システム開発
２０－Ｆ－５

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２１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ＪＫＡ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２１年３月

財団法人 機械システム振興協会

はじめに

東西冷戦の終焉後間もない 1993 年、CERN（セルン、欧州原子核研究機構）のWWW（ワールド・ワイド・ウェブ）の無料開放を契機に、爆発的に普及したインターネットは、またたく間に世界中の人々をつなぎ、情報の共有と拡大を促進しました。こうしたインターネットの普及は、動き始めた経済のグローバル化を一気に加速させ、また、情報伝達のベースである言語の標準化への対応を促してきました。そして、グローバルな経済社会での情報は、事実上の世界標準言語といえる英語をベースに動いています。米欧先進国は、このような英語の標準化への対応として、いわゆる制限英語に例示されるように、英語を他の言語に翻訳しやすくするための取り組みを多く行ってきました。

一方、日本語によるコミュニケーションをベースとする我が国においても、過去には、日本語の規格化ないし標準化が提唱されたことがあり、また、実践的な規定集等が関連団体や企業で集積されてきました。しかしながら、米欧のような業界横断的な取り組みは未だみられず、本格的な普及には至っておりません。そして、我が国産業界では、世界に向けて情報を発信する際に、翻訳という負荷を必然的に負うことになるため、この翻訳コストを削減するために、機械翻訳の品質と効率をいかに高めるかが大きな課題となっています。

他方、ICT(Information and Communication Technology)の進化は、留まることを知らず、現在では、ネットワーク上に存在するコンピュータ資源をサービスとして提供するクラウドコンピューティングが、サービスの利用者と提供者の双方にとってのコストの削減と人的資源活用の効率化を達成するための技術として注目されています。こうした状況のもと、我が国においても、コストを削減しつつ産業活性化を図るための ICT 革新に向けた具体的取り組みとして、クラウドコンピューティングが推進されつつあります。ネットワーク上にある様々な情報は、その内容が正確に伝達され、また、機械処理し易い形で伝達されることにより、その効率的な活用が期待されます。このような点を踏まえますと、伝達すべき情報を記述する日本語を標準化し、機械処理し易くすることが、我が国の ICT 革新の成功に欠かせないものであると言えます。

本スタディの報告書は、このような現状認識のもと、我が国経済の活性化と国際競争力の強化に資する「技術用日本語プラットフォーム」の開発の必要性を産業界に初めて示した平成 19 年度のスタディをベースに、「技術用日本語プラットフォーム開発計画」をとりまとめたものです。このプラットフォーム開発計画が、機械処理にも適した標準的な日本語である「技術用日本語」を産業界に普及するとともに、日本型の ICT 革新を加速することに貢献し、日本企業の国際競争力強化の一助となれば幸いであります。

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感謝いたしますとともに、本スタディにご協力をいただいた関係各位に心から敬意を表する次第であります。

平成２１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
守屋　敏道

目　次

１　スタディの目的

情報化社会が進展するなか、日本の知識基盤となるのはいうまでもなく日本語である。一方、グローバル社会における知識基盤に対しては、英語が大きな役割を担っている。英語圏における長年の努力によって、英語はそのような役割を担えるようになったのである。例えば、1978 年のカーター米国大統領令による奨励によって、Plain English が法律、金融、公共の場面に広く浸透し利用されている。

更に、産業用文書や商業用文書の作成に用いられる英語に関しては、あるがままの英語を使用するのではなく、国際言語としてノンネイティブを含む多民族が意思疎通できる正確で分かり易い英語とするための実践がなされてきた。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などの多国籍企業が、英文マニュアルなどを記述するための制限英語をそれぞれが定め活用している。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）を定めている。Web コンテンツに対しても、HTML や XML による文書の構造化と一体となって、標準的な英語のライティングスタイルがほぼ固まり、それを利用した高度な情報サービスへの移行が始まっている。

他方、わが国における日本語への取り組みについては、確かに、ここ 10 年ぐらいは、客観的に分かり易く日本語を用いるための努力が各方面で行われてきた。

用語の規定を中心にした産業界各分野における努力、日本語能力試験や日本留学試験に代表される外国人への日本語教育分野における努力、日本語テクニカルライティングや日本語テクニカルコミュニケーションに関する関連団体の努力、日本語ワープロや機械翻訳などの日本語処理技術に関わる分野における努力、これらの努力が技術用日本語に向けての多くの知見を蓄積してくれた。しかしながら、今のところこれらの蓄積は、個別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ことが緊急の課題となっている。

技術用日本語のプラットフォームシステムにとっては、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、プラットフォーム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、相当の規模と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みが整備され、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。

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ねばならない。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、日本語に関する知見の蓄積をプラットフォームシステム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。

先に指摘したように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ギ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。すでに実用化され利用されている制限英語は、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留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(文書の作成と編集)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。

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に用いる日本語をインターネット時代に適応した日本語に変革すること、すなわち、日本語の情報力を格段に強化することが技術用日本語プラットフォームの目標である。この目標を実現するために平成 19 年度後期におけるスタディでは、開発計画基本案を策定し、技術動向調査に基づいて計画全体のフィージビリティを実証し、記述・評価実験に基づいて技術用日本語共通基盤仕様（第 0 版）のフィージビリティを実証した。

[開発計画]

1. 開発課題
   1.1 技術用日本語共通基盤仕様
   1.2 技術用日本語プラットフォームシステム
      1.2.1 技術用日本語オーサリングシステム
      1.2.2 技術用日本語言語知識集合知サーバ
   1.3 技術用日本語アプリケーションシステム

本スタディは、技術用日本語プラットフォーム開発計画実施案の最終案を策定することを目的としている。その裏付け資料を得るために、技術用日本語プラットフォームでは、オーサリング用実験ソフト及び言語知識集合知サーバ用実験データを開発し、コンピュータを用いた実証を行うものである。更に、入力テキストの技術用日本語変換のために独自の日本語表層概念記述言語 CDL.jpn.sf の仕様を策定するとともに、テキスト形式以外にグラフ形式のオーサリングについても併せて実験するものである。

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、実証実験ソフトに関わる作業（機械翻訳系、概念記述言語系）の実施は再委託した。

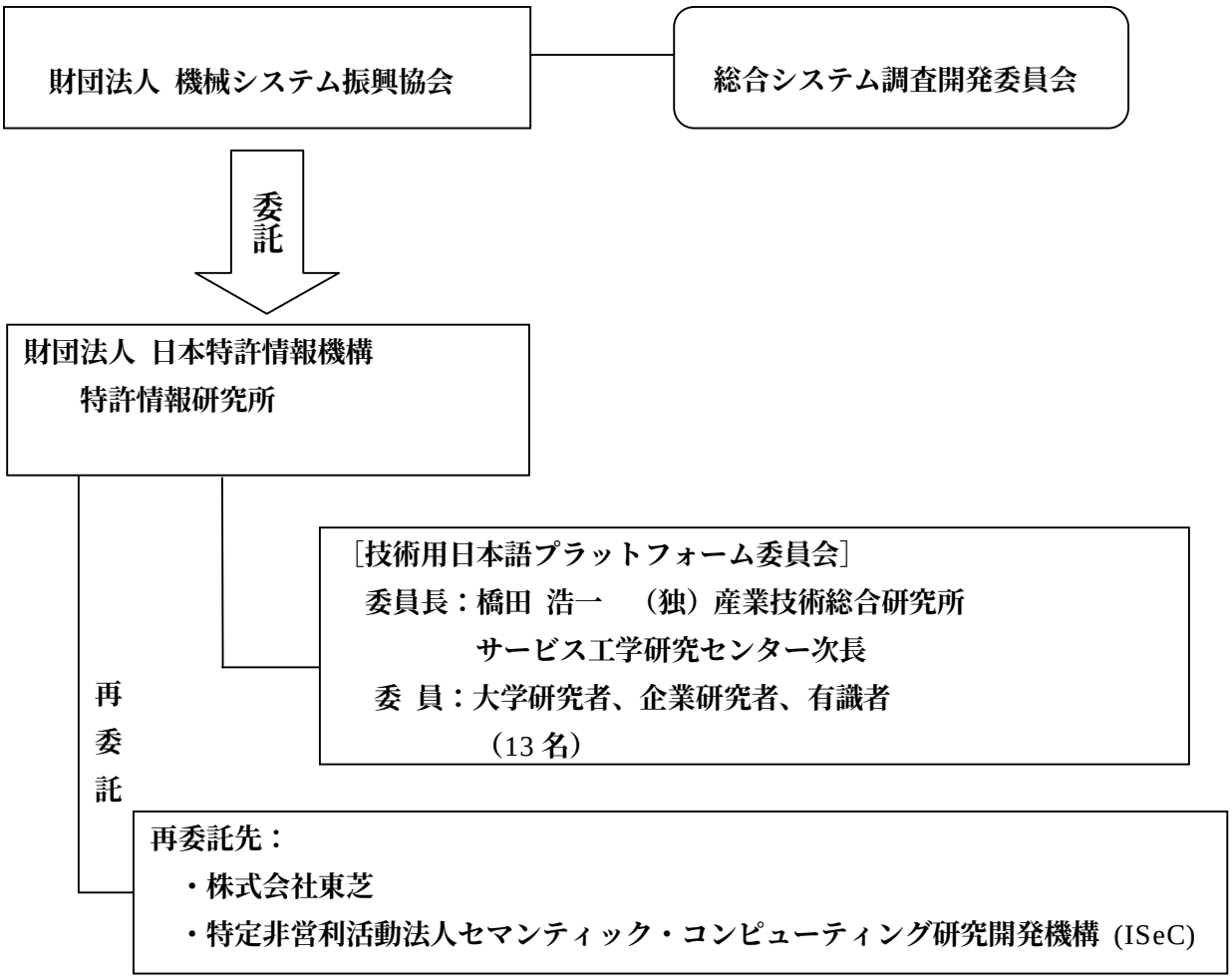


図１：委託事業実施体制

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 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 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 デジタルものづくり研究センター 招聘研究員	志　村　洋　文
委　員	東北大学大学院 工学研究科 教授	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	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 サービス工学研究センター　次長	橋　田　浩　一
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	いすゞ自動車株式会社 知的財産部　知的財産第一グループ（JIPA 特許第１副委員長）	長　池　将　幸
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　第１特許室　主幹	服　部　博　生
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	富士通株式会社 法務・知的財産権本部　特許部　担当部長　弁理士	横　山　淳　一
委　員	(株)日立製作所 システム開発研究所 uValue イノベーション研究部　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
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	守　屋　敏　道
事務局	(財)日本特許情報機構 特許情報研究所　顧問	横　井　俊　夫
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	渡　邊　豊　英
事務局	(財)日本特許情報機構 特許情報研究所　研究企画課長	大　塩　只　明
事務局	(財)日本特許情報機構 特許情報研究所　研究管理課長	塙　　金　治
事務局	（NPO）セマンティックコンピューティング研究開発機構 理事	安　原　　宏

３　スタディ成果の要約

本スタディは、平成 20 年 4 月から平成 21 年 2 月にかけて、以下の３つのテーマについて実施した。

１．開発計画実施案の策定
２．技術用日本語オーサリングシステム用実験ソフトの開発と動作・評価実験
３．技術用日本語言語知識集合知サーバ用実験データの開発と動作・評価実験

以下に、各テーマの成果の概要を述べる。

３－１　開発計画実施案の策定

平成 19 年度スタディで策定した 3 年間の計画案をベースに、本スタディ終了後の開発フェーズを想定した技術用日本語プラットフォームの開発計画実施案（最終案）を策定した。

開発課題は、①技術用日本語プラットフォームシステム、②プラットフォームアプリケーション、③プラットフォーム上のモデル運用サービスの三階層に分けた。

プラットフォームシステムでは、技術用日本語オーサリングシステム及び技術用日本語集合知サーバの計画を策定した。策定にあたっては、３－２「技術用日本語オーサリングシステム用実験ソフトの開発と動作・評価実験」及び３－３「技術用日本語言語知識集合知サーバ用実験データの開発と動作・評価実験」と連携して作業を進めた。

プラットフォームアプリケーションでは、特許文書の機械翻訳、文書検索への応用の計画策定に重点的に取り組んだ。

モデル運用サービスでは、知財関連のステークホルダー（技術者、特許庁、知財サービスベンダー、知財サービスプロバイダーなど）の視点を取り込んだ知財サイクル・ワンストップサービス及び、先進的知識マネジメントサービスの計画策定を実施した。なお技術用日本語共通基盤仕様（第 1 版）は、Japio 内の特許版産業日本語委員会で策定した。

