人工知能学会
第12回知識流通ネットワーク研究会
SIG-KSN-012-03

自己記述性に基づく専門用語辞書作成支援手法

Term Dictionary Construction based on Self-Descriptiveness

竹内広宜[1*]　中村大賀[1]　荻野紫穂[1]
Hironori Takeuchi[1]　Taiga Nakamura[1]　Shiho Ogino[1]

[1] 日本アイ・ビー・エム株式会社 東京基礎研究所
[1] IBM Research - Tokyo

Abstract: In this paper, we considered to construct a term dictionary needed in SI projects by applying document analytics technologies to existing artifact documents. We proposed a method considering the self descriptiveness of the extracted terms. We compared the new method with existing methods and found that we can reduce the costs to create an initial term dictionary with a fixed-size.

1　はじめに

システム構築などのプロジェクトでは，要求仕様書をはじめとして様々な文書成果物が作成される．これらの成果物の間には，関連性があるが，文書成果物は大量に作成され，その書き手はそれぞれ異なるといった場合が多い．そして，それぞれの文書成果物に変更が入るため，矛盾や不整合が生じる可能性がある．また，要求仕様書などは自然言語で書かれることが多く，曖昧性を持った記述が含まれる可能性もある．一般に，テストなどの後工程に行くほど，発見された不具合の修正にかかるコストは高い．そのため，プロジェクトの早期に作成される文書成果物の品質を高めることは重要であり，文書成果物中の問題を発見する試みが行われている [4, 5, 8].

文書成果物の品質を高めるにあたって，用語の整理や統一は重要な活動の１つである．ここで用語とは，構築するシステムが適用されるビジネスに関する専門用語をはじめ，アクター・エンティティ・画面といったシステムの要件や設計に関わる重要な概念の名称である．通常，用語は，記述中の名詞，表や式などの一部として文書中に現れる．用語はシステム構築において重要な概念を指し示す名称であるため，システム構築に関係する人が用語について共通の理解を持つ必要がある．そのため，重要な用語については， プロジェクト内で意味定義とともに用語辞書として整理し，プロジェクトメンバー間で共有することが行われている．もし，同じ概念を示す異表記を用いると，関係者間で異なる理解が生じる可能性がある．たとえば，[4] では，ある組み込みソフトウェア開発における不具合の原因を調べた結果，文書成果物中の表現に対する誤解が原因であるものが 16% であり，用語の不統一が不具合の一因となっていたことが報告されている．そこで，プロジェクトでは，用語辞書内の各用語について異表記を禁止語として定義し，作成された文書成果物の用語を統一することが行われている．

用語辞書作成では，用語をその意味定義などとともに，主に人手で登録することになる．本論文では過去の類似プロジェクトの成果物といった既存文書に対して，自然言語処理などのテキスト分析技術を適用し，用語辞書を作成することを試みる．通常の用語抽出技術では，出現頻度を元に対象分野に固有な語を抽出することが行われている．これに対し，本論文ではさらに，抽出された語が，意味定義を付与するのに十分な情報をもっているかどうかという「自己記述性」という観点で用語抽出を行うことを試みる．そして，実プロジェクト文書への適用を通して，頻度情報のみを利用した従来手法に比べ，効果的に用語辞書が作成できることを示す．

本論文の構成は以下の通りである．まず 2 章ではシステム構築プロジェクトなどにおける既存文書からの，文書構造解析技術，自然言語処理技術を用いた用語抽出について述べる．3 章では用語抽出の実践における課題と，関連する従来手法 について述べる．そして，4 章で自己記述性を用いた提案手法を説明し，その評価実験について 5 章で述べる．6 章で考察を行い，最後に 7 章で本論文のまとめを行う．

*連絡先：日本アイ・ビー・エム・株式会社 東京基礎研究所
〒 135-8511 東京都江東区豊洲 5-6-52
E-mail: HIRONORI@jp.ibm.com

2　自然言語技術を用いた既存文書からの用語辞書作成

多くのプロジェクトでは，対象とするビジネスや構築するシステム固有の知識をドメイン知識として整理・共有する必要があり，様々なドメイン知識の表現が検討されている [10]．その表現形式の 1 つとして，概念（キーワードや表現）を，その説明記述とともに用語辞書として登録することが行われている．自然言語処理技術の 1 つに，入力テキストを単語に分割し，各単語に品詞を付与する形態素解析技術がある．プロジェクト固有の用語のほとんどは名詞であるため，既存文書成果物に形態素解析技術を適用し，名詞を頻度とともに抽出することで用語辞書の元となるデータ（用語候補リスト）を作成することができる．現在，ChaSen[2], 言葉 Web[9] といった形態素解析ツールや専門用語抽出ツールが公開され，広く利用可能となっている．しかしながら，既存文書成果物から用語候補リストを抽出するにあたって，以下の課題がある．

