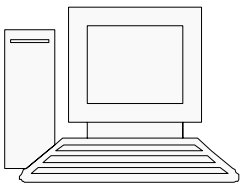
特許公開 2005-000031＊	7,344 文字
特許公開 2005-003747＊	20,714 文字
特許公開 2005-004500	8,549 文字
特許公開 2005-004707	1,147 文字
特許公開 2005-010347	12,351 文字
特許公開 2005-025312	5,470 文字
特許公開 2005-025474＊	8,113 文字
特許公開 2005-025653	15,303 文字
特許公開 2005-025659	13,258 文字
計	92,249 文字

原文
この式にベイズの定理を適用する事により、^E＝argmax「〈E〉」P(E)P(J｜E)が得られる。
この式のうちP(E)は言語モデルと呼ばれるものであり、ターゲット言語Eの文における単語の出現尤度を示すものである。

a. 原文を機械翻訳システムに入力して英語の「原文翻訳結果」を得る。

原文	原文翻訳結果
この式にベイズの定理を適用する事により、^E＝argmax「〈E〉」P(E)P(J｜E)が得られる。	^E=argmaxEP(E) P(J|E) is obtained by applying Bayes' theorem to this formula.
この式のうちP(E)は言語モデルと呼ばれるものであり、ターゲット言語Eの文における単語の出現尤度を示すものである。	Inside P (E) of this formula is called a language model, and shows the appearance likelihood of the word in the sentence of target language E.

機械翻訳システム



b.. 機械翻訳の専門家がこの原文翻訳結果を確認して良い機械翻訳であるかどうかを判定する。

c. 良い機械翻訳で無い場合は、機械翻訳用規則に基づき機械翻訳の専門家が原文を「技術用日本語文」に変換する。

機械翻訳の専門家　機械翻訳用規則（7規則）

判定により技術用日本語文なし

原文	技術用日本語文	原文翻訳結果
この式にベイズの定理を適用する事により、^E＝argmax「〈E〉」P(E)P(J｜E)が得られる。		^E=argmaxEP(E) P(J|E) is obtained by applying Bayes' theorem to this formula.
この式のうちP(E)は言語モデルと呼ばれるものであり、ターゲット言語Eの文における単語の出現尤度を示すものである。	この式のP(E)は言語モデルと呼ばれるものであり、ターゲット言語Eの文における単語の出現尤度を示す。	Inside P (E) of this formula is called a language model, and shows the appearance likelihood of the word in the sentence of target language E.

d. 技術用日本語に変換された文を機械翻訳システムに入力して英語の「技術用日本語翻訳結果」を得る。

評価対象データ

原文	技術用日本語文	原文翻訳結果	技術用日本語翻訳結果
この式にベイズの定理を適用する事により、^E＝argmax「〈E〉」P(E)P(J｜E)が得られる。		^E=argmaxEP(E) P(J|E) is obtained by applying Bayes' theorem to this formula.	
この式のうちP(E)は言語モデルと呼ばれるものであり、ターゲット言語Eの文における単語の出現尤度を示すものである。	この式のP(E)は言語モデルと呼ばれるものであり、ターゲット言語Eの文における単語の出現尤度を示す。	Inside P (E) of this formula is called a language model, and shows the appearance likelihood of the word in the sentence of target language E.	P (E) of this formula is called a language model, and shows the occurrence likelihood of the word in the sentence of target language E.

評価者
（知財の専門家）
（言語処理システムの専門家）
（日本語の専門家）

英語の評価

原文翻訳結果	技術用日本語翻訳結果	知財③ 原文翻訳結果	知財③ 技術用日本語翻訳結果	知財③ 比較評価	知財③ コメント（英）知財③
^E=argmaxEP(E) P(J|E) is obtained by applying Bayes' theorem to this formula.					
Inside P (E) of this formula is called a language model, and shows the appearance likelihood of the word in the sentence of target language E.	P (E) of this formula is called a language model, and shows the occurrence likelihood of the word in the sentence of target language E.	×	◎	◎	原文の係り受けエラーが解消し、ほぼ正確な英文となった。

日本語の評価

原文	技術用日本語文	技術用日本語編集に関するコメント	知財① 技術用日本語正確性	知財① 技術用日本語読解容易性	知財① コメント（日）知財①
この式にベイズの定理を適用する事により、^E＝argmax「〈E〉」P(E)P(J｜E)が得られる。					
この式のうちP(E)は言語モデルと呼ばれるものであり、ターゲット言語Eの文における単語の出現尤度を示すものである。	この式のP(E)は言語モデルと呼ばれるものであり、ターゲット言語Eの文における単語の出現尤度を示す。	○「ものであり」は意味的に不要であるが、連用中止用法の動詞句には残した。「ものである」が残ることで、ここで「切れ目」であることが読み手に伝わりやすい。	○	○	「示すものである」→「示す」とするならば、「呼ばれるものであり」→「呼ばれ」とすべき。