計画策定にあたり、最近コンピュータ業界で注目されているクラウドコンピューティングは本テーマにとっても基盤システムであると判断し、開発課題に積極的に取り入れた。

開発計画実施案は、以下の章立てでまとめた。

1　開発計画
1.1　目標と課題
1.2　目標設定の先進性と妥当性
1.3　開発スケジュール
1.4　開発体制

2　開発課題

2.1　技術用日本語共通基盤仕様（第１版）

2.2　プラットフォームシステム

2.3　プラットフォームアプリケーション

2.4　モデル運用サービス

３－１－１　開発計画

（1）目標と課題

「日本の産業活動の基盤となる産業情報サイクルの活性化のために技術用日本語プラットフォームを開発するとともに、このプラットフォームを基幹技術としたモデルサービスの実運用化を提案し、次世代 IT 環境として推進されつつあるクラウドコンピューティングの本格的な普及に備え、日本版クラウドコンピューティングの普及に向けた環境整備の一翼を担う。」を目標として掲げた。

技術用日本語は、旧来からの閉じた日本語から以下の要件を満たす開かれた日本語へと脱皮することが求められることになる。

①　世界に開かれた日本語

②　分野間に開かれた日本語

③　人間からコンピュータへと開かれた日本語

一方、クラウドコンピューティングは、技術用日本語プラットフォームをシステムとして実装し運用していくための基本的なシステムアーキテクチャとしても採用し得るものである。IT はその時代に応じた、コンピューティング資源の最適活用･有効活用を図るという点で、常に一つの方向性をもって進歩してきている。

当然ながらその方向性は、アプリケーションサービスに対するニーズの動向に対応している。その動向はニーズの多様化と人間中心指向であろう。つまり、利用者の多様化に伴ってニーズが個別化・多様化するとともに、人間にわかりやすいサービスや扱いやすいインタフェースへのニーズが高まっており、それに応えるためにコンピューティング資源をより柔軟に活用できる技術が求められているのである。

人間にわかりやすい個別的で多様なサービスや扱いやすいインタフェースを提供するには、情報システムが人間と意味を共有する必要がある。そのためには情報コンテンツの意味がコンピュータにも「理解」できるように構造化されていなければならない。とりわけ、情報コンテンツのうち知的生産活動の最重要基盤である自然言語の意味内容をコンピュータにも扱いやすくするための構造化に関する標準が必須である。あらゆる自然言語に共通の国際標準と並んで個別言語の特徴に応じた規格が必要であり、日本語の意味構造に関する規格の策定とそれを活用するための基盤の構築がわが国の責務であることはいうまでも

ない。

クラウドコンピューティングをベースに技術用日本語プラットフォームシステムを開発し、このプラットフォームを基幹技術としたモデルサービスの実運用化を目指すことは、例えば、特許明細書などや企業の営業機密文書などを取り扱うことによる強固なセキュリティ対策の実現など、現状のクラウドコンピューティング関連技術における技術課題を刺激し、その開発を促すなど、日本におけるクラウドコンピューティングの発展にも寄与するものである。

すなわち、技術用日本語の産業界への導入は、日本におけるクラウドコンピューティングの本格的な普及のための環境整備の一翼を担い、かつ、その発展にも貢献するという相乗効果が期待されるのである（図 3-1-1 参照）。

策定した課題を表 3-1-1 に示す。平成 19 年度のスタディに対して項番 D「モデル運用サービス」を追加した。これで、共通基盤となる言語から運用サービスまでを含めた総合的なシステムの開発計画となった。

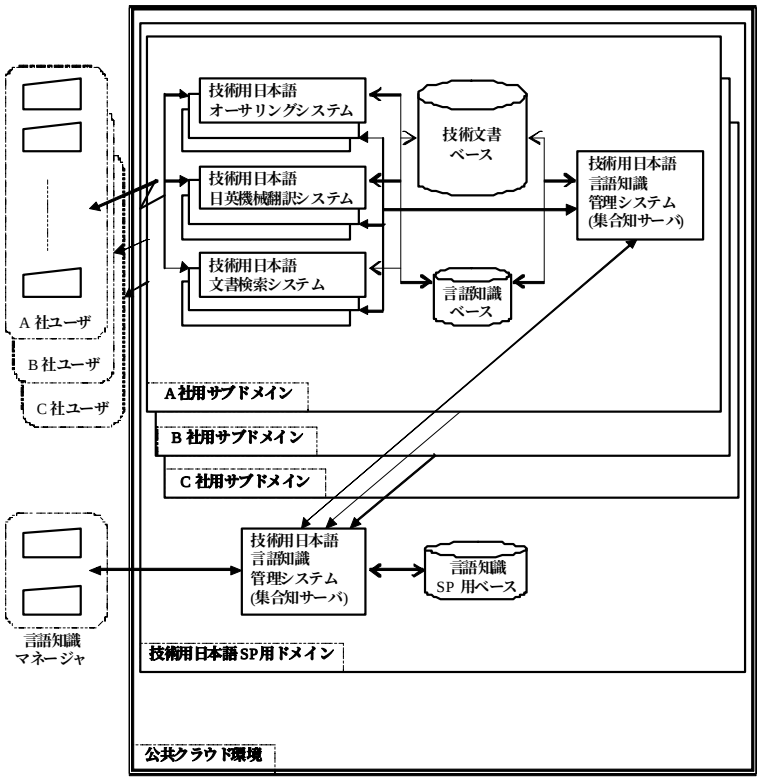


図 3-1-1 公共クラウド上での技術用日本語プラットフォーム実現形態

表 3-1-1　技術用日本語プラットフォームの開発課題と開発スケジュール

<table>
<tr><th>開発課題</th><th>1 年度</th><th>2 年度</th><th>3 年度</th><th>4 年度～</th></tr>
<tr><td>A.　技術用日本語共通基盤仕様</td><td>改良拡張</td><td>改良拡張</td><td></td><td></td></tr>
<tr><td>B.　プラットフォームシステム</td><td rowspan="3">プロトタイプ</td><td rowspan="3">運用システム</td><td rowspan="3">改良拡張</td><td></td></tr>
<tr><td>b-1 技術用日本語オーサリングシステム</td><td></td></tr>
<tr><td>b-2 技術用日本語言語知識集合知サーバ</td><td></td></tr>
<tr><td>C.　プラットフォームアプリケーション</td><td rowspan="3">プロトタイプ</td><td rowspan="3">運用システム</td><td colspan="2" rowspan="3">改良拡張</td></tr>
<tr><td>c-1 技術用日本語日英機械翻訳システム</td></tr>
<tr><td>c-2 技術用日本語文書検索システム</td></tr>
<tr><td>D.　モデル運用サービス</td><td rowspan="3">プロトタイプ1</td><td rowspan="3">プロトタイプ2</td><td rowspan="3">運用システム</td><td rowspan="3">運用サービス</td></tr>
<tr><td>d-1 知財サイクル・ワンストップサービス</td></tr>
<tr><td>d-2 先進的知識マネジメントサービス</td></tr>
</table>

（2）目標設定の先進性と妥当性

産業技術文書において、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。このいまだな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、技術的な観点からの要点を列挙する。

① 言語学の知見及び言語処理の経験やノウハウの十分な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。
② 機械翻訳、日本語ワープロ、校正支援システムなどの実績により適切なシステム実現技術に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。
③ テキスト処理、図解・数式・プログラム言語、メディア処理などの広範な応用への連携を支える概念記述言語 CDL 技術の利用

（3） 開発スケジュール

開発スケジュールは、大きくスタディフェーズ(平成 19 年度から平成 20 年度)、開発フェーズ（平成 21 年度から平成 23 年度)、運用・改良・拡張フェーズ(平成 24 年度以降)の 3 段階とした。スタディフェーズは、本スタディを含めた 1 年半のフィージビリティスタディである。開発フェーズは、実運用システムを開発する期間であり、プラットフォームシステムの開発期間として 3 年間を、プラットフォームアプリケーションの開発期間としては 3～5 年程度を見込む。開発フェーズにおける各開発課題の開発状況に応じ、順次、運用・改良・拡張フェーズに移行することとした（表 3-1-1 参照)。

（4） 開発体制

技術用日本語プラットフォームの開発を成功させ、その先の運用フェーズへの移行をスムーズなものとし、技術用日本語を我が国産業界へ速やかに普及させていくためには、関係省庁を始め、IT ベンダー、大学や研究機関、ユーザ企業・団体、サービス企業・団体といった関係者・関係団体の連携が不可欠である。具体的には、言語処理、機械翻訳、検索、通信・ネットワーク関連技術といった、技術用日本語プラットフォーム開発に必要な各種要素技術に強みを持つ IT ベンダー企業や研究機関からなるプラットフォームシステム・サービスの開発体制を組織するとともに、特許情報を中心とした産業技術情報のユーザ企業や、大学などを中心とした研究機関をも集結した開発体制を構築する必要があると考えた。

そのために、開発フェーズのスタートにあたっては、技術用日本語コンソーシアムと技術用日本語フォーラムを立ち上げるとともに、適切な機関が中心となりプロジェクトマネジメントを行う体制を構築する。技術用日本語コンソーシアムは、ユーザ企業・団体やサービス企業・団体を中心に、サービス事業の立ち上げや普及を目的に活動する。技術用日本語フォーラムは、大学や研究機関を中心に、技術用日本語に対する学術的な裏付けやプラットフォームシステムを研究開発ツールとして活用する新技術研究開発の促進などの活動を行う（図 3-1-2 参照）。

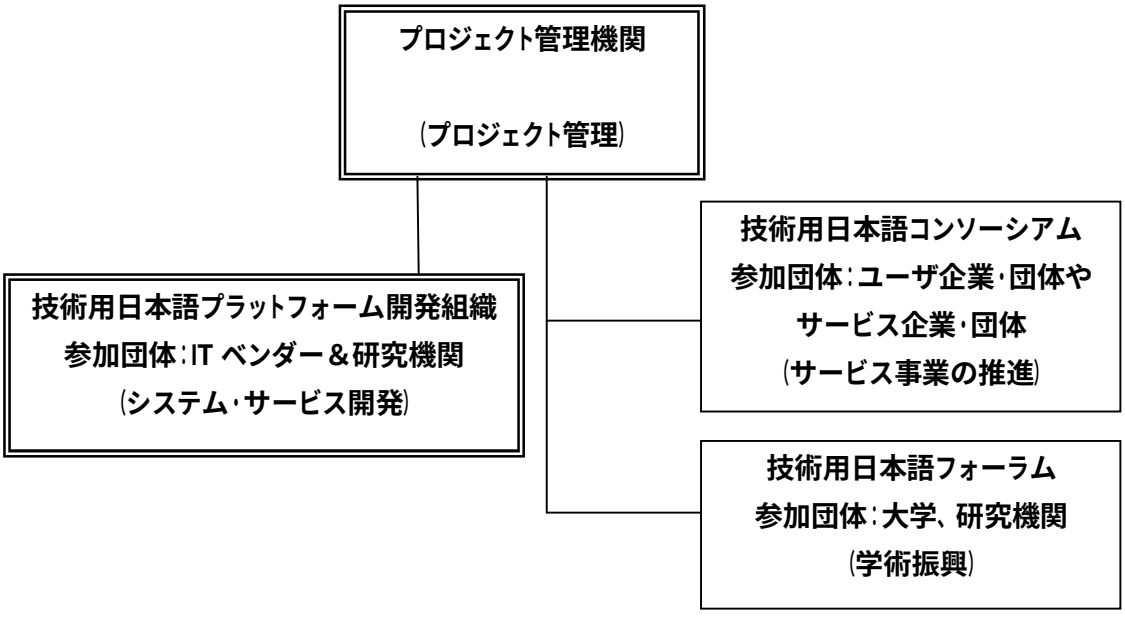


図 3-1-2 開発計画の実施体制

３－１－２　開発課題

開発課題は、以下の 4 項目からなる(表 3-1-1 参照)。

（１）技術用日本語共通基盤仕様（第 1 版）

（２）プラットフォームシステム

（３）プラットフォームアプリケーション

（４）モデル運用サービス

これらの課題に対して開発フェーズの計画内容を明確にするために、システム概要、システム構成と利用イメージ、開発項目、開発線表について各々まとめた。以下その骨子をまとめる。

（１）技術用日本語共通基盤仕様（第１版）

技術用日本語共通基盤仕様は、Japio 内の特許版産業日本語委員会で策定した。そこから、現在 4 つのファミリー言語の仕様を作っている。特許文書を対象にした特許版技術用日本語、機械翻訳精度を高めるための日英機械翻訳技術用日本語、文レベル検索を指向する文書検索技術用日本語、テキストを図式でオーサリングするための図式技術用日本語である。