1. 入力となる文書成果物は Microsoft®Word, Excel, PowerPoint などで，プロジェクト固有の形式で書かれていることが多い．
2. 文書成果物中には，更新履歴などプロジェクト固有の概念が出現しないセクションなどが存在する．
3. 形態素解析技術は，一般的な用語が中心である新聞記事を用いて学習しているものが多い．そのため，用語候補として抽出したい長い専門用語が必ずしも名詞 1 語として認識されない．

これらの課題を解決するために構築した用語候補リスト作成のフローを図 1 に示す．

まず，上記課題の 1,2 に対応するため，文書成果物の構造をモデル化する技術 [5] を用いて，プロジェクト固有のフォーマットで書かれた文書成果物について，特定の箇所からのみ記述を抽出する．次に抽出されたテキストに対して形態素解析技術および構文解析技術といった自然言語処理技術を適用する．形態素解析では，ツールが持っているシステム辞書によって分野固有の長い複合名詞が短い名詞で分割されることがある．そのため，上記課題の 3 に対応するため，本研究では構文解析ツールを適用して得られた文節のうち名詞を含む文節から付属語（助詞，句読点）を取り除くことで用語候補を抽出した．例えば,「中途解約受取額は預金期間から決まる.」という入力テキストから,「中途解約受取額は」「預金期間から」「決まる.」が文節として得られ，最終的に「中途解約受取額」「預金期間」が用語候補として得られる．

これら文節の同定および複合語の抽出を行う前に，前処理として，入力テキストを句読点だけでなく，括弧

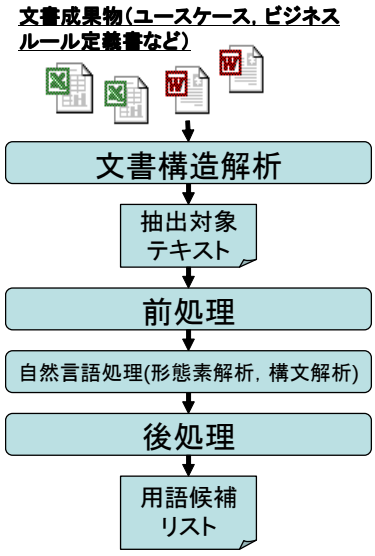

図 1: 用語候補リスト作成フロー

などの記号を区切り文字として分割する．こうすることで,「機能としてお試しモード（仮称）がある」といった記述から括弧つきの語 (お試しモード（仮称）) が抽出されることや，数式などが用語として抽出されることを回避する．そして，得られた自然言語処理の結果に対して，名詞を含む文節（名詞節）から，付属語を取り除き名詞および複合名詞を抽出する．これにより，形態素解析ツールでは分割されてしまう，プロジェクト固有の長い複合名詞を 1 語として抽出することができる．また，抽出された名詞および複合名詞に対し，出現箇所の前方の文字を走査する．そして，前方に隣接する文字が漢字である場合は，前方に追加することで抽出語を更新する．この過程を後処理と呼ぶ．これにより，文節区切りの際，誤って分割されてしまった未知語を抽出することができる．これらの処理で得られた結果を，用語候補リストとして出力する．

3　用語候補リストの抽出手法と課題

2 章で述べた既存文書内にある名詞の抽出は単純な手法であるが，膨大な数の名詞が用語候補リストとして抽出される．そのため，辞書作成では，候補リストから人手で登録すべき語を選択する必要があるが，プロジェクトにおける重要語の見落としが起こる可能性がある．抽出語を頻度順に並べる手法が考えられるが，対象分野に関係がない一般語が上位に来ることが多く，用語辞書作成の効率化につながらないことが多い．