図 3-3-2　実験の流れ

実験の評価者は 表 3-3-4 評価者リスト の 6 名である。特許の専門家を評価者とした理由は、正確性の判断とユーザとしての評価を期待したためである。言語処理システムの専門家を評価者とした理由は、技術用日本語関連のシステム提供者の立場で、人に理解しやすいこと と コンピュータにとってわかりやすいこと のバランスをとるようなコメントを期待したためである。日本語の専門家を評価者とした理由は、技術用日本語仕様が一般の人に理解されやすい日本語となるようなコメントを期待したためである。

表 3-3-4 評価者リスト

専門分野	名前
特許の専門家	知財①、知財②、知財③
言語処理システムの専門家	システム①、システム②
日本語の専門家	日本語

（2） 評価基準と評価シート(日本語)

評価対象データの原文と技術用日本語文を比較して日本語の正確性と日本語の読解容易性の評価を行った。日本語の正確性の判断は機械翻訳用規則（7 規則）の問題点の洗い出しの資料、日本語の読解容易性の評価は人にとっての読みやすさの指標となる。日本語で行った正確性と読解容易性の評価基準を 表 3-3-5 に示す。評価に使ったシートを 表 3-3-6 日本語評価シートに示す。

表 3-3-5 日本語の評価基準

① 日本語の正確性	技術用日本語への変換で「原文の技術内容が過不足なく正確に反映されている／反映できていない」という観点で評価した。	○：原文の技術内容が過不足なく正確に反映されている
		×：原文の技術内容を正確に反映できていない
		－：○ではないが、×と判断するほどではない
② 日本語の読解容易性	技術用日本語に言い換えた文で読みやすさ（分かりやすさ）が変化したかどうかを評価した。	○：読みやすくなった、分かりやすくなった
		×：読みにくくなった、分かりにくくなった
		－：読みやすさ、分かりやすさは変わらない

表 3-3-6 日本語評価シート

原文	技術用日本語文	技術用日本語編集に関するコメント	知財①		
			技術用日本語正確性	技術用日本語読解容易性	コメント(日)知財①
この式にベイズの定理を適用する事により、^E＝argmax「<E>」P(E)P(J｜E)が得られる。					
この式のうちP(E)は言語モデルと呼ばれるものであり、ターゲット言語Eの文における単語の出現尤度を示すものである。	この式のP(E)は言語モデルと呼ばれるものであり、ターゲット言語Eの文における単語の出現尤度を示す。	○「ものであり」は意味的に不要であるが、連用中止用法の動詞句には残した。「ものである」が残ることで、ここで「切れ目」であることが読み手に伝わりやすい。	○	○	「示すものである」→「示す」とするならば、「呼ばれるものであり」→「呼ばれ」とすべき。
後者のP(J｜E)が翻訳モデルと呼ばれるものであり、第2の言語の文Eから第1の言語の文Jが生成される確率を表す。	後者のP(J｜E)が翻訳モデルと呼ばれるものであり、「{第2の言語の文Eから第1の言語の文Jが生成される確率}」を表す。	○埋め込み表現の範囲を明確化。	○	－	同上。
この言語モデルと翻訳モデルとを用いて、入力文Jに対し前述した条件付確率P(E)P(J｜E)が最大となる翻訳文^Eを生成する。	この言語モデルと翻訳モデルとを用いて、「{入力文に対して条件付確率P(E)P(J｜E)が最大である}」翻訳文^Eを生成する	○「～となる」は意味が多様であり適切に訳すことが難しいことがあるので書き換え。 ○「に対して」も翻訳において適切に訳すことが難しい場合が多い。 ○埋め込み文のスコープを明確化。	○	－	大差なし。

（3） 評価基準と評価シート(英語)

評価対象データの原文翻訳結果と技術用日本語翻訳結果を見て、それぞれの翻訳結果の評価と技術用日本語翻訳結果が良くなったかどうか評価した。この評価はコンピュータにとってのわかりやすさの指標となる。英語の評価基準を 表 3-3-7 英語の評価基準に、評価に使ったシートを 表 3-3-8 英語評価シートに示す。

表 3-3-7 英語の評価基準

① 原文翻訳結果の評価	原文の翻訳結果を評価する。	◎：原文技術内容がほぼ正確に理解でき、英文としても自然である
		○：致命的な誤訳がなく、原文の技術内容が推測できる
		×：意味不明、もしくは本質的な誤訳を含む
		－：評価せず
② 技術用日本語翻訳結果の評価	技術用日本語の翻訳結果を評価する。	◎：原文の意味がほぼ正確に理解でき、英文としても自然である
		○：致命的な誤訳がなく、原文の大体の意味が推測できる
		×：意味不明、もしくは本質的な誤訳を含む
		－：評価せず