（２）プラットフォームシステム

技術用日本語のオーサリングシステムと言語知識集合知サーバからなる。オーサリングシステムはカナ漢字変換の日本語ワードプロセッサーの次の技術である言い換え変換プロセッサーである。言語知識集合知サーバは、電子化辞書の形態をとるが言い換え規則、機械翻訳知識、検索知識のための言語資源である。このレイヤは、アプリケーションとサービスから呼び出される。

（３）プラットフォームアプリケーション

技術用日本語向き日英機械翻訳システム及び技術用日本語向き文書検索システムを主要アプリケーションとした。日英機械翻訳システムは、技術用日本語を入力文とすることで訳文の精度向上を目指す。文書検索システムは、単語からセンテンスへの検索パラダイムのシフトを行う。

（４）モデル運用サービス

知財サイクル・ワンストップサービスと先進的知識マネジメントサービスを主要サービスとした。知財サイクル・ワンストップサービスは、特許版技術用日本語を利用して知財サイクルで使用する特許文書を効率よく機械支援するものである。先進的知識マネジメントは、企業内、企業間における知識共有型の高度な産業サービスを提供する。

上記の策定した各開発課題のシステム構成図と開発線表を図 3-1-3～図 3-1-8、表 3-1-2～表 3-1-7 に掲げる。これらは、各システム構成図に記載されているように、アプリケーションはプラットフォームシステムを呼び出し、モデルサービスはアプリケーションを呼び出す。特に言語知識集合知サーバは全てのシステムから利用される共通の言語資源である。クラウドコンピューティングの環境はこのような相互依存した融合システムを構築することに適していると考えている。

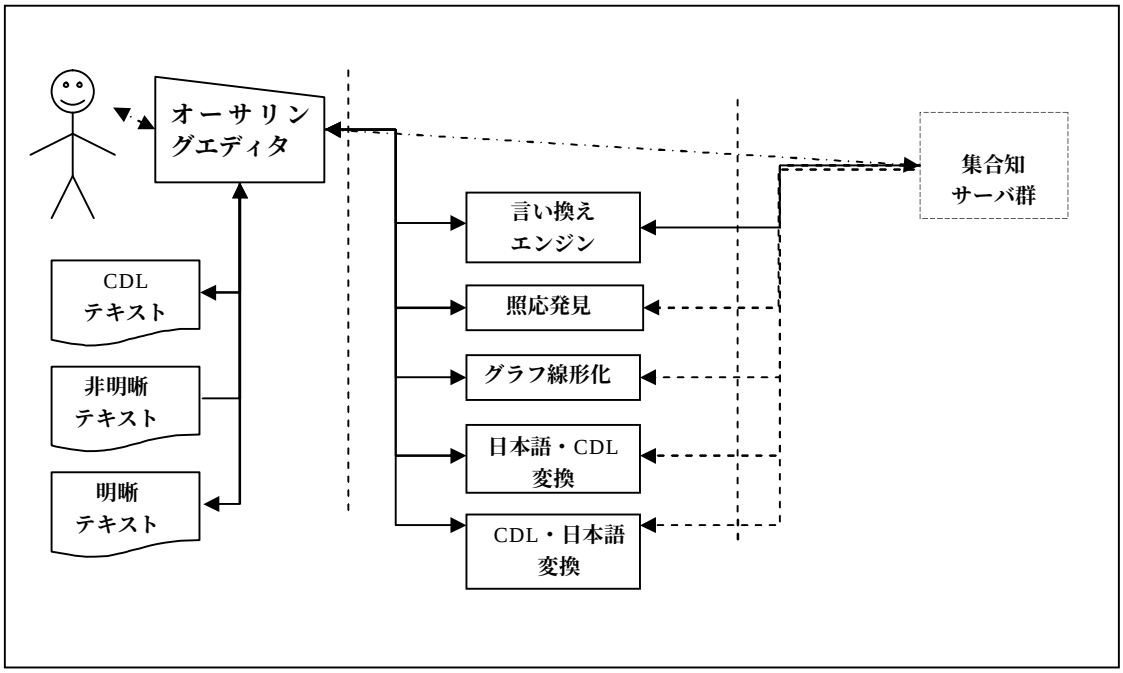


図 3-1-3 技術用日本語オーサリングシステムの構成

表 3-1-2 技術用日本語オーサリングシステムの開発線表

開発項目			第１年度	第２年度	第３年度
A 技術用日本語オーサリングエディタ	設計	ユースケース設計、モジュールインタフェース設計、UI 設計、クラス設計	→		
	実装	基本エディタ機能	プロトタイプ	完了	
		① 表現の明晰化	文章長/ターム	正規表現	構文パターン
		② 学習機能	〔設計〕	プロトタイプ	
		③ 照応関係の明示化	〔設計〕	プロトタイプ	精度向上
		④ セマンティックオーサリング	〔組込設計〕	組込プロト	完了
		⑤ 既存テキストのインポート	Text/XML		ワープロなど
		⑥ 文書エクスポート	Text/XML		ワープロなど
		⑦ モジュール管理	プロトタイプ	β	完了
B. 言い換えエンジン			ターム	正規表現	構文パターン
C. 照応発見モジュール			設計	プロトタイプ	精度向上
D. グラフ線形化モジュール			設計	プロトタイプ	完了
E. 諸文書形式 CDL 変換モジュール			文書構造	日本語解析	精度向上
F. CDL 諸文書形式変換モジュール			Text/XML		ワープロなど

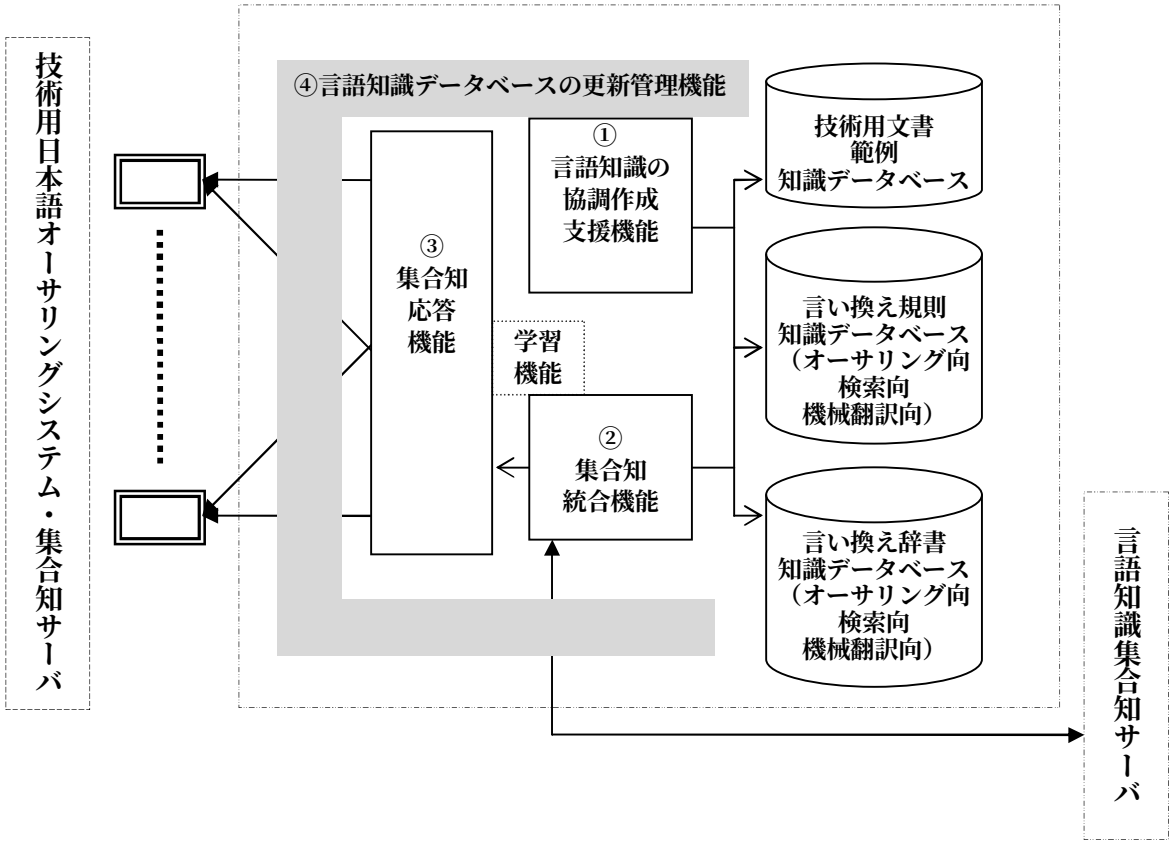


図 3-1-4 技術用日本語言語知識集合知サーバの構成

表 3-1-3 技術用日本語言語知識集合知サーバの開発線表

開発項目		第1年度	第2年度	第3年度
設計	アプリケーションシステムUI	→		
	言語知識集合知サーバが連動する機能		→	→
	言語知識の協調作成支援機能	→	→	→
実装	A 言語知識の協調作成支援機能			
	・言い換え規則の効率的な抽出機能	文章長/ターム	正規表現	構文パタン
	・未知語候補の効率的な抽出機能	プロトタイプ	精度向上	精度向上
	・同義語、類義語、関連語候補の効率的な抽出機能	プロトタイプ	精度向上	精度向上
	・対訳候補の抽出機能	プロトタイプ	精度向上	精度向上
	・既存辞書の集合知変換機能	プロトタイプ	精度向上	精度向上
	B 集合知統合機能			
	・集合知のキャッシュ機能		設計	プロトタイプ
	・言語知識集合知サーバの整理統合機能		設計	プロトタイプ
	・言語知識集合知サーバの多重化機能			設計
	・学習機能		設計	プロトタイプ
	C 集合知応答機能			
	・集合知応答機能	プロトタイプ	精度向上	精度向上
	・認証・課金機能		設計	プロトタイプ
	D 言語知識データベースの更新管理機能			
	・言語知識データベースの更新管理機能	プロトタイプ	精度向上	精度向上
	・集合知公開管理システム		設計	プロトタイプ
	E 言語知識の開発(特許ドメイン)		運用	運用・リリース

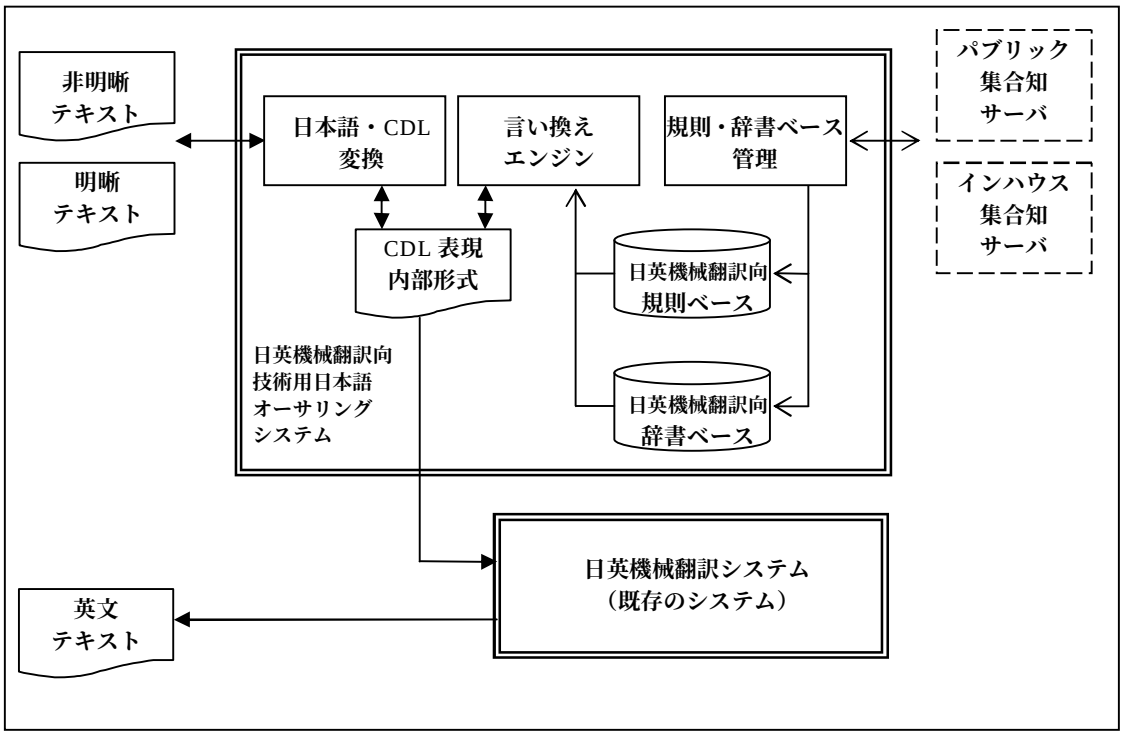


図 3-1-5 技術用日本語日英機械翻訳システムの構成

表 3-1-4 技術用日本語日英機械翻訳システムの開発線表

開発項目	第 1 年度	第 2 年度	第 3 年度
A.日本語解析エンジン	設計、基本機能開発	機能改良	―
B. 日本語生成エンジン	設計、基本機能開発	機能改良	―
C. 技術用日本語オーサリングシステム・インタフェース	設計、基本インタフェース開発	インタフェース改良	インタフェース改良
D.日英トランスファエンジン	設計、基本機能開発	機能改良	―
E. 英語生成エンジン	基本機能開発	―	―
F. 日英機械翻訳向き言い換えエンジン	設計、基本機能開発	機能改良	実例に応じた改良
G. 日英機械翻訳向き言い換え規則ベース	設計、基本機能開発、知識開発	知識開発	知識改良拡張
H. 日英機械翻訳向き言い換え辞書ベース	設計、基本機能開発、知識開発	知識開発	知識改良拡張
技術用日本語日英機械翻訳システム 全体	プロトタイプシステム	実運用システム	実用的な運用システム