このような課題に対する用語抽出技術として参照コーパス (例えば新聞記事の集合) との比較を用いた手法がある [6]．これは抽出された語について，対象文書集合

(仕様書などの既存の文書成果物) における出現頻度と参照コーパスにおける出現頻度を比較し，対象文書集合に特有な（対象文書集合のみに多く出現する）語を抽出する手法である．対象文書集合での総単語出現数を n_d，参照コーパスでの総単語出現数を n_c とし，それぞれの文書集合において，ある語（w）の出現数を w_d，w_c とする．ここで，

$$W_d = \frac{n_d(w_d + w_c)}{n_d + n_c} \quad (1)$$

$$W_c = \frac{n_c(w_d + w_c)}{n_d + n_c} \quad (2)$$

とする．そして抽出指標として以下の対数尤度 (LL_w) を計算する．

$$LL_w = 2(w_d \log \frac{w_d}{W_d} + w_c \log \frac{w_c}{W_c}) \quad (3)$$

各語について対数尤度 LL_w を計算し，高い値を持つ対象文書集合に特に多く出現する語を用語とし抽出する．本手法を仕様書に適用し対象分野の専門用語を抽出する研究が行われている [3, 7].

上述の手法では，対象文書で特に多く出現する語が抽出される．しかし，対象文書に多く出現する語の中には，語のみを見て意味定義を付与することが困難な場合がある．ここで，意味の定義が困難な語には 2 種類あると考えられる．1 つ目は前後の文脈を見ないと意味が定まらない語である．これは，例えば,「連携情報」という語が該当する．通常，連携先は複数存在すると考えられ，対象分野の深い知識があったとしても，前後の記述を読まないと一意に理解できないことが多い．2 つ目は，語自身に曖昧性があり意味を付与できない動作名詞である．ここで動作名詞とは，動作を表す名詞であり,「する」をつけることでサ変動詞となる名詞である．例えば「無効処理」という語が該当する動作名詞となる．なぜなら，無効の場合，停止するのか，削除するのかといった具体的な動作についての情報を語から取得することが困難であるからである．

本論文では，このような,(対象分野の知識がない場合に）語からだけでは意味付与が困難な語を非自己記述的用語と定義する．そして，このような非自己記述的用語を抽出し，用語候補リスト作成に活用する手法を提案する．

4 非自己記述的用語の抽出を利用した用語候補リストの作成

前章で述べたように非自己記述的用語には前後の記述を読まないと意味付与が困難な名詞と，具体的な動作情報を含まない動作名詞がある．前者を文脈依存非自己記述的名詞，後者を非自己記述的動作名詞と定義する．本章では，これら 2 種類の語を抽出し，用語候補リストを作成する手法を提案する．

4.1 文脈依存非自己記述的名詞の抽出

まず，最初に参照コーパスに対し，形態素解析を適用し，名詞，複合名詞を抽出する．参照コーパスデータとして，本研究では，新聞記事 (2002 年の産経新聞 10,000 記事) を用いた．例えば「顧客の情報として受付情報がある」という記述からは「顧客」,「情報」,「受付情報」が抽出される．また，抽出された複合語をさらに単単語に分割する（受付情報 → 受付, 情報）．次に，単単語の名詞 (kwd) にそれぞれに対し,「連体修飾され度合い (S_1)」と「名詞修飾度合い (S_2)」を計算する．まず，コーパス中の kwd（例えば「データ」）の出現回数を M とする．そして，kwd についてコーパスデータ内で出現箇所における係り受け情報を抽出し，係り元および係り先の情報を取得する．kwd の出現箇所における係り元について，以下の場合の頻度をそれぞれ m_1, m_2, m_3 として求める．

- m_1 係り元が助詞（の，より）(例：勘定ＤＢ内の → データ)
- m_2 動詞，形容詞，形容動詞の連体修飾　(例：ユーザーが作成する → データ)
- m_3 名詞の連接 (例：顧客 → データ)

同様に，係り先として，以下の場合の頻度を m_4, m_5 として求める．

- m_4 係り先が助詞 (の) (例：データ → の保存)
- m_5 名詞の連接 (例：データ → 転送)

得られた係り受け情報を元に，kwd についての S_1, S_2 を以下のように求める．

$$S_1 = \frac{m_1 + m_2 + m_3}{M} \quad (4)$$

$$S_2 = \frac{m_4 + m_5}{M} \quad (5)$$

得られた S_1, S_2 が閾値 τ_{S_1}, τ_{S_2} を越えた場合，kwd をそれぞれ $\{N_{S_1}\}$, $\{N_{S_2}\}$ として抽出する．閾値 τ_{S_1}, τ_{S_2} は調整可能であるが両方とも 0.7 と固定した．$\{N_{S_1}\}$ には前に情報が連体修飾などの形で付加されやすい語が含まれる．参照コーパスから抽出された例として,「発生」,「記述」,「設計」,「性質」などがある．$\{N_{S_2}\}$ には後続の名詞に対して修飾しやすい語が含まれる．抽出された例として,「小型」,「合同」,「耐熱」などがある．$\{N_{S_1}\}$ および ,{Ns2} の持つ性質から，$\{N_{S_1}\}$ を