③ 英訳の比較評価	原文翻訳結果と技術用日本語翻訳結果を比較して、その正確さ及び分かりやすさ、読みやすさ、英文としての自然さ等、総合的に判断してどちらの翻訳精度が高いか評価する。	◎:技術用日本語翻訳結果の方が圧倒的に優れている
		○:技術用日本語翻訳結果の方が優れている
		×:技術用日本語翻訳結果の方が劣っている
		ー:変化なし又は双方とも意味をなしていない

表 3-3-8 英語評価シート

原文翻訳結果	技術用日本語翻訳結果	知財③			
		原文翻訳結果	技術用日本語翻訳結果	比較評価	コメント(英)知財③
^E=argmaxEP(E) P(J|E) is obtained by applying Bayes' theorem to this formula.					
Inside P (E) of this formula is called a language model, and shows the appearance likelihood of the word in the sentence of target language E.	P (E) of this formula is called a language model, and shows the occurrence likelihood of the word in the sentence of target language E.	×	◎	◎	原文の係り受けエラーが解消し、ほぼ正確な英文となった。
The latter P (J|E) is called a translation model and the probability that sentence J of the 1st language will be generated from sentence E of the 2nd language is expressed.	The latter P (J|E) is called a translation model and expresses the probability that sentence J of the 1st language will be generated from sentence E of the 2nd language.	○	◎	○	原文でも意味は大体通じるが、技術用日本語は後半へのつながりが明解になり改善がみられる。
Conditional probability P(E) P(J|E) mentioned above to input sentence J generates translation^E used as the maximum using this language model and a translation model.	Translation^E whose conditional probability P(E) P(J|E) is the maximum to an input sentence is generated using this language model and a translation model.	×	○	○	原文は意味不明なレベル。技術用日本語は分かりづらいが誤訳ではない。

（4） 人にとっての読解容易性　評価集計表

３つの特許公開番号の「読解容易性 平均」の平均を算出したのが表 3-3-9 日本語・評価集計表（百分率）の「読解容易性 全文献平均」である。ここから機械翻訳用規則（７規則）の人に対する読解容易性の変化を読み取る。

５３％　　○：わかりやすくなった、読みやすくなった

６％　　×：わかりにくくなった、読みにくくなった

４１％　　－：わかりやすさ、読みやすさは変わらない

読解容易性の変化は○の 53% から ×の 6% を減じた値とすると、機械翻訳用規則（７規則）は人にとっての読解容易性に４７％の向上をもたらしたと考えられる。

また、×の６％は人にとっての読解容易性を下げていることであり、機械翻訳用規則（７規則）による言い換えの予期しないマイナス効果（以下、副作用という。）と見られる。また、特に特許関係の評価者が正確性の評価で他の評価者に比べて辛い評価を行っておりコメントも多い。このことは特許特有の注意点があり、何らかの回避策が必要であることを示している。

表 3-3-9 日本語・評価集計表（百分率）

				評価者→	知財①		知財②		知財①		システム①		システム②		日本語						
特許公開番号	原文数	変換文数	技術用日本語の文数	日本語の評価	正確性	読解容易性	正確性	読解容易性	正確性	読解容易性	正確性	読解容易性	正確性	読解容易性	正確性	読解容易性	日本語の評価	正確性 平均	読解容易性 平均	正確性 全文献平均	読解容易性 全文献平均
2005-31	77	43	112	○	34%	89%					80%	74%			96%	44%	○	70%	69%		
				×	24%	11%					20%	2%			1%	7%	×	15%	6%		
				－	42%	0%					0%	23%			4%	50%	－	15%	24%		
2005-10347	142	93	160	○							88%	40%	88%	78%	94%	13%	○	90%	44%		
				×							13%	1%	10%	6%	6%	3%	×	10%	4%		
				－							0%	59%	3%	16%	0%	84%	－	1%	53%		
2005-25474	119	100	125	○	72%	87%	79%	28%	74%	36%	93%	53%			90%	29%	○	82%	47%	80%	53%
				×	8%	9%	17%	5%	18%	10%	5%	6%			10%	9%	×	11%	8%	12%	6%
				－	19%	4%	4%	67%	9%	54%	2%	41%			0%	62%	－	7%	46%	8%	41%