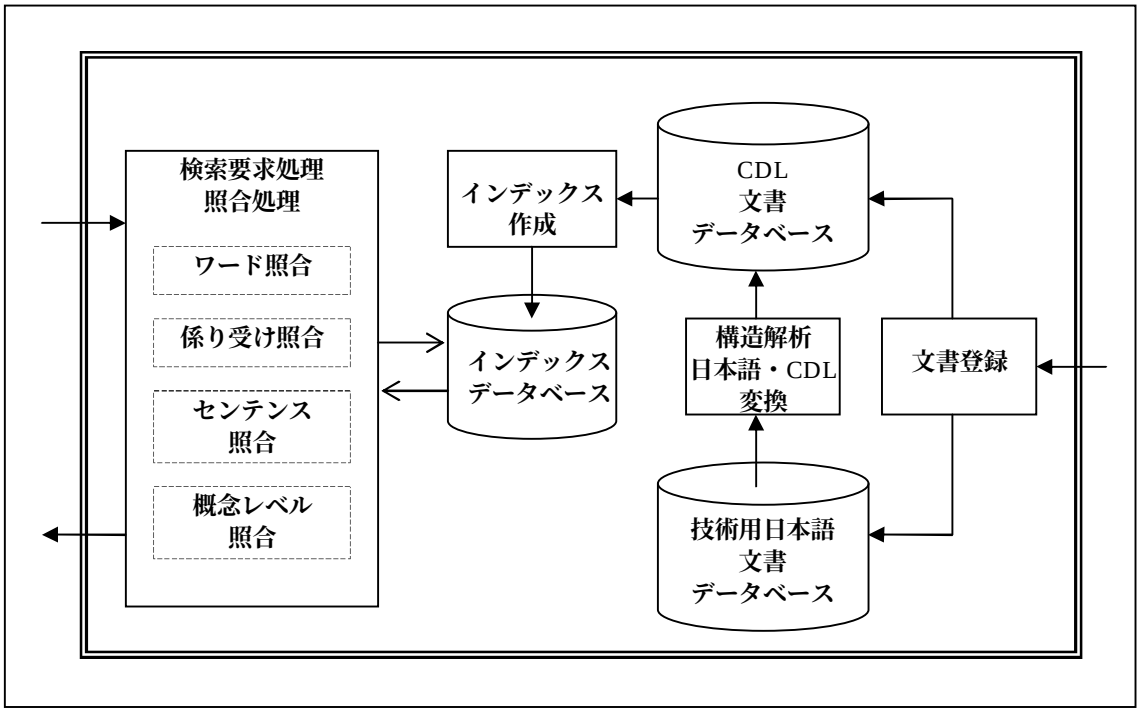


図 3-1-6 技術用日本語文書検索システムの構成

表 3-1-5 技術用日本語文書検索システムの開発線表

開発項目	第１年度		第２年度	第３年度	第４年度~
（開発フェーズ）	Ⅰ		Ⅱ		Ⅲ
A. 文書検索精度向上のための技術用日本語の仕様策定	→				
B. 技術用日本語を前提としたフレーズや文の解析方式					
基礎検討、プロト検証		→	→	→	
設計・開発					→
C. 柔軟な類似性判定方式					
基礎検討、プロト検証		→	→	→	
設計・開発					→
D. システム実装方式					
d-1 変換精度検証と向上策		→	→		
d-2~3 インデクシング／UI			→	→	→

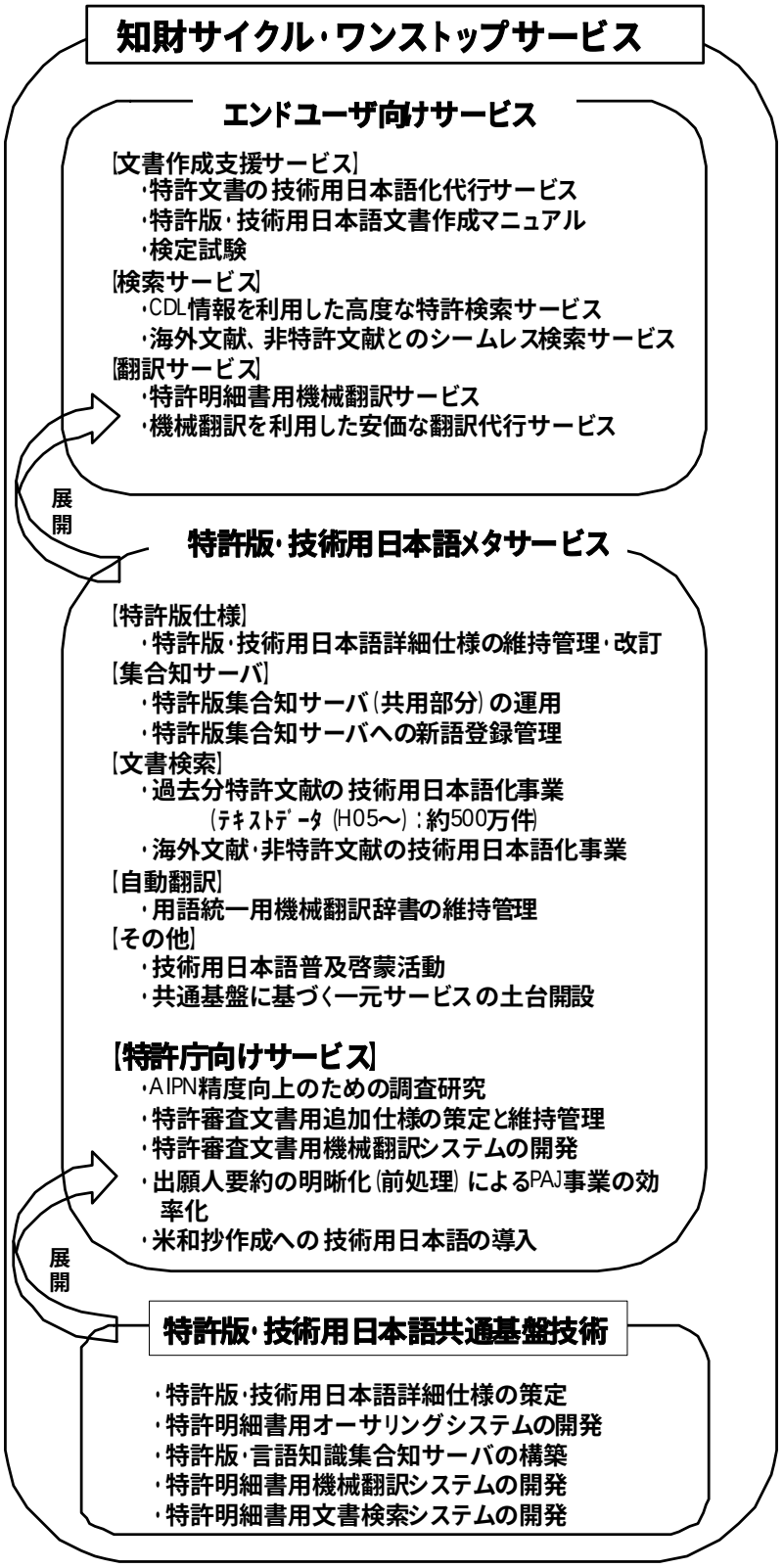


図 3-1-7 知財サイクル・ワンストップサービスの構成

表 3-1-6 知財サイクル・ワンストップサービスの開発線表

開発項目	第１年度	第２年度	第３年度
A-1.特許版・技術用日本語詳細仕様	→		
A-2.特許明細書用オーサリングシステム	→	→	→
A-3.特許版言語知識集合知サーバ	→	→	→
A-4.特許明細書用日英機械翻訳システム	→	→	→
A-5.特許明細書用文書検索システム	→	→	→
B.特許版・技術用日本語メタサービス	⇢	⇢	⇢
C.エンドユーザ向けサービス	⇢	⇢	⇢
D.特許庁向けサービス	⇢	⇢	⇢
E.クラウドコンピューティングプラットフォーム	→	→	→

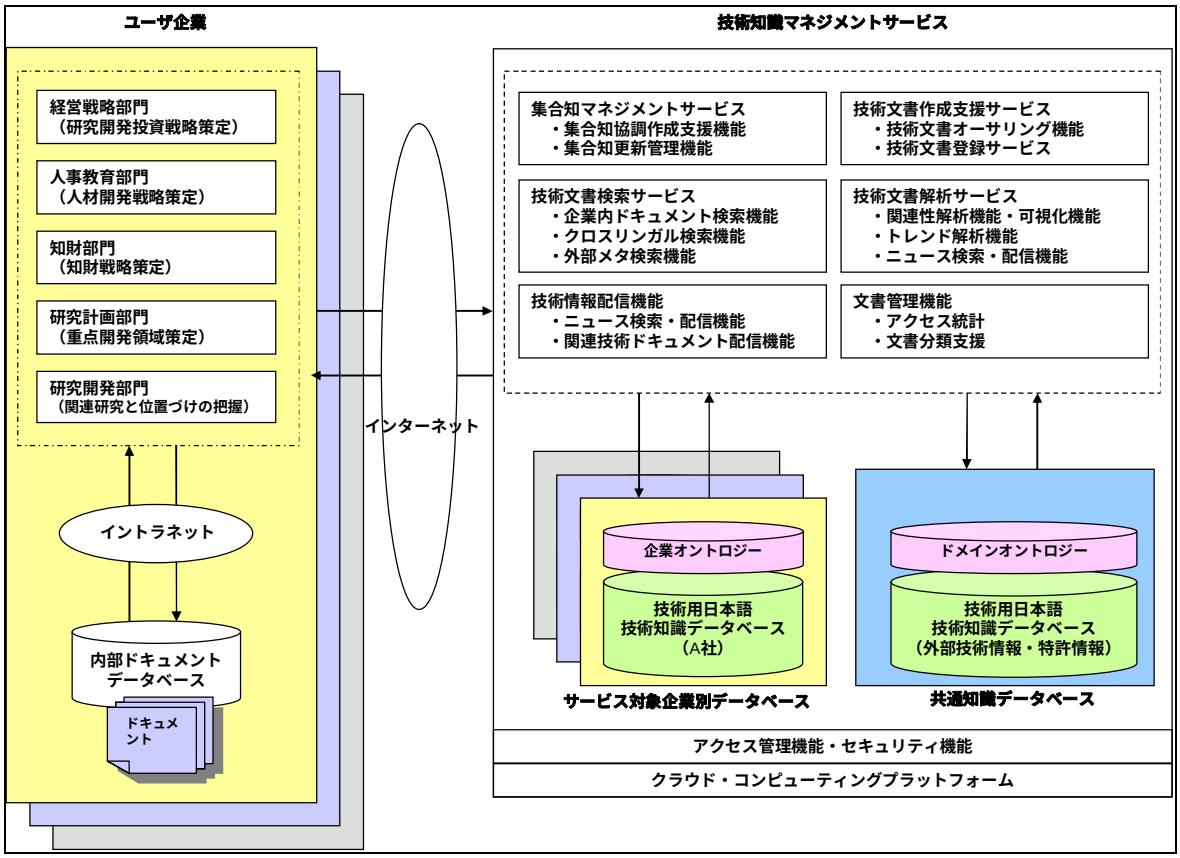


図 3-1-8 先進的知識マネジメントサービスの構成

表 3-1-7 先進的知識マネジメントサービスの開発線表

開発項目	第１年度	第２年度	第３年度
A-1.集合知マネジメントサービス	→	→	→
A-2.技術文書作成支援サービス	→	→	→
A-3.技術文書検索サービス		→	→
A-4.技術文書解析サービス		→	→
A-5.技術情報配信機能		→	→
A-6.文書管理機能		→	→
B.クラウドコンピューティングプラットフォーム	→	→	→
C.知財サイクルワンストップサービスとの連携			→