先頭語として含む語（「発生 XX」など）および $\{N_{S_2}\}$ を末尾語として含む語（「YY 合同」など）が非自己記述的である語と考えられる．

以上のように参照コーパスから得た $\{N_{S_1}\}$, $\{N_{S_2}\}$ を用いて文脈依存非自己記述的名詞を抽出する．まず，対象となる文書成果物から名詞，複合名詞する．そして抽出された名詞，複合名詞のうち，先頭語が $\{N_{S_1}\}$ に含まれているもの，または，末尾語が $\{N_{S_2}\}$ に含まれているものを文脈依存非自己記述的名詞として抽出する．その結果，例えば「発生頻度」,「性質調査」などが非自己記述的名詞として得られる．

4.2 非自己記述的動作名詞の抽出

まず，事前に CRUD 動作名詞リストを作成する．CRUD とは Create, Read, Update, Delete の 4 つの動作を示し，ほとんどソフトウェアアプリケーションが備えるべき機能を表す．CRUD に相当する動作名詞 31 語を CRUD 動作名詞リストとして事前に作成した．以下にその一部を示す．

- Create: 作成，生成，定義
- Read: 参照，読込
- Update: 変更，更新，変換
- Delete: 消去，削除，解除

次に，参照コーパスから動作名詞の連続（例：完了通知 → 完了+通知，作成処理 → 作成+処理）を抽出する．そして，得られた動作名詞の連続のうち，CRUD 動作名詞リストの後ろに出てくる動作名詞を $\{VNa\}$ として抽出する．例えば,「作成処理」,「更新手続」などから「処理」「手続」が得られる．CRUD 動作名詞が動作に関する十分な情報を持っているため，$\{VNa\}$ は曖昧な動作名詞の集合となる．

以上のように参照コーパスから得た曖昧な動作名詞 $\{VNa\}$ を用いて，非自己記述的動作名詞を抽出する．まず，対象となる文書成果物から名詞，複合名詞する．そして抽出された名詞，複合名詞のうち,「動作名詞以外の名詞＋ VNa」を含む語または VNa が先頭語となっている語を非自己記述的動作名詞として抽出する．その結果，例えば「データ処理」,「手続画面」などが得られる．

4.3 用語候補リストの作成

用語候補リストの作成では，まず 3 章で述べた従来手法を用いて，参照コーパスにおける頻度との比較に基づき対象文書集合にのみ多く出現する語を対数尤度を用いてランキングする．得られた用語候補リストから 4.1 節で得られた文脈依存非自己記述的名詞および 4.2 節で非自己記述的動作名詞を取り出す．これら文脈依存非自己記述的名詞および非自己記述的動作名詞を再び対数尤度順にランキングし，用語候補リストの最後に加え，最終的な用語候補リストとして出力する．

5 評価実験

本章では，非自己記述的用語の抽出を利用した用語候補リスト作成の評価を行う．評価実験では，入力として，あるシステム構築プロジェクトで作成された，ユースケース (Excel 文書)，ビジネスプロセス定義書 (Word 文書)，事務手続書 (Word 文書) など 8 種類の文書成果物 (合計 177 文書) を用いた．2 章で述べた用語抽出システムを用いて入力文書から名詞および複合名詞を抽出し，以下の 3 種類の手法でランキングされた用語候補リストを作成した．

1. ランダムに並べた用語候補リスト (Baseline)
2. 3 章で述べた既存手法でランキングした用語候補リスト (Existing)
3. 4 章で提案した手法を用いてランキングした用語候補リスト (New)

プロジェクトで用いられた用語辞書に登録された語を正解データとし，各手法で得られた結果を評価した．平均精度 (Average Precision)[1] を表 1 に示す．一般的に