（5） コンピュータにとっての明晰性

「表 3.3-10 英語・評価集計表（百分率）」から コンピュータにとってのわかりやすさ を読み取る方法に下記の３とおりが考えられる。

① 英語・評価集計表（百分率）の全文献平均の改善の評価から機械翻訳用規則（７規則）のコンピュータにとってのわかりやすさの向上を読み取る方法である。

７２％　　◎○：技術用日本語翻訳結果の方が（圧倒的に）優れている

３％　　×：技術用日本語翻訳結果の方が劣っている

２５％　　－：変化なし、又は双方とも意味をなしていない

このことから、機械翻訳用規則（７規則）はコンピュータにとってのわかりやすさの向上に６９％(72%-3%)程度の向上をもたらしていると考えられる。

② 原文翻訳結果と技術用日本語翻訳結果それぞれのプラス評価の◎と○の全文献平均の値を比べる方法である。全文献平均の◎○合算の欄を見ると原文翻訳結果：２３％ が 技術用日本語翻訳結果：８２％ に変化しており、５９％の向上があった。

③ 原文と技術用日本語文それぞれのマイナス評価×の全文献平均の値を比べる方法である。全文献平均の×の欄を見ると原文翻訳結果：６９％　が 技術用日本語翻訳結果：１６％ 変化しており、５３％の改善があった。

これらのことから、機械翻訳用規則（７規則）は少なくとも５３％、高く見ると６９％程度、コンピュータにとってのわかりやすさの向上をもたらしていることが判明した。

表 3-3-10 英語・評価集計表（百分率）

評価者→					知財①			知財③			システム①			システム②				文献別平均			文献別平均 ◎と○を合算					全文献平均		
特許公開番号	原文数	変換文数	技術用日本語の文数	英語の評価	原文	技術用	改善	原文	技術用	改善	原文	技術用	改善	原文	技術用	改善	英語の評価	原文	技術用	改善	英語の評価	原文	技術用	改善	英語の評価	原文	技術用	改善
2005-31	77	43	112	◎	0%	0%	0%				1%	53%	49%				◎	1%	27%	24%								
				○	0%	74%	74%				13%	20%	20%				○	7%	47%	47%	◎○	7%	74%	71%				
				×	97%	24%	0%				53%	22%	6%				×	75%	23%	3%	×	75%	23%	3%				
				－	3%	3%	26%				32%	4%	26%				－	17%	4%	26%	－	17%	4%	26%				
2005-10347	142	93	160	◎										1%	12%	11%	◎	1%	12%	11%								
				○										44%	68%	43%	○	44%	68%	43%	◎○	45%	80%	54%				
				×										55%	20%	6%	×	55%	20%	6%	×	55%	20%	6%				
				－										0%	0%	40%	－	0%	0%	40%	－	0%	0%	40%				
2005-25474	119	100	125	◎	0%	0%	0%	0%	23%	18%							◎	0%	11%	9%								
				○	2%	97%	97%	30%	64%	64%							○	16%	80%	80%	◎○	16%	92%	90%	◎○	23%	82%	72%
				×	98%	3%	0%	58%	10%	0%							×	78%	6%	0%	×	78%	6%	0%	×	69%	16%	3%
				－	0%	0%	3%	12%	3%	17%							－	6%	2%	10%	－	6%	2%	10%	－	8%	2%	25%

３－３－３　セマンティックオーサリングによる言い換え

「セマンティックオーサリングによる言い換え」は報告書(本編)の「1.1.4 文書制作高度化のためのセマンティックオーサリング」で述べた方法を用いて実際に手作業で特許原文をグラフィカル表示に変換した。目的は既存のセマンティックエディタで特許明細書のグラフィカルなオーサリングが可能かどうかを確認する事である。グラフィカルなオーサリングが可能であれば、技術用日本語プラットフォームシステムで述べた技術用日本語オーサリングシステムにも応用する事が可能となる。

具体的にはセマンティックオーサリングによる言い換えは人手でセマンティックエディタを用いて特許明細書のテキストをグラフ表示に変換した。今回、グラフ表示から実際に技術用日本語文を生成する事は実施していない。

なお、この実験により生成したグラフ表示を「参考資料：実験結果 3.セマンティックオーサリングによる言い換え」にまとめた。

本実験の結果、分かった良い点と課題を記す。

良い点

① グラフィカル表示で特許明細書が記述できる。
　実験対象とした８文献の特許明細書の原文をグラフィカル表示に言い換えできた。

② 特許明細書の作成支援にグラフィカル表示のユーザインタフェースが有効。
　オーサリング作業そのものが特許の専門家でない人の特許明細書の理解を助ける。
　グラフィカル表示から技術用日本語を生成できれば専門家でなくとも特許明細書が書ける。

課題

① グラフィカル表示から技術用日本語文の生成が必要。

② 正確性の評価が難しい。
　理由は原文とグラフィカル表示から生成された文が対応しないためである。

グラフィカル表示に言い換える規則を「基本仕様と詳細仕様」にまとめる。

（1）基本仕様

①　単文化
　述語は一つ(複文、重文を単文に分解する)