３－２　技術用日本語オーサリングシステム用実験ソフトの開発と動作・評価実験

非明晰な日本語を明晰な技術用日本語に変換するオーサリングシステムのスタディとして、実験ソフトを用いて非明晰な日本語を構文解析し、CDL 化し、CDL レベルで言い換えを実行し、CDL から日本語生成を行い、機能面、操作面、性能面などから評価を行った。ここで CDL とは、ISeC で開発された概念記述言語でハイパー構造を持つネットワーク型言語である。

以下に実験ソフトの概要を記す。

３－２－１　開発仕様

（1）システム構成

実験ソフトは、図 3-2-1 に示すように、原文(日本語)テキストを入力して、CDL 表現に変換し、言い換えエンジンが言い換え規則と言い換え辞書を利用して別の CDL 表現に変換する。このプロセスは日－日機械翻訳のトランスファに相当する。変換された CDL を生成すると改善された（技術用日本語に近づいた）テキストが出力される。また、グラフ型オーサリングシステムからの日本語入力も可能とした。

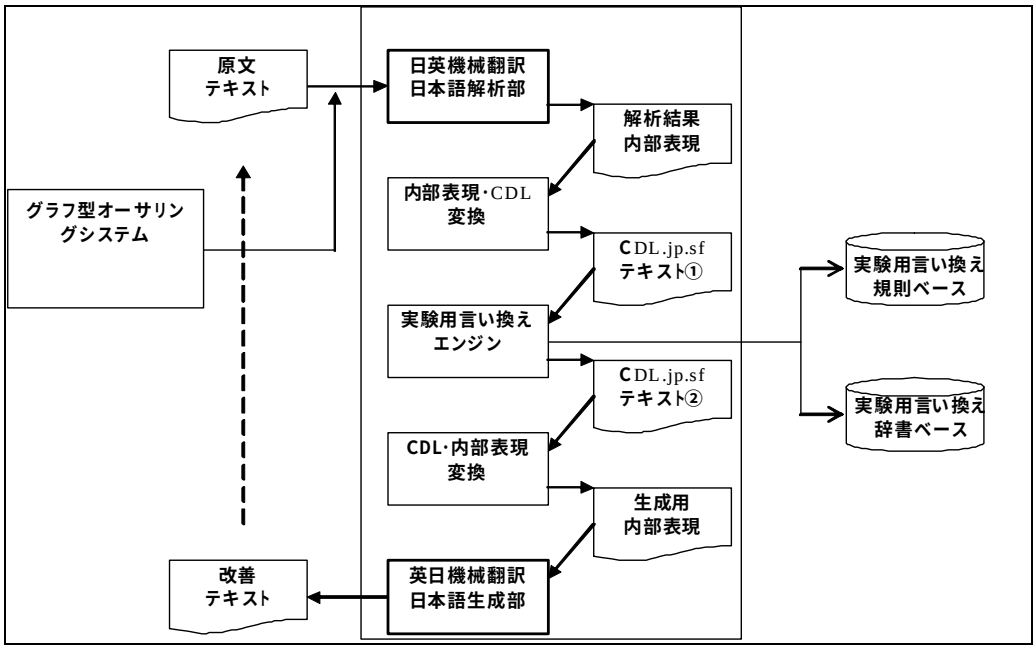


図 3-2-1　技術用日本語オーサリングシステム用実験ソフトの構成

（2）CDL.jpn.sf の仕様

実験ソフトは、日本語を CDL.jpn.sf（sf は surface を意味する）に変換して処理を行う。そのために CDL.jpn.sf の言語仕様を設定した。CDL は図 3-2-2 に示すようにノードと呼ぶ実体概念とアークと呼ぶ関係概念を基本要素とする。以下で、【】はメタカテゴリーを意味する。今回は概念として表層レベルの語彙を用いることにした。CDL.jpn.sf はフラット型 fCDL.jpn.sf とモジュール型 mCDL.jpn.sf の 2 種類を定めた。その理由は、言い換え処理のソフトを人間の自然な感覚で記述できるようにするためである。

```
ノード（実体概念）：　{【ヘッド】;【ボディ】}　　//ボディはノード、アークの入れ子構造
アーク（関係概念）：　　［【from ラベル】　【関係名】　【to ラベル】］
```

図 3-2-2 CDL の基本要素

図 3-2-3 はビジュアル化したもので、外側のノード（○）がヘッド、内側のノードとアーク（→）がボディとなる。ボディの要素は空集合でも可能。

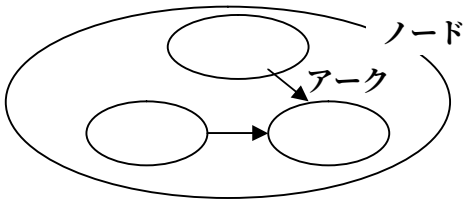


図 3-2-3 ノード／アークの概念図

また、図 3-2-4 にて、(a)は単語のリスト、(b)は単語と単語の係り受け関係のリストとなる。つまり、fCDL.jpn.sf では、単語リストと係り受けリストが分離して記述される。

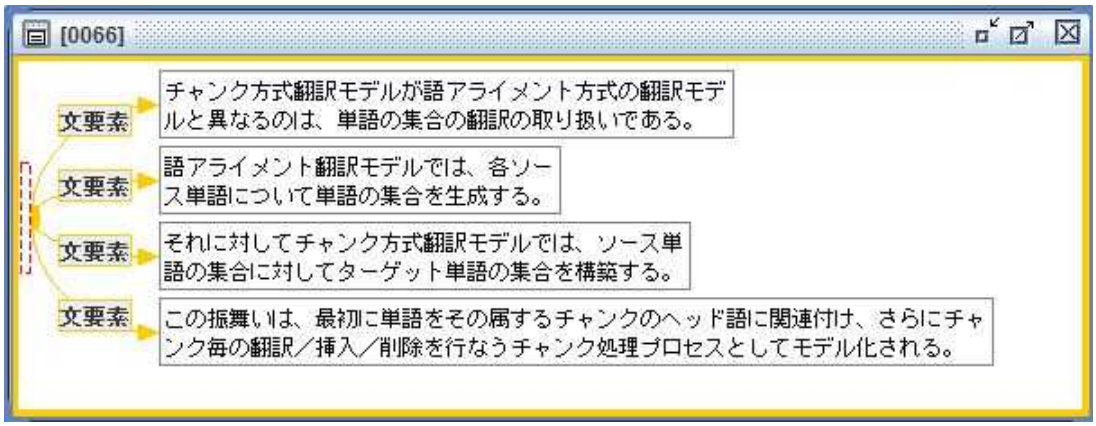
```
{#s【ID】　文　sent=<【文の文字列】>;【fCDL.jpn.sf かmCDL.jpn.sf によって異なる】}
        //文(センテンス)の定義
 {#s【ID】　　文　　sent=<【文の文字列】>;
    {#【ID】　　【単語見出し】　　【属性情報】…;}。。。・・・(a)
    ［#【from-ID】　　【関係名】　　#【to-ID】］。。。　・・・(b)　　}
       //【fCDL.jpn.sf】の定義　。。。(1 個以上の繰り返し) …(0 個以上の繰り返し)

 {#s【ID】　　文　　sent=<【文の文字列】>;　【文成分】。。。}
    【文成分】::= {#【ID】　　【単語見出し】【属性情報】…;　【文成分】...}
    ［#【ID】　　【関係名】　　#【to-ID】］
       //【mCDL.jpn.sf】の定義
```

図 3-2-4 CDL.jpn.sf の「文」対応の仕様

（3）グラフオーサリングの仕様

グラフオーサリングは、技術用日本語オーサリングシステムに文を超えた文章を網状言語で記述してそれを CDL.jpn.sf で実装し、最終的に日本語を生成する。その日本語が技術用日本語オーサリングシステムに入力される。変換ステップを以下に示す。

①セマンティックオーサリングを用いて文書を記述する。（以下は「段」の記述例）



↓

②図式（網状）表現を CDL 表現に展開する。

↓

③その CDL から日本語文を生成する。

（上記①から生成された文を以下に示す。）

> **チャンク方式翻訳モデルが語アライメント方式の翻訳モデルと異なるのは、単語の集合の翻訳の取り扱いである。**
>
> **語アライメント翻訳モデルでは、各ソース単語について単語の集合を生成する。**
>
> **それに対してチャンク方式翻訳モデルでは、ソース単語の集合に対してターゲット単語の集合を構築する。**
>
> **この振舞いは、最初に単語をその属するチャンクのヘッド語に関連付け、さらにチャンク毎の翻訳／挿入／削除を行なうチャンク処理プロセスとしてモデル化される。**

プロトタイプとして、特許明細書の作成を想定して特許明細書のオントロジーを図 3-2-5 に示すように策定した。

請求項のオントロジーは、図 3-2-6 に示すように、「請求項」と「構成要素」のプロパティを作成し、プロパティの対象はラベル付きハイパーノードで記述する。

1 特許請求の範囲（請求項）
2 明細書
2.1 発明の名称
2.2 発明の詳細な説明
2.2.1 技術分野
2.2.2 背景技術
2.2.3 発明の開示
2.2.4 発明が解決しようとする課題
2.2.5 課題を解決するための手段
2.2.6 発明の効果
2.2.7 発明を実施するための最良の形態
2.2.8 産業上の利用可能性
2.2.9 図面の簡単な説明
3 図面
4 要約書
4.1 要約
4.2 課題
4.3 解決手段
4.4 選択図

図 3-2-5 特許明細書のオントロジー

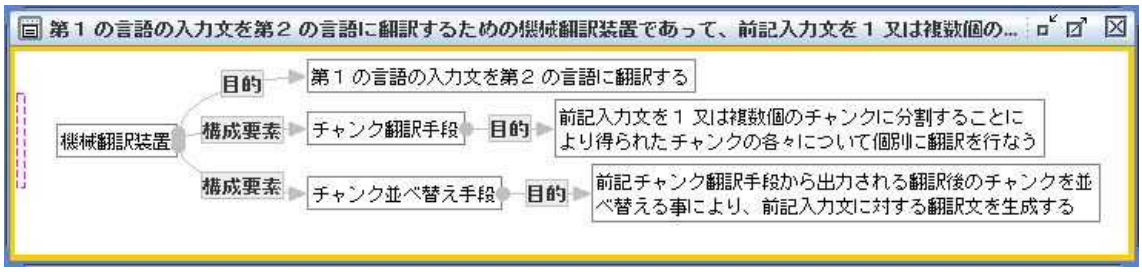


図 3-2-6 請求項の記述例

３－２－２　開発ソフト

（1）実験ソフトのシステム

実験ソフトの目的は、機械支援による言い換え処理のフィージビリティスタディであり、解析処理と生成処理は特許の機械翻訳で実績のある既存のソフトを活用した。日本語を解析し、その出力をモジュール化し、言い換えエンジンで言い換えを行い、言い換え出力を非モジュール化し、日本語生成に渡す。この一連の処理を結合して、言い換え実験ソフトを開発し、動作及び評価実験を実施した。開発したソフトウェアは、図 3-2-7 に示す。