表 1: 平均精度

手法	平均精度
Baseline	0.0601
Existing	0.0785
New	0.0895

ランキングリストの上位に正解データがあるほど，平均精度は高くなる．このことから，提案手法では，用語辞書に登録すべき語をより多く上位で検出できることがわかる．

平均精度を用いることで手法間の比較をすることができるが，この結果から，実プロジェクトへの適用で得られる効果を評価することは難しい．そこで，実プロジェクトでの適用を想定した評価を考える．プロジェクトにおいて，初期に一度作成された辞書が十分であることは少なく，プロジェクトが進むにつれて辞書は頻繁に更新される（主に辞書への用語の追加が行われる）．更新が起こる場合の例として以下がある．

- スコープの変更によりシステム化するビジネスが変更
- 新規ベンダーが設計時に参画したことにより，問い合わせや確認作業が発生
- テストケース作成時に問い合わせや確認作業が発生

このように，辞書の更新が頻繁に起こるため，プロジェクト開始時に完全な辞書を作成することは難しい．そのため，まずプロジェクトの初期段階では，コストや作成期間を考え，ある規模の辞書を初期辞書として作成し，プロジェクト期間内で随時更新することが行われることがある．そこで，このような初期辞書の作成を想定し，ある決められた語数の用語辞書を作成することを考える．そして，得られた用語候補リストから初期辞書を作成する際にかかるコストを評価する．

用語候補リストから設定したサイズの辞書を作成するにあたって，辞書に登録すべき用語かどうか精査する候補語の数をコストと定義し，各手法について調べた．図 2 に 3 種類の手法で得られた候補語リストを用いた場合における辞書サイズとコストの関係を示す．Baseline が最もコストがかかる手法であるため，その結果よりも下側であるほど，効果的に辞書構築ができることを示している．グラフから，提案手法 (New) が既存手法 (Existing) より，一定サイズの辞書を作成する際に，効果があることがわかる．また Baseline 手法でのコスト

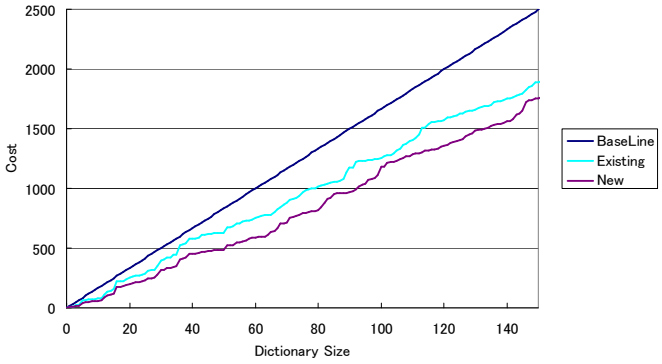

図 2: 辞書サイズとコストの関係.

を 100 とした時の比較を表 2 に示す．従来手法より，コストを約 15% 下げることができることがわかる．

最後に提案手法を用いて用語候補リストを作成する際に得られた非自己記述的用語をいくつか示す．まず文脈依存非自己記述的名詞として，

有無区分，不備情報，処理区分，募集情報， 連携情報，回数情報，情報取扱，変更有無

表 2: コストの比較

辞書サイズ	Baseline	Existing	New
20	100	77.7	61.6
40	100	89.9	70.2
60	100	78.0	60.7
80	100	78.4	63.5
100	100	78.1	72.9
120	100	78.5	68.2
140	100	75.1	67.1

などが得られた．語が出現する記述を読まないと意味定義を付与できない語が抽出されていることがわかる．次に非自己記述的動作名詞として，

手続画面，満期処理，当該処理，無効処理， 単独処理，手続内容

などが得られた．具体的な処理および手続きについての情報が含まれていないため，複数の意味が考えられる語が抽出されていることがわかる．

6 考察

提案手法を用いることで，効果的に用語辞書が作れるだけでなく，同時に非自己記述的用語が得られる．このうち，文脈依存非自己記述的名詞については，文書中の出現箇所の前後を読むことで一意に理解できるかどうかをレビュー時に注意すべき語として整備して活用することが考えられる．また，非自己記述的動作名詞は具体的な動作情報が不足している語であるため，使用すべきでない禁止語リストとして利用することが考えられる．

評価実験では，提案手法は，候補語リストの上位に辞書に登録すべき語を集められる語を集められることがわかった．これにより，ある特定のサイズの辞書を作るというタスクにおいて，人が候補語を精査するコストを下げられることが期待される．しかし，提案手法では，辞書に登録すべき語を文脈依存非自己記述的名詞と判定しまう場合もある．そのような語はランキングの下位で扱われるため，高い網羅率 (Recall) を目指し大きなサイズの辞書を考える場合にはコストが高くなる可能性がある．例えば，5 章の評価実験では,「取引先」が $\{N_{S_1}\}$ に含まれていた．これは，新聞記事においては,「石油の取引先」,「売買取引先」といったように使われているため,「取引」に関する連体修飾され度合いが高く計算されたからである．一方，システムが対象とする特定のビジネスでは「取引先」は一意に決まるため,「取引先」を先頭語として含む語のいくつかが