②　制限用法、非制限用法の明確化
　関係節の制限用法と非制限用法との区別を行う

③　談話関係による単文の結合
　談話関係オントロジを使い単文間の関係を明確化する

④　　共参照

共参照を使い、意味構造をグラフ化する

（2）セマンティックオーサリング使用における詳細仕様

今回の言い換えは手作業で行ったため、文章を構造化する時に人によるばらつきが生じる。これを防ぎある程度の成果物の標準化を行うために詳細仕様を規定する。これはプログラム言語によるソフトウエア開発時のコーディング規約に対応する。また逆にこの詳細仕様が構造化文章のノウハウになる。

① 文章をよく読み全体を理解する
② 原文章をそのまま構造化するのではなく意訳する
③ 対象領域に応じて構造化
④ 単文化して談話関係等(45 種類)で結ぶ
⑤ 使用述語の制約
⑥ $0、$1 等で統語的構成素を示す
⑦ #tm、#0、#1 等で照応・共参照を示す
⑧ 【数式】対応

セマンティックオーサリングの例を図 3-3-3 に示す。

【0034】
好ましくは、チャンク方式の翻訳モデルは、第2の言語をソース言語、第1の言語をターゲット言語とする、チャンクの並べ替えモデルを含む。そしてチャンク並べ替え手段は、チャンク翻訳手段から出力される出力チャンク列の各々について1又は複数通りのチャンクの並べ替えを行ない、各チャンクの並べ替えモデルから算出されるチャンク並べ替えの尤度と、当該出力チャンク列に含まれる出力チャンクの各々に対して算出されている尤度とから、各並べ替えの尤度を算出し、最も高い尤度を持つチャンクの配列を翻訳文として出力するための手段をさらに含んでもよい。

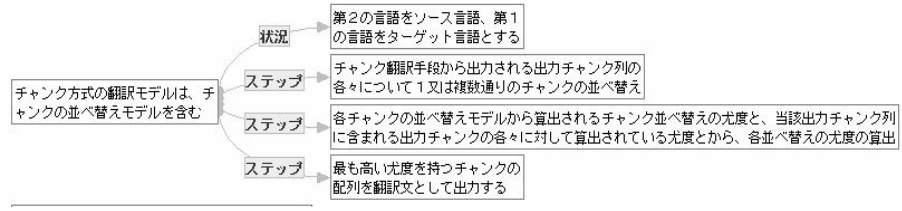


図 3-3-3 セマンティックオーサリングの例

３－４　市場調査、事業シミュレーション

知的生産性の向上及び経済活性化の観点から、国内グローバル企業等へのヒアリング結果等に基づき、経済効果をまとめた。

３－４－１　技術用日本語の適用対象と効果

「人に対する明晰性」については、ドキュメントの内容が正確かつ速やかに理解されるという時間面での効率性向上に加えて、専門知識が豊富とはいえない者、例えば出資者らが技術の内容を把握し易くなることから、産業技術の流通の活性化が図られることを示した。

「コンピュータに対する明晰性」については、実験で検証したとおり、機械翻訳精度の向上への効果は明らかで、大きな非関税障壁となっている、産業技術ドキュメントの翻訳に係るコスト及び品質の状況を大きく改善することを示した。加えて、技術用日本語の仕様を前提とした機械翻訳エンジンの開発やチューンナップも期待されることを示した。

特に、特許文書は、独特の表現形態や記載様式で特に読解や機械処理が困難であるが、二重投資の回避、技術流通の活性化を図る上で、「不特定多数の者」による利用の促進が重要で、検索効率向上の観点からも、技術用日本語を適用する効果がとりわけ大きな分野であることを示した。

更に、読解支援や自動要約作成への展開可能で、こうした技術進歩が我が国の知的生産性の向上、ひいては産業の発展にもたらすことを示した。

これらの経済効果については、「創造」、「保護」、「利用」からなる技術用日本語利用サイクルのステップ毎にまとめた。加えて、ステップ間のドキュメントの循環と再利用が有効で、各ステップ内で完結させず、技術論文を特許明細書のベース、特許明細書をライセンス情報や外国出願の原文といった形で、次に効率よく循環させていくことで更に効率化が促進され、我が国の知的生産性の向上に資することを示した。

３－４－２　市場規模と経済的効果

年間の作成件数が最も正確に把握可能な特許文書を中心に、効果を試算した。

「外国出願に伴う翻訳作業の年間市場規模」を約 1,168 億円と試算し、原文に比して、技術用日本語の翻訳は、少なくとも２０％程度のコスト削減効果は見込めることから、年間 230 億円強のコストダウンとなることを示した。

また、翻訳の品質が不完全であることに起因する権利取得失敗の損失についても示した。その損失が仮に 1%程度であるとしても、特許出願が急増している中国に関して、貿易額は 1,324 億ドル（2003 年）年間約 13 億ドルの損失となる。

加えて、翻訳コスト削減等により我が国の外国出願が活性化し、グローバル出願率 22%（米国 44%、欧州 60%）の大きな引き上げによる国際競争力強化の可能も示した。