図から分かるように、日本語解析、言い換えエンジン、日本語生成は共通のインタフェースとして fCDL.jpn.sf によって情報交換を行う。

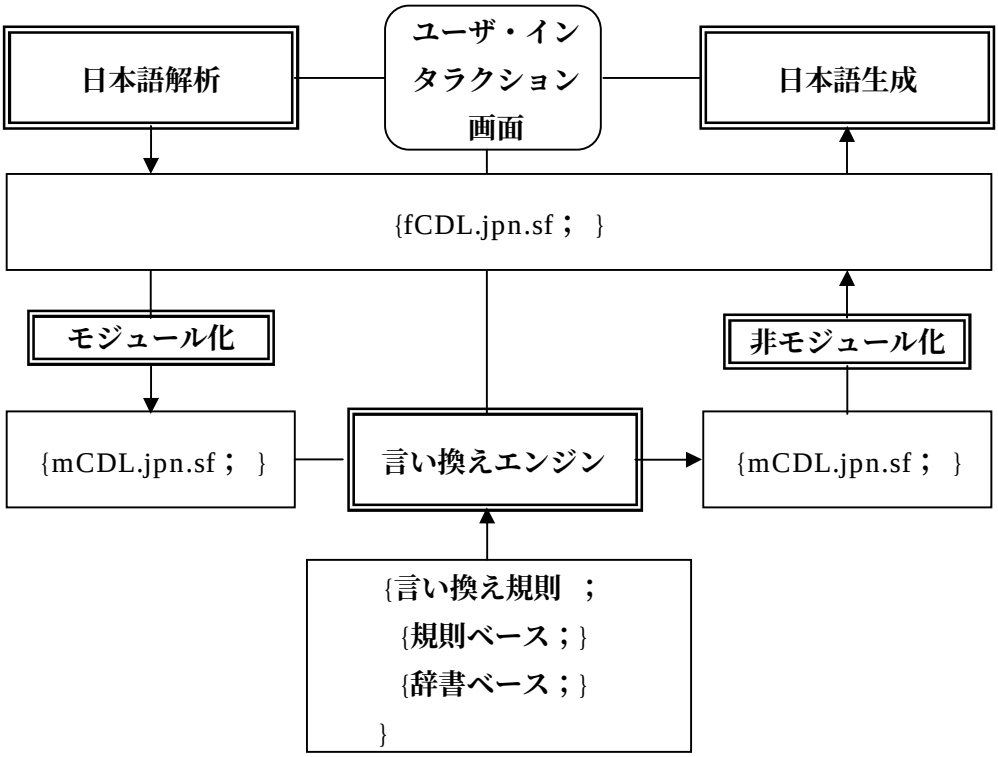


図 3-2-7 技術用日本語オーサリングシステム用実験ソフトの構造

（2）日本語解析処理の例

特許明細書の例文と、それを解析して CDL 表現内部形式に出力した例を示す。

{#s081224-2 **文** sent=<**暗渠用ブロック１０に、暗渠長手方向に相隣る暗渠用ブロック１０同士を互いに合決で接続するための切欠きを施し、これにより暗渠長手方向に突出した合決部１１，１１を形成する。**>;

{#0 **暗渠**	pos=<**名詞**>	fw=<>	info=<.....>;}
{#1 **用**	pos=<**接尾辞**>	fw=<>	info=<.....>;}
{#2 **ブロック**	pos=<**サ変名詞**>	fw=<>	info=<.....>;}
{#3 **１０**	pos=<**数字**>	fw=<**に、**>	info=<.....>;}
{#4 **暗渠**	pos=<**名詞**>	fw=<>	info=<.....>;}
{#5 **長手**	pos=<**固有名詞**>	fw=<>	info=<.....>;}
{#6 **方向**	pos=<**名詞**>	fw=<**に**>	info=<.....>;}
{#7 **相隣る**	pos=<**サ変名詞**>	fw=<>	info=<.....>;}
{#8 **暗渠**	pos=<**未知語**>	fw=<>	info=<.....>;}
{#9 **用**	pos=<**接尾辞**>	fw=<>	info=<.....>;}
{#10 **ブロック**	pos=<**サ変名詞**>	fw=<>	info=<.....>;}
{#11 **１０**	pos=<**数字**>	fw=<>	info=<.....>;}
{#12 **同士**	pos=<**名詞**>	fw=<**を**>	info=<.....>;}
{#13 **互い**	pos=<**形容動詞**>	fw=<**に**>	info=<.....>;}
{#14 **合決**	pos=<**未知語**>	fw=<**で**>	info=<.....>;}
{#15 **接続**	pos=<**サ変名詞**>	fw=<**する**>	info=<.....>;}
{#16 **ため**	pos=<**名詞**>	fw=<**の**>	info=<.....>;}
{#17 **切欠き**	pos=<**未知語**>	fw=<**を**>	info=<.....>;}
{#18 **施**	pos=<**動詞**> inf=<**五さ**>	fw=<**し、**>	info=<.....>;}
{#19 **これにより**	pos=<**副詞**>	fw=<>	info=<.....>;}
{#20 **暗渠**	pos=<**名詞**>	fw=<>	info=<.....>;}
{#21 **長手**	pos=<**固有名詞**>	fw=<>	info=<.....>;}
{#22 **方向**	pos=<**名詞**>	fw=<**に**>	info=<.....>;}
{#23 **突出**	pos=<**動詞**> inf=<**五さ**>	fw=<**した**>	info=<.....>;}
{#24 **合決**	pos=<**未知語**>	fw=<>	info=<.....>;}
{#25 **部**	pos=<**接尾辞**>	fw=<>	info=<.....>;}
{#26 **１１**	pos=<**数字**>	fw=<，>	info=<.....>;}
{#27 **１１**	pos=<**数字**>	fw=<**を**>	info=<.....>;}

{#28 形成　　pos=<サ変名詞>　　fw=<する。>　　info=<.....>;}

[#0　　連体(.....)　　#1]
[#1　　連体(.....)　　#2]
[#2　　隣接(.....)　　#3]
[#2　　ニ格(.....)　　#18]
[#4　　連体(.....)　　#5]
[#5　　連体(.....)　　#6]
[#6　　ニ格(.....)　　#15]
[#7　　連体(.....)　　#8]
[#8　　連体(.....)　　#9]
[#9　　連体(.....)　　#12]
[#10　　連体(.....)　　#11]
[#11　　連体(.....)　　#12]
[#12　　ヲ格(.....)　　#15]
[#13　　連用(.....)　　#15]
[#14　　デ格(.....)　　#15]
[#15　　連体(.....)　　#16]
[#16　　ノ格(.....)　　#17]
[#17　　ヲ格(.....)　　#18]
[#18　　連用(.....)　　#28]
[#19　　連用(.....)　　#23]
[#20　　連体(.....)　　#21]
[#21　　連体(.....)　　#22]
[#22　　ニ格(.....)　　#23]
[#23　　連ヲ(.....)　　#25]
[#24　　連体(.....)　　#25]
[#25　　隣接(.....)　　#27]
[#25　　ヲ格(.....)　　#28]
[#26　　並列(.....)　　#27]
}

（3）言い換え処理

言い換え処理の内部構造は、図 3-2-8 に示すように、並列に実行できることに配慮して 3 つにグルーピングしている。言い換え規則の実行制御部が入力文の mCDL.jpn.sf を 1 文ずつ入力し、各グループが規則ベースと辞書ベースを適用して行く。言い換え結果は

mCDL.jpn.sf で出力し、次のステップに移動する。

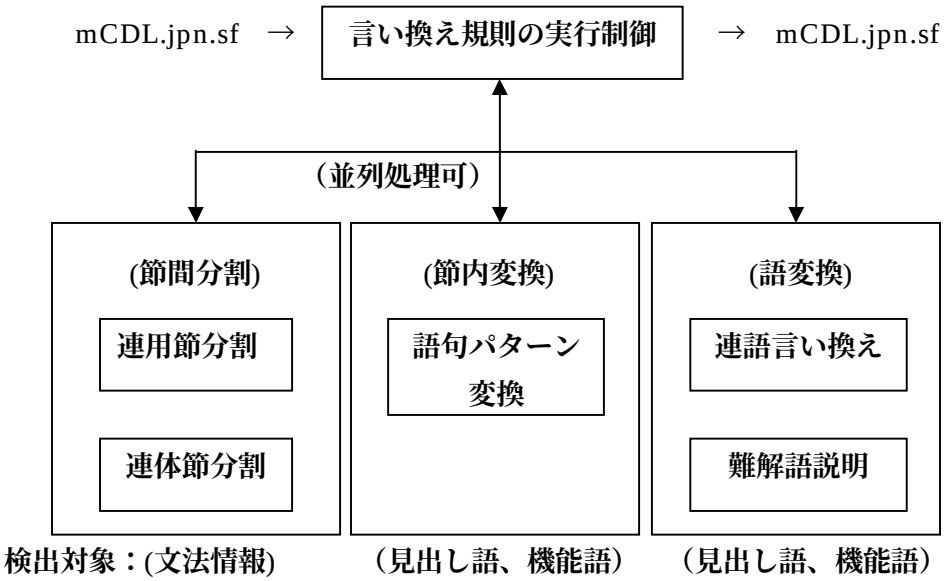


図 3-2-8 言い換え処理の構成

（4）言い換え規則の表現

言い換え規則は、規則ベースと辞書ベースに分かれる。規則ベースは、構文的な情報を利用した言い換え変換の集合であり、辞書ベースは語彙的な情報による言い換え変換の集合である。

規則の記述は図 3-2-9 に示すように、＜自然言語による記述＞、＜CDL.jpn.sf 構造を想定した処理記述＞、＜実行関数＞の 3 種類の記述を併記すること（前 2 者は省略可能）とした。先ず、＜自然言語による記述＞は、形式的な言語表現に通じていない言語専門家でも解読できるようにしておく。これは規則のメンテナンスで重要な情報源となる。＜CDL.jpn.sf 構造を想定した処理記述＞は、自然言語記述を形式化した記述で、言い換え処理を関数実装するときに CDL グラフを処理するときの方法を設計情報として記述しておく。これを元に＜実行関数＞を記述する。

【言い換え規則データベースのスキーマ】

```
{#pproot 言い換え規則；
 {#pproot1  規則ベース；
     {#r【pattern】-【number】   【規則名】；
                {【自然言語による記述】；}
                {【DL.jpn.sf 構造を想定した処理記述】；}
                {【実行関数】；}    }。。。
 }
```

```
{#pproot2　辞書ベース；
    {#d【number】　【見出し語】；
            {【言い換えタイプ】；}
            {【タイプ毎の言い換えデータ】；}　}。。。
  }
}
```

図 3-2-9 言い換え規則のスキーマ表現

言い換え規則は、表 3-2-1 に示すように 5 つのクラスに大分類した。その中でクラス１、２、３は単純な規則ではなく、種々の条件下で変化するもので規則ベースとした。上述したように規則は形式的な記述に限定することはしていない。クラス４、５は変換パターンが単純であるので以下に例示するように変換方法を直接辞書ベースに記述することとした。

表 3-2-1　言い換え分類と処理方式

言い換えクラス	言い換え処理の説明
（節間分割） １．連用節分割	(自然言語記述）文節数が一定以上のときに長文と見なし、連用節を分離する。両端に近い分割は避ける。分割候補は用言（動詞、形容詞、形容動詞、サ変名詞）で、終止形にする。分割の際、文頭に提題(～は、)があれば、主節、従属節の両方にかける。用言付属の接続助詞などによって、分割後の文に接続詞を付与する。分割終了後、未だ分割すべき点があるかもしれないので繰り返す。 （CDL 想定記述）連用修飾のアークを削除し、２分割した節を文形式の CDL に整形する。提題があれば従属節の用言に修飾させる。分割後の文 ID は“ -シリアル番号 ”を繋げる。主節と従属節の順序や結束表現などは研究課題である。属性情報は変更しない。全て生成側で修正する。
(節間分割） ２．連体節分割	(自然言語記述）文節数が一定以上のときに長文と見なし、連体節を切り出す。被修飾された体言は切り出された節の格成分に入れる。元の位置の体言には「その」などの指示詞を追加する。連体修飾のパターンは非制限的/制限的用法の分類があるがここでは非制限的なものとする。 （CDL 想定記述）注目する体言をコピーし、連体節に埋め込む。その際の関係概念名は係り受け関係で記載されているときはそれを使用し、記載されていないときはデフォルトとして「ハ格」とする。注目体言には、「その」あるいは「この」などを結合させることも検討する。連体節の分割は、主節と従属節の順序や指示詞の指定、結束表現など、研究課題が多い。

（節内変換） ３．語句パターン	(自然言語記述)「コミュニケーションを取る→コミュニケーションする」、「A を B とする→A が B である」、「N しか、V しない→N だけ V する」といった文内の単語や句の構文的な共起関係の変換規則のクラスである。これは、自立語や格助詞の変更を伴う。第 2 の例であれば、「する」がありそれに「ヲ格」と「ニ格」が修飾している場合に言い換えを実施する。 （CDL 想定記述）これは規則のキーワードとなる関係名（上記第 2 例ではヲ格、ニ格)、fw 情報などをトリガーにしてグラフ変換及び被修飾ノードの処理を行う。（上記第 2 例では、hw「する」を「である」に変換する。）
（語変換） ４．１語／連語⇔句節	(自然言語記述) 複合語、連語や難解語を辞書ベースに登録してある句や節の言い換えデータに言い換える。 一般的には、名詞は名詞句に置き換わる。サ変名詞は「する」が付くか否かで節あるいは句に置き換わる。難解語は“（説明文）”のように補足説明として難解語の後ろに挿入する。 （CDL 想定記述)複合語などは辞書に言い換える句の fCDL.jpn.sf 表現が記録されているとする。難解語は文字列である。原文中の該当する連語ノードが置き換えられる。言い換えタイプで判断する。
(語変換) ５．形態素レベル	冗長語の削除など、構文的な変換が必要でないもので、辞書に登録されている通りに置換あるいは、削除する。言い換えタイプで判断する。