正解データ（プロジェクトで作成された用語辞書）に含まれていた．これらの語は，本手法では下位にランキングされてしまうため，網羅率を大幅に上げた場合に精度を保つためには何らかの工夫が必要である．例えば，連体修飾され度合い（S_1）および名詞修飾度合い（S_2）の算出に用いるコーパスデータの選択や抽出された非自己記述的用語の再評価などが考えられる．これは，今後の課題である．

文脈依存非自己記述的名詞は，ある複数の用語の共通概念を示す名詞と考えられる．そのため，抽出された文脈依存非自己記述的名詞を，階層構造を持った知識体系を構築する際に活用できると考えられる．これも今後の課題である．

7 まとめ

本論文では，システム構築プロジェクトなどで重要となる用語辞書を，テキスト分析技術を用いて既存文書から作成することを行った．従来手法では仕様書などの対象文書における頻度情報を用いて候補語を得ることが行われるが，本論文では抽出された語が意味定義を付与するのに十分な情報を含んでいるかという自己記述性という視点を導入した．そして，非自己記述的用語を抽出し，用語候補リスト作成に活用する手法を提案した．評価実験を通した従来手法との比較で，提案手法では平均精度が高いことがわかった．また，実プロジェクトへの適用を想定し，事前に設定した数の初期用語辞書を作成する際のコストを比較した．その結果，提案手法を用いることで従来手法より候補語から初期辞書を作成するコストを約 15% 削減できることがわかった．網羅性が非常に高い辞書を作成する場合においては，高精度で用語候補リストを抽出するためにさらなる工夫が必要である．今後の課題として，非自己記述的かどうかの判定に使用する参照コーパスデータの選択や抽出された非自己記述的用語の再評価や，階層構造を持った知識体系の構築における文脈依存非自己記述的名詞の利用が考えられる．

参考文献

[1] C. Buckley and E. M. Voorhees. Evaluating evaluation measure stability. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 33–40.

[2] ChaSen. http://chasen-legacy.sourceforge.jp/.

[3] R. Gacitua, P. Sawyer, and V. Gervasi. On the effectiveness of abstraction identification in requirements engineering. In *Proceedings of the 18th IEEE International Requirements Engineering Conference*, pp. 5–14, 2010.

[4] T. Nagano, Y. Sakamoto, S. Haraguchi, H. Takeuchi, S. Ogino, and A. Fukuda. Critiquing rules and quality quantification of development-related documents. In *Proceedings of the Joint Conference of the 21st International Workshop on Software Measurement and the 6th International Conference on Software Process and Product Measurement (IWSM-MENSURA)*, pp. 30–37, 2011.

[5] T. Nakamura, H. Takeuchi, F. Iwama, and K. Mizuno. Enabling analysis and measurement of conventional software development documents using project-specific formalism. In *Proceedings of the Joint Conference of the 21st International Workshop on Software Measurement and the 6th International Conference on Software Process and Product Measurement (IWSM-MENSURA)*, pp. 48–54, 2011.

[6] P. Rayson and G. Garside. Comparing corpora using frequency profiling. In *Proceedings of ACL Workshop on Comparing Corpora*, pp. 1–6, 2000.

[7] P. Sawyer, P. Rayson, and K. Cosh. Shallow knowledge as an aid to deep understanding in early phase requirements engineering. *IEEE Transactions on Software Engineering*, 31(11):969–981, 2005.

[8] G. Zu, H. Taira, K. Makino, T. Kano, and S. Matsumoto. The supporting technology of business document proofreading based on intercultural differences. In *Proceedings of IEEE Joint Conference on E-Commerce Technology and Enterprise Computing, E-Commerce and E-Services (CEC/EEE)*, pp. 91–98, 2007.

[9] 言選 Web. http://gensen. dl.itc.u-tokyo.ac.jp /gensenweb.html.

[10] 長田晃, 小澤大伍, 海谷治彦, 海尻賢二. 要求獲得におけるドメイン知識表現の役割. 情報処理学会論文誌, 48(8):2522–2533, 2007.