その他の産業技術ドキュメントについても、その規模が大きいことを示した。例えば、

研究論文については、その被引用状況から、日本語が低からぬ非関税障壁として存在していることを示した。また、取扱説明書については、翻訳コストだけでも年間 200 億円のコスト削減効果が見込めることを示した。

３－４－３　予測されるサービス事業展開

技術用日本語の普及により知的生産性の向上に資する、種々のサービス事業の新規創出や高度化が見込まれ、二つの例を示した。いずれも文書をベースとした事業であり、技術用日本語プラットフォームにより新しい活力を与えることが期待できる。

（１）先進的知識マネジメント

多くの企業では、単なる文書共有システム、全文検索システムの導入のレベルにとどまっている知識マネジメントに対して、ドキュメント自体を知識として共有可能な形態で記述することを可能にする技術用日本語が大きなブレークスルーを与えることを示した。

（２）技術用日本語文書の作成支援－知財サイクルのワンストップサービス－

技術用日本語の使用をサポートする新規サービス事業として、「その作成を支援するツールやサービスの提供や作成代行」「技術用日本語環境における翻訳関連事業」及び「特許明細書等のバックファイルの技術用日本語化による利用性向上」などの展開の可能性を示した。

３－５　まとめ

以上、本年度のスタディの成果をステップ毎にまとめると以下の図のようになる。

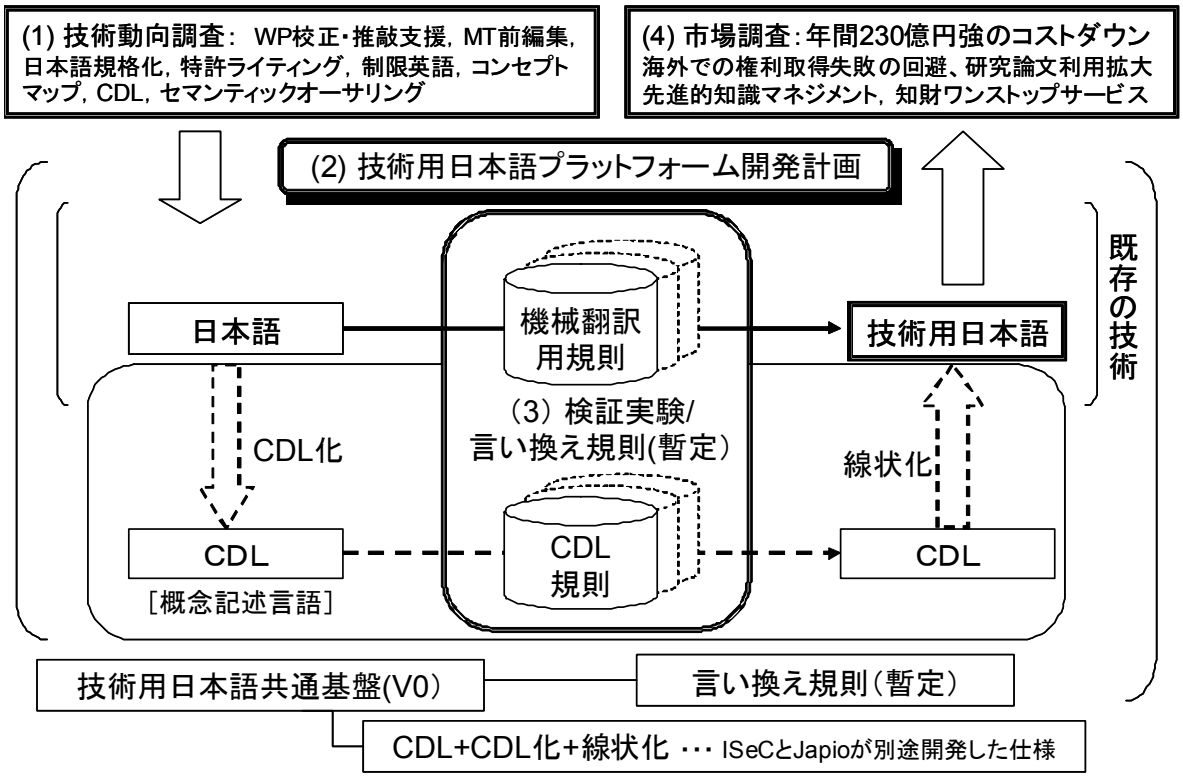


図 3-5-1　スタディのステップ対応の成果

本スタディが掲げる経済活性化とは、海外取引をする企業関係者に、知的基盤技術の充実の大切さに警鐘を鳴らすものであり、日本が国際化するグローバル経済を安心、安全に推進するために国家として政策的に遂行すべき課題を提起することを主たる目的として行ってきた。