（5）実験結果

長文分割の実験に使用したテキストは、機械翻訳分野の特許明細書から 50 文、電気自動車分野の特許明細書から 50 文をとり、機械処理で長文分割（連用節と連体節）を実施した。また、節内のパターン変換及び複合語と難解語の言い換えも実験し、言い換え規則が指示した所望の結果を得た。表 3-2-2 に、言い換え結果のサンプルを示す。

表 3-2-2 長文分割の実行例

（連用節分割）語アライメント方式の翻訳モデルの生成では、ソース文に含まれる単語の集合の各々について個別に翻訳語を生成してターゲット単語の集合を生成し、さらにそれらターゲット単語の、翻訳文内での位置を決定する事により翻訳を行う、という戦略を採っている。	・語アライメント方式の翻訳モデルの生成では、ソース文に含まれる単語の集合の各々について個別な翻訳語を生成してターゲット単語の集合を生成する。 ・語アライメント方式の翻訳モデルの生成では、それらターゲット言語の翻訳文内の位置を決定する事により翻訳するいう戦略を採る。

（連体節分割）統計的機械翻訳では、第１の言語の文と第２の言語の文との多数の対訳文を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う。	・学習が第１言語の文と第２言語の文の多数の対訳文を含む対訳コーパスを利用した。 ・統計的機械翻訳では、予め学習することで翻訳モデルを作成しこの翻訳モデルを用いて翻訳する。

クラス３、４、５については、個別のパターンを定めて実験した。これらのパターンを辞書に格納することで、本格的な言い換え知識となる。以下にクラス３（語句パターン）の実行結果を示す。

A. 言い換え対象の原文

{#3A-a 文 sent=<情報手段の進歩により、海外の人々と外国語でコミュニケーションを取る機会が増えている。>;}

B. 解析結果（CDL 表現内部形式）

{#0 情報 pos=<名詞> fw=<>;}
{#1 手段 pos=<名詞> fw=<の>;}
{#2 進歩 pos=<サ変名詞> fw=<により、>;}
{#3 海外 pos=<名詞> fw=<の>;}
{#4 人々 pos=<名詞> fw=<と>;}
{#5 外国語 pos=<名詞> fw=<で>;}
{#6 コミュニケーション pos=<サ変名詞> fw=<を>;}
{#7 取 pos=<動詞> inf=<五ら> fw=<る>;}
{#8 機会 pos=<名詞> fw=<が>;}
{#9 増え pos=<動詞> inf=<下わ> fw=<ている。>;}
[#0 連体 #1]
[#1 ガ格 #2]
[#2 連用 #9]
[#3 ノ格 #5]
[#4 並列 #5]
[#5 連用 #7]
[#6 ヲ格 #7]
[#7 連体 #8]
[#8 ガ格 #9]

C. 言い換え結果(CDL 表現内部形式)

{#0　情報　pos=<名詞>　fw=<>;}
{#1　手段　pos=<名詞>　fw=<の>;}
{#2　進歩　pos=<サ変名詞>　fw=<により、>;}
{#3　海外　pos=<名詞>　fw=<の>;}
{#4　人々　pos=<名詞>　fw=<と>;}
{#5　外国語　pos=<名詞>　fw=<で>;}
{#7　コミュニケーション　pos=<サ変名詞>　fw=<する>;}
{#8　機会　pos=<名詞>　fw=<が>;}
{#9　増え　pos=<動詞> inf=<下わ>　fw=<ている。>;}
[#0　連体　#1]
[#1　ガ格　#2]
[#2　連用　#9]
[#3　ノ格　#5]
[#4　並列　#5]
[#5　連用　#7]
[#7　連体　#8]
[#8　ガ格　#9]

（6）文構造の表示

言い換えを行う前に、システムの日本語解釈の内容をユーザに示すことは有用である。例えば、次の日本文を処理する場合について示す。

> 以下の説明では、ソース言語の文及び単語をそれぞれソース文及びソース単語、ターゲット言語の文及び単語をそれぞれターゲット文及びターゲット単語と呼ぶ。

この文をシステムが以下のように解釈したとする。

以下の説明では、ソース言語の文及び単語を　それぞれソース文及びソース単語、ターゲット言語の文及び単語を　それぞれターゲット文及びターゲット単語と　呼ぶ。

文の構造を示すために、ユーザに以下のような表示を行うことは有効である。

<…説明では、…単語を…単語を…単語と呼ぶ>

この表示により、ユーザは容易にシステムの解釈の誤りを判断することができ、効率よ

く日本語の言い換えを行うことができる。

ユーザは「ソース単語」の直後に「と呼び、」を加えると、システムが新たに解釈し、次のように表示する。

以下の説明では、ソース言語の文及び単語を　それぞれソース文及びソース単語と呼び、
ターゲット言語の文及び単語を　それぞれターゲット文及びターゲット単語と　呼ぶ。

同様に、ユーザに以下のような構造表示を行う。

<…説明では、…単語を…単語と呼び、…単語を…単語と呼ぶ>

これで、システムが正しい構造を認識していることが分かる。

３－２－３　実験の評価

（1）全体評価

解析・言い換え・生成を繋げた実験では特に大きな課題は見つかっていない。

実験では、特許の実文に対する表 3-2-1 の 5 種類の言い換えクラスに対して、概ね正しい日本語表層文を出力することができた。また、1 文ずつのオーサリングとは異なる既存のテキストファイルをバッチ的に言い換え処理するための実験としてソフト系の特許文 50 文と機械系の特許文 50 文の長文分割を実施した。いずれの場合も、言い換え後の出力は言い換える前の文に対して、明晰な表現であるといえる。

言い換え文で、一部の単語の語尾活用が正しくない事例があった。これは、日本語解析モジュールで解析されたデータに含まれる属性（進行、受身、など）の表現が、日本語生成モジュールで利用するデータに期待する属性の表現と一致しないものがあったからである。これはデータ変換時に属性名・属性値を適切に変換し、解析モジュールと生成モジュールで属性を共有することにより解決するものである。

また、言い換えた部分以外で、語順が変わってしまうものもあった。これは、CDL 表現内部形式が、表層の語順を反映したデータであるのに対して、生成モジュールはその語順を利用しないで文法の記述に従って語順を決めるからである。これは、生成モジュールの処理方針にかかわる問題である。原文の語順や格助詞を最大限再現することが求められるなら、解析モジュール・生成モジュールと、言い換えモジュールのインタフェースを変更する必要が生じる。

今回は解析モジュールで係り受け解析まで行う中で、一部意味解析に及ぶ処理も含んでいる。これは、解析モジュールが日英機械翻訳用に設計・構成されたものであるからである。正しい英訳を出力するために、早期の段階で一部意味的な処理を行っている。原文の語順や格助詞を最大限再現するなら、少し浅い解析処理に抑え、意味的な変換処理を省いた解析結果を CDL 表現内部形式に出力するべきである。この場合、生成モジュールも表層表現を最大限利用する機能を追加する必要がある。

（2）言い換え処理の評価

A. 規則の記述

言い換え処理の設計を進めた結果、当初計画を変更した点は、言い換え処理の共通のデータ交換フォーマットを XML ベースから CDL.jpn.sf ベースへ変更したことである。変更理由は、構文レベルの言い換えを実装する際に、構文木を処理することになるのだが、そのときの処理単位は構文的な固まり(モジュール)のコピー、挿入、削除、変更などの操作が必要なことにあった。XML でそのような構文木操作をコーディングするには直感的なロジックを反映しにくくなり、言い換え規則の共有やメンテナンスコストを考えるとき、言い換え規則のロジックの理解容易性を維持できる方式がベストであると判断したからである。実験の結果、言い換えエンジンは直感的かつかなり平坦なロジックで記述でき、動作することが実証できた。本方式が XML ベースよりも高次レベルの記述能力があることが示せたと考える。今回の言い換え規則の仕様では、自然言語による記述、CDL.jpn.sf を意識した記述及び、実際の実行関数の 3 点セットとしたが、この仕様の評価は規則毎に評価データを収集する必要がある。

B. 言い換え品質

今回の実験ソフトでは、言い換え処理の動作実験のソフト開発が主たる目標であり、言い換えられた日本語の品質に関しての評価は課題としていなかった。言語的な評価は機械翻訳の訳文評価と類似の作業になるが評価文は母国語であるので機械翻訳に比べれば難度は低くなるはずである。機械翻訳などと同様に言い換え規則数が増大すると、そこから新たに技術用日本語の言語的評価に対する課題が生じることが考えられる。また、文脈や意味情報を踏まえた言い換え規則の言語的評価は研究課題である。

C. 言い換え処理の実行性能

言い換え処理は、日-日機械翻訳と見なせるので、実行時間も日英機械翻訳と本質的な差はないと考える。即ち、実行時間に対する性能的な課題は生じない。ただし、大量の文章を言い換える場合、例えば、10 年分の特許を全て言い換えるような場合は、プラットフォームのアーキテクチャの再検討が必要となる。並列的な仕組みを取り入れているので、バッチ処理に対しても十分対応できるものと考えている。

３－２－４　技術用日本語オーサリングシステムの今後の課題

技術用日本語オーサリングシステム用実験ソフトを開発する過程で得られた検討課題を以下に挙げる。FS としては実験していないが、本格的な開発フェーズで検討すべきと考えられることも含む。

・言い換えの実行ログを蓄積して言語知識集合知に格納するメカニズム
・個別の規則毎に言い換え変換の評価を行う必要があるがその仕掛けや方法の検討
・規則は次第に成長して行くことになるがそのときの管理方法
・集合知サーバと関連する「集合知プラント」のような知識創出専用モジュールの設計
・特許を対象とすると過去文を一括処理するときの方法
・特許あるいは、特許の特定分野に限定したドメイン固有の規則知識の抽出方式
・入力文によって発生する副作用の回避策
・過去の修正履歴を活用できる仕組みを導入すること。
・言い換え辞書ベースはユーザがカスタマイズできるようにする。（仮名漢字変換のユーザ辞書のようなもの）
・言い換えではなく、「診断」だけの機能も提供できるようにする。
・文脈や文体を考慮した、より高度な言語知識に基づく変換アルゴリズムを検討する。
・技術用日本語プラットフォームが全体として 1 つの窓で操作できる環境技術用日本語

オーサリングシステムの一つの目的は、技術用日本語を用いて高精度の日英機械翻訳システムを提供することであった。

昨年度のフィージビリティスタディでは、非明晰な日本語を明晰化することによって英訳の精度が向上することを示した。本年度は、非明晰な日本語に対して、日本語解析処理、生成処理、言い換え処理を開発して機械支援で明晰化できることを示した。今後、この 2 つのスタディを継承して、日本語オーサリングシステムから生成された技術用日本語が日英機械翻訳システムの高精度化につながることを実証する必要がある。

３－３　技術用日本語言語知識集合知サーバ用実験データの開発と動作・評価実験

本スタディでは、言い換え辞書ベースの基本仕様を策定し、基本辞書項目（複合語、多義語、難解語、結合価パターン）に関する数十語の実験データを開発し、当該実験データが言い換え変換で機能するかの実験を行った。

本スタディにおける言語知識集合知サーバの開発と動作・評価実験は、わかりやすい特許文を作成する支援システムの開発や、公開された特許文を読む者が理解しやすいように特許文自体を構造化して捉えるための手がかりとなるシステムの開発や、より性能のよい特許機械翻訳システムを作るための辞書の開発などに関係する。

総合科学技術会議から平成 20 年 5 月に公開された「革新的技術戦略」において、高速大容量通信網技術、電子デバイス技術、高度画像技術、組込みソフトウェア技術、地球温暖化対策技術、知能ロボット技術、医療工学技術、再生医療技術、創薬技術、検知技術、食料生産技術、希少資源対策技術、グリーン化学技術、新材料技術の 14 項目が革新的技術に挙げられている。これら新規技術開発には新技術用語、異表記の統一、対訳といった作業が発生する。言語知識集合知サーバはそのような語彙辞書に関する仕様策定と作業の効率性を目指して実験データの検証を行っている。ここでは、複合語、多義語、難解語、結合価パターンを基本辞書項目とし、以下のように整理してデータの収集を実施した。更にこれらのデータを実験ソフトで利用する実験を実施した。

検討項目は辞書記述にかかわる以下の範囲である。

① 複合語の分析と記述案の作成

複合語を構成する語に分解し、複合語の構造などを中心に検討する。これは、複合語を分解して構成語の組み合わせで理解することによって、より理解しやすい語彙構造を得て言い換え表現を獲得したり、対訳語が登録されていない場合の専門語などについて、とりあえずの翻訳語を得るための一手段になったりするものである。