その趣旨に叶い、本スタディの成果は特定の業界に経済的効果をもたらすというより、日本全体の経済活動の活性化に貢献する知的基盤技術の開発提案であり、技術用日本語を産業日本語と見なせば、産業ドキュメントの制作、翻訳、検索、要約、出版、知識処理の各々に対して高度化、高効率化、高精度化をもたらすことが実現でき、グローバルな産業活動の知的基盤強化になり、海外ビジネスを積極的に推進する日本の企業にとって文書処理技術からの支援となることが明らかになった。

一例として、知的財産分野を見ると、特許明細書は特許の権利化における唯一のドキュメントであり、日本語の言語処理技術の強みで、特許庁では、特許の全文検索や機械翻訳サービスを提供している。しかし、特許戦略をグローバルに展開する企業には、翻訳のコストは非常に高いため、海外出願はなかなか進まないという日本語の弱みが出る。本スタディでは、海外出願に係るコストダウンの実現が可能であることが確認できた。これは、知的財産権取得の世界レベルでの競争激化が進むなか、日本の知的財産権を保護するため

にきわめて重要な成果である。

こうした技術用日本語の効果を更に発展させれば、産業活動全体に大きな経済効果をもたらすことが期待できる。例えばWebサービスや、行政サービス、特に言語情報が絡む電子政府サービスにおいて効果的であることは明らかであろう。

本スタディでは、技術用日本語プラットフォームの開発を進め、技術用日本語の利用を広くわが国産業界に浸透させていくために、大きくスタディフェーズ、開発フェーズ及び運用フェーズの３つの段階からなる開発スケジュールを策定し、予定として考えられる開発経費及び開発体制を含め、基本的な開発計画を示した。

まず、スタディフェーズは、本スタディを含む１年半のフィージビリティスタディの段階であって、開発課題として掲げた「共通基盤仕様」を第１次仕様に仕上げていき、「プラットフォームシステム」及び「アプリケーションシステム」の基本設計を行う。その上で、詳細な開発計画を策定する。

次に開発フェーズは、3年間を見込んでおり、この段階で実運用システムを開発し、その先の運用フェーズに移行することを予定している。具体的には、「共通基盤仕様」の改良とともに「プラットフォームシステム」及び「アプリケーションシステム」のプロトタイプシステムの開発、そして運用システムの開発・改良へと進め、平成２４年度以降は、各種サービスを立ち上げ、技術用日本語をベースとする知の生産環境の統合的な整備への活動を展開することを計画している

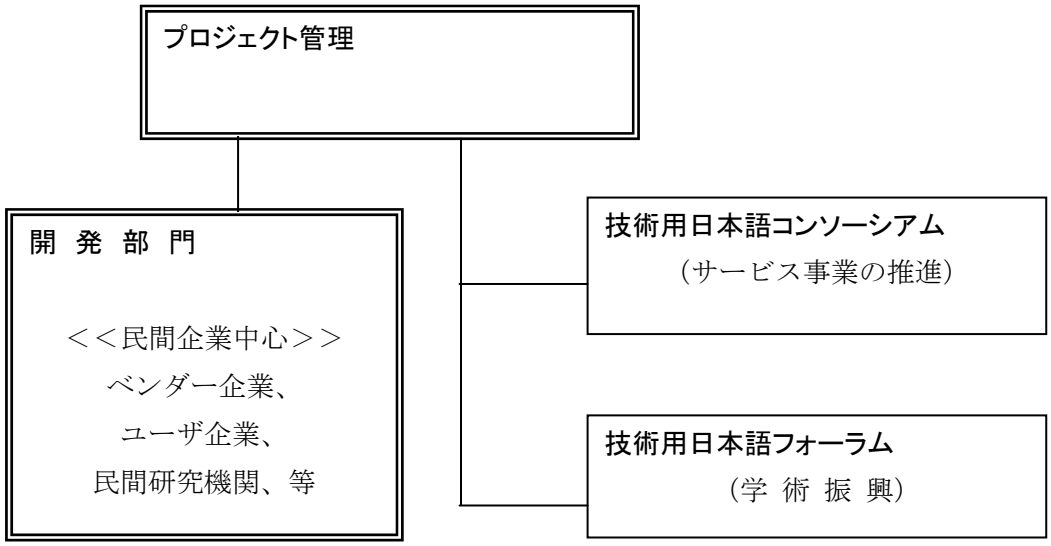


図 3-5-2 技術用日本語プラットフォームの開発体制

表 3-5-1 技術用日本語プラットフォーム開発スケジュール

年度	H19 年度	H20 年度	H21 年度	H22 年度	H23 年度	H24 年度
フェーズ	スタディ		開発			運用
A. 技術用日本語共通基盤仕様 A-1 言い換え規則（暫定版） A-2 （CDL+CDL 化+線状化は ISeC と Japio が H19 年度開発）	第 0 次 ---------→ (------→)	第 1 次 ----------→	改良拡張 ----------→	改良拡張 ----------→		
B. 技術用日本語プラットフォーム システム B-1 技術用日本語オーサリングシステム B-2 技術用日本語言語知識集合知サーバ	基本検討 ---------→	基本検討 ----------→	ﾌﾟﾛﾄﾀｲﾌﾟ ----------→	運用システム ----------→	改良拡張 ----------→	・技術用日本語コンソーシアム
C. 技術用日本語アプリケーション システム C-1 技術用日本語日英機械翻訳システム C-2 技術用日本語文書検索システム	基本検討 ---------→	基本検討 ----------→	ﾌﾟﾛﾄﾀｲﾌﾟ ----------→	運用システム ----------→	改良拡張 ----------→	・知の生産環境の整備
D. モデル運用サービス D-1 特許ワンストップサービス D-2 先進的知識マネージメントサービス				プロトタイプ 運用形態開発 ----------→	実運用形態開発 ----------→	

４　スタディの今後の課題及び展開

長期的には、本スタディにおいてまとめた、スタディフェーズ、開発フェーズ及び運用フェーズからなる開発スケジュールに沿って、仕様の改良・拡張、各種システムの開発・改良・拡張、モデル運用サービスへの展開といった大きな課題に取り組んでいかなければならないが、開発フェーズの初年度に予定しているプロトタイプシステムの開発に進む前に、スタディフェーズにおいて取り組んでおくべき、多くの課題が残された。それは、以下のとおりである。

４－１　実験ツールによる技術検証

本年度のスタディでは、技術用日本語の言い換え規則（暫定）を策定し、そのうちの機械翻訳用規則に基づいて手作業で特許文献中の日本語を技術用日本語に変換し、原文と変換された文の両方の機械翻訳結果を比較評価することを通じて、技術用日本語によって機械翻訳の訳質が向上することを確認した。また、技術用日本語は機械処理に適応できる簡潔さを持つことも確認した。

しかし、技術用日本語プラットフォームの開発には、手作業から機械処理へと、その検証のステップを進めていくことが不可欠である。

具体的には、技術用日本語プラットフォーム技術を検証するために、技術用日本語の処理をコンピュータ上で行うことができる実験ツールを作成し、機械翻訳及び知識管理を想定して評価実験し、技術用日本語そのものの優位性と、どのように機械処理を実現すればよいかについての確認を行う必要がある。

技術用日本語プラットフォームのエンジンは、日本語を解析して規則データベースを参照しながら技術用日本語を自動的、あるいは半自動で生成する。このプラットフォームの技術的評価・検討のために、技術用日本語を機械支援で生成する実験ツールを構築する。実験ツールを構築する基本的な作業として、まず、制限規則が技術文献を広くカバーできるように、語彙や文法を制限する規則データを揃え、また、広く収集した規則を計算機の中に埋め込むために、規則データのフォーマットを設定する。フォーマットされた規則集合は規則データベースとなり、実験ツールに組み込まれる。

実験の手順としては、これらの規則が組込まれた実験ツールに、特許文献、論文、マニュアル等の文章を入力し、生成された文書と原文について、日本語文章そのものの比較評価及びそれぞれの機械翻訳英文の比較評価を行う。この評価結果を規則データベースにフィードバックする。この作業を繰り返すことによって、規則が見直され、技術用日本語における曖昧性や解りにくさが解消していくことになる。

４－２　知識検索及び要約生成への技術用日本語の応用に係る考察

技術用日本語を解析すると、XML におけるツリー構造（ネスト構造）と RDF における関係構造の両方を含んだハイパーグラフ構造になる。ここから、単語間の関係知識や、意味構造を使った要約知識などが抽出できることが期待できる。この点について考察を行い、その可能性の評価が必要である。

４－３　プロジェクト計画の策定調査

今後の展開として、技術用日本語プラットフォームの開発を、スタディフェーズから開発フェーズに円滑に移行させるための準備を行っていく必要がある。具体的には、特許文献、論文、及びマニュアルの各々の文書作成に対して、技術用日本語の導入が効率的かつ的確に行われるための環境整備を念頭に置きつつ、技術用日本語プラットフォーム開発の次期プロジェクトを関係各機関に提案できるようにするために、その開発計画を立案する。

システム開発　　１９－Ｆ－１２

経済活性化のための
技術用日本語プラットフォームの開発
に関するフィージビリティスタディ
（ 要　旨 ）

平成２０年３月

作　　成　　財団法人機械システム振興協会
東京都港区三田一丁目４番２８号
TEL 03-3454-1311

委託先　　財団法人日本特許情報機構
東京都江東区東陽四丁目１番７号
TEL 03-3615-5511