② 多義語

語義弁別の方策について検討するものである。

まず機能語（主に格助詞）で複数の役割を持っているものを一覧し、多義の可能性を示唆する。

③ 難解語

①の複合語の検討にも関わる部分であるが、ここでは特許特有の難解語を試験的に選び、言い換え語による記述を試みた。

④ 結合価パターン

特許文を自然言語処理技術によって解析する場合、係り受け関係の判定ができること

が前提となる。同じ文内に述語が複数あった場合、格成分が後続のどの述語にかかるかの判定には、通常、結合価パターンによる判定が用いられる。

また、特許文記述の時点において誤った格助詞を用いて表現している場合など、基準とする格助詞パターンによって誤りと思われるものを抽出し、自動的に警告することも可能である。本スタディでは、そういった解析に有効な結合価パターンを辞書記述の中に含めることを想定し、結合価パターンについて検討した。更に動詞の結合価パターンを得る方法をまとめた。

３－３－１　複合語

特許文献においては意味を限定するために、複数の単語（主に漢語）をつないで複合語を作るケースが多い。これは意味を限定するには有効であるが、機械翻訳などにおいて、その複合語に対応する訳語がなければ、「語－対－語」ではなく、「語－対－句」で表現しなければならないこともある。そこで複合語については、特許文と言う性格上、そのまま使うとしても内部辞書においては、分解して表現した形で橋渡しすることも必要である。ここでは、複合語の各要素を係り受け関係のレベルに分解して記述し、辞書の中に複合語を分解したレベルで情報を記載することについて検討した（表 3-3-1、表 3-3-2 参照）。

表 3-3-1 複合語の分解例

複合語	構成語を用いて句レベルに展開した形				
組み換えプラスミド	組み換え	た	プラスミド		
該組み換えプラスミド	該	組み換え	た	プラスミド	
形質転換され	形質	が	転換され		
ラン藻シネココッカス	ラン藻	の	シネココッカス		
形質転換体	形質	を	転換	する	体
該形質転換	該	形質	を	転換	する

表 3-3-2 複合語の記述形式案

項目	内容
登録番号	1
見出し	金属加工
読み	きんぞくかこう
全体品詞	名詞
語構成	金属/加工
言い換え	金属【を/で】加工する

形態素情報	**金属**(N)/**加工**(SA)
構文関係	(**金属**(**格関係**)/(**加工**(SA))
意味関係	**金属**←［**材料**/**道具**］←**加工**

３－３－２　多義語

自立語と付属語に分けて多義語の考え方を以下のとおり検討した。

（１）複数の意味を持つ多義語は文脈によって判断せざるを得ない。意味レベルまで言語処理をする場合には、もしその語の意味を限定できる他の単語に置き換えられるなら、記述時に限定して書くのが賢明である。「取る」という語を用いてこの記載例を以下に示す。（　）内は置き換える語を示す。記述形式を表 3-3-3 に示す。

「吸引ルーメン断面積を大きく【取る】（確保する）ことによる吸引力の向上」
「レバー部材の回転角度を大きく【取る】（確保する）こと」
「スクリーン上の被印刷剤を掻き【取る】（掻き取り去る）」
「前記各システムA，Xは相互に連携を【取る】（持つ）ことで」
「次に【取る】（選択する）べき運転情報」

（２）同じ文字列の助詞でも、格助詞、接続助詞のように異なる機能を持つ品詞になるものもある。更に格助詞内でも、文脈により異なる機能になるものがある。特に格助詞については、述語の係り受け解析にあたって重要な役割を担うため、該当文の中での機能が単独で決まることが望ましい。付属語「で」の多義性の例を示す。

①格助詞：操舵角が左右【で】等しくなる（場所），パネル【で】構築(材料)，モータ【で】デッキを（道具）
②接続助詞：煩雑な操作を必要としない【で】読み取る
③形容動詞の語尾、断定助動詞の「だ」の活用形：トリガ信号を高速【で】伝送する

表 3-3-3 多義語の記述形式案

項目	内容
見出し語	取る
品詞	動詞
語義	①領域、時間、数量などを確保する
例	断面積を大きく【取る】
言い換え（単語表現）	確保する
見出し語との意味的関係	＞
結合価パターン	ガ・ヲ
体言部分の意味マーカ	ガ（主体）
体現部分の意味マーカ	ヲ（領域、場所、空間、エリア、隙間）

３－３－３　難解語

難解な特許用語は、専門語によって、より明確に定義する必要があるので必ずしも専門語が不要というわけではない。ただし、広く知識を知らせるという用途においては、一般語での解説も必要である。ここでは特許特有の難解語を試験的に選び、言い換え語による記述を試みた(表 3-3-4 参照)。

表 3-3-4 多義語の記述形式案

項目	内容
見出し	合決
読み	あいじゃくり
品詞	N
英訳	Straight scarf joint
語義	貼り合わせる板の厚さをそれぞれ半分ずつ欠き取ること。
言い換え（句表現）	厚みを半分ずつ削った組み合わせた接合
構文係り受け関係	（（（（１厚み）２を）（（３半分ずつ）４削っ））５た）（（６組み合わせ）７接合）
意味関係	（（（（１厚み）２を）⇒[obj]⇒（（３半分ずつ）４削っ）５た）⇒[modify]⇒(（６組み合わせ）７接合）
言い換え（単語表現）	接合
見出し語との意味的関係	＜

３－３－４　結合価パターン

結合価とは、用言にかかる「体言＋格助詞」の組み合わせによって表現されたものである。特許文を自然言語処理技術によって解析する場合、係り受け関係の判定ができることが前提となる。そういった解析に有効な結合価パターンを辞書記述の中に含めることを想定し、結合価パターンについて検討した。表 3-3-5 は、「配送」の結合価の例である。

表 3-3-5 多義語の記述形式案

中核文	格助詞	体言	格助詞	体言	格助詞出現順パターン
分割鍵を各ユーザに配送する	を	分割鍵	に	各ユーザ	_を_に
該アクチュエータにエネルギーを配送する	に	該アクチュエータ	を	エネルギー	_に_を

３－３－５　言い換え実験

言い換え実験は、難解語と複合語の一種である臨時一語に対して実施し、言い換え処理で機能していることを確認した。

①難解語例：　表 3-3-6 のテーブル形式で関係データベースに登録し、hw（見出し語）にマッチすると def 部分を補足説明として追加する。

表 3-3-6 難解語登録データ例

hw	def
合決	貼り合わせる板の厚さをそれぞれ半分ずつ欠きとること

②複合語例：　表 3-3-7 のテーブル形式で関係データベースに登録する。複合語「冷媒減圧時」は句レベル「冷媒を減圧する時」で言い換えるために、CDL.jpn.sf による記述で登録しておく。

表 3-3-7 複合語登録データ例

hw	wnum	word_lis	CDL.jpn.sf
冷媒	3	「減圧」 「時」	{#9000　冷媒　pos=<名詞>　fw=<を> ;} {#9001　減圧　pos=<サ変名詞>　fw=<した>;} {#9002　時　pos=<名詞>　fw=<>　;} [#9000　ヲ格　#9001] [#9001　連体　#9002]

３－４　まとめ

経済活性化のための技術用日本語プラットフォームに関するスタディを実施した。技術用日本語共通基盤仕様に則り、技術用日本語プラットフォームシステムの開発計画を策定し、実験ソフトと実験データによりオーサリングシステムの実装について検証した。

技術用日本語それ自体は、アプリケーション分野により種々の言語クラスに最適設計されるが、プラットフォームの基盤は図 3-4-1 に示すような 3 階層からなる共通のアーキテクチャである。

図 3-4-1　技術用日本語プラットフォームのアーキテクチャ

産業技術文書の典型として特許文書を中心に技術用日本語のオーサリングシステムや言語知識集合知サーバの検討を行ってきたが、並行して機械翻訳、文書検索への応用検討を進めたことで技術用日本語プラットフォームでのサービスの全体像を示すことができた。

技術用日本語は、人間の自然言語と機械用の概念記述言語 CDL が連携して機能するオープンな言語であり自然言語による分かり易さ、正確さ、簡潔さを具備したコミュニケーション力を提供し、今後産業界において標準的な言語として使用され、グローバルな経済産業活動における競争力強化への貢献が期待できる。

本プラットフォームのもう一つの特徴は、日本版クラウドコンピューティング環境上で実装することを想定して検討したことである。そのことにより、技術用日本語の言語仕様及びプラットフォームシステムが、機能面、性能面、規模の点でスケーラビリティが担保されることになる。産業分野の自然言語は技術の進展と共に進化成長してゆくものであるが、それに対して柔軟に対応できるというメリットは長期的にも大きな経済効果をもたらすものと期待できる。

４　スタディの今後の課題及び展開

４－１　技術用日本語の普及

国土と資源が限られた我が国が、今後も、国際競争力を維持、向上させていくためには、新しい知財を創出し続けることが重要である。そして今、知財立国の時代にあって、我々の知恵を、正確かつ明解に表現し、確実に伝達し、無駄なく活用していくことが、今まで以上に重要になってきている。技術用日本語は、我が国産業界全体の国際競争力の強化に資することを目的に、産学官連携で取り組むべき重要かつ広範囲なテーマである。

そして、技術用日本語の導入の対象となる産業技術情報は、産業活動のすべてにかかわる情報である。そして、産業活動全体を通じて、あらゆる側面で、情報がなめらかに流れることによる効果は、計り知れないインパクトを秘めている。

技術用日本語プラットフォーム開発計画の実施にあたり、最も大切なことは、まず、技術用日本語とそのプラットフォーム開発計画を産業界に広く深く普及させることである。そのためには、３－１－１開発計画の中で言及している「技術用日本語フォーラム」や「技術用日本語コンソーシアム」といった、関係各省庁、大学や研究機関、ユーザ企業・団体、サービス企業・団体と連携し、技術用プラットフォーム開発計画を共有しつつこれを実施するための具体的な枠組み作りを早急に行う必要がある。

４－２　開発への着手

技術用日本語プラットフォームの開発では、3 年間の開発フェーズに沿って開発を推進することを目指すが、当初計画のような総合的な開発に着手するにはタイミング的な調整が必要なことも予想される。そのような事態も想定して、図 4-2-1 に示すように先ずは特許関連文書を技術用日本語の普及及びプラットフォームシステム・プラットフォームアプリケーションの開発のための牽引役とするべく研究開発に着手することを考えている。このプロトタイプの開発により技術用日本語プラットフォームシステムのノウハウ（特に、オーサリングシステム、言語知識集合知サーバ、機械翻訳、文書検索）を蓄積し、知財サイクルのモデルサービスや産業技術文書、科学技術論文への展開を目指す。

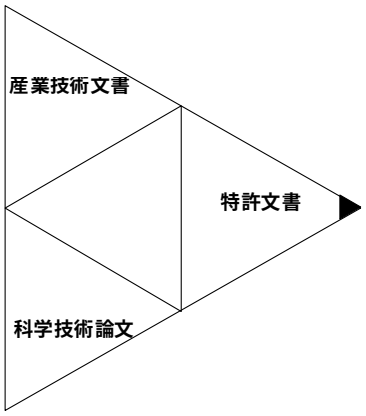


図 4-2-1　技術用日本語の牽引役：特許文書

４－３　学会での研究推進

技術用日本語は、その網羅すべき分野の広さから、多数の研究課題を含んでおり、また、技術用日本語プラットフォームの開発を支え、これを拡充していくための基礎・応用研究自体も推進していく必要がある。そこで、言語処理研究の中心的存在である言語処理学会などの関係学会の協力を仰ぎ、大学・研究機関などを主たる会員とする「技術用日本語フォーラム」を定期的に開催し、技術用日本語に関する研究の深化と学術的な発展を図る。また、フォーラムの開催を介して、技術用日本語に関する研究成果を大学・研究機関の間で共有することに加え、年に１回程度その活動成果を公開発表することにより、広く世論を喚起する。

４－４　産業界の協力体制作り

技術用日本語を広く産業界に根付かせるために、「技術用日本語フォーラム」の開催と並行して、ユーザ企業・団体やサービス企業・団体など、産業界を主たる会員とする振興機関「技術用日本語コンソーシアム」の設立を目指す。「技術用日本語コンソーシアム」は、関係省庁・関係団体の協力の下、「技術用日本語フォーラム」の会員・大学・研究機関と連携し、技術用日本語に関するサービス事業を立ち上げ、普及することを目的に活動する。

システム開発　　２０－Ｆ－５

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２１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