Vol.2013-SE-179 No.9
2013/3/11

出現単語の抽象度を用いたソフトウェアドキュメントの具体化の評価

遠藤 充[1]　森崎 修司[1]

概要：ソフトウェアドキュメントの詳細化，具体化を計測することを目指し，辞書による抽象度の定義を用いてドキュメントに含まれる語の抽象度を計測した．国語辞典の抽象度の定義を用いて同一ソフトウェアのフェーズの異なるドキュメント間で，含まれる単語の抽象度ごとの出現頻度を求めたところ，大きな違いはみられなかった．基本設計書に含まれる語よりも詳細設計書に含まれる語，要件定義書に含まれる語よりも基本設計書に含まれる語のほうが国語辞典に抽象度の定義のない単語が多く含まれており，それらの単語はソフトウェア開発の観点からみて，具体的な単語が多かった．

1. はじめに

ドキュメントによる詳細化を中心とするソフトウェア開発では，要求定義書，基本設計書，詳細設計書といったドキュメントを作成しながら，利用者の要求を段階的に詳細化した上で，コーディング，テストを実施する．コーディングとテストにおいては実際にプログラムを動作させながらプログラムが期待通りに作成されているかどうかを判断することができる．しかしながら，特定の記法やモデル化ができていない限り，要求定義書，基本設計書，詳細設計書といったドキュメントは動作させることができないので，内容を目視して妥当性を確認する必要がある．

目視によるドキュメントの妥当性確認は強力であるもののコストが大きいという問題があり，自動化等の方法により，小さいコストでドキュメント評価を試みようとする手法が提案されている [1][7]．これらの研究は大きく 2 つのアプローチに大別できる．1 つは，ドキュメント作成時に計算機が解釈しやすいルールを設け，作成者がそれに従うことにより，評価をしやすくするものである．UML 等の記述形式をはじめとして，図表や体裁のチェック [6] がある．これらの研究はドキュメント全体の評価を低コストで実施できるというメリットがある一方で，作成者に記法やタグ付けなどの制約を強いるというデメリットがある．もう一つのアプローチは，ドキュメント作成には特に制約を設けず，事前に設定したルールに違反する部分を問題点の候補として作成者に提示したり [8]，ルール違反の数によってスコアリングしたりするものである．これらの研究は，ドキュメントが何らかの記法に沿っていなかったりタグ付けされたりしていなくても評価が可能であるため，適用範囲が広くドキュメントの作成コストを増やさないというメリットがある一方で，ドキュメントの評価が局所的なものとなったり，汎用的なルールの定義が難しかったり，ルールの定義にコストがかかったりする点に課題がある．

本研究では，計算機によるドキュメント評価の自動化を目指し，作成時に記述方法等の制約を課すことなく，ドキュメント全体にわたって網羅的な定量化が可能かどうかを試行する．具体的には，同一ソフトウェアの異なるフェーズのドキュメントに出現する単語の抽象度の分布を調査する．抽象度は国語辞典で定義されているものを使う．

以降，2 章では評価の方法を述べる．3 章で 2 種類のソフトウェアのドキュメントの概要と抽象度の分布を示す．4 章で考察し，5 章でまとめる．

2. 評価方法

2.1 概要

本評価で前提としているドキュメントは，自然言語で記述され，形態素解析ツール等を用いて単語に分解できるものを対象とする．ドキュメントの記述言語に特に前提は置かないが，単語の抽象度が定義された辞書が存在することを想定している．辞書は専用に構築しても構わないが，辞書構築コストが大きいことが予想されるため，一般の国語辞典の利用を想定している．

ドキュメントを D に含まれる単語の集合を W とする．

$$W = w_1, w_2, \cdots, w_n$$

w_k は単語を表す．n はドキュメント D に含まれる単語か

[1] 静岡大学
Shizuoka University

ら重複を取り除いた単語数である．

単語 w の抽象度 w_l は辞書 Dic で定義され，単語 w に対する関数 f として次のように定義する．

$$(w_{l1}, w_{l2}, \cdots, w_{lm}) = f(w)$$

w_{lk} の値が大きいほど具体的な単語であることを示す．m が 2 以上の場合には，多義語であることを示し，w_{lk} はそれぞれの意味での抽象度である．

w が m 種類の意味を持つことを示している．多義語でない場合には $k = 1$ となり，$w_l = w_{1l}$ とする．単語 w の抽象度が辞書に定義されていない場合には $f(w) = 0$ となる．多義語の場合に，i 番目の意味で用いられていることが識別できる場合には，$w_l = w_{li}$ とする．どの意味で用いられているかが識別できない場合には，後述する方法で計算する．

2.2 手順

評価は以下の手順で実施する．手順 3 において W_1 のほうが抽象度の高い単語が多ければ，ドキュメントを段階的に詳細化する際に，単語も具体化していると考えることができる．

(1) 同一ソフトウェアの異なるフェーズのドキュメント D_1, D_2 に含まれる単語の集合 W_1, W_2 を得る．(D_1 は D_2 よりも前の開発フェーズのドキュメントとする)
(2) W_1, W_2 の各単語の抽象度を算出する．多義語である w は，w_{lk} の平均値，中央値，最小値の 3 種類の算出方法で結果を比較する．
(3) W_1, W_2 に含まれる単語の抽象度ごとの出現頻度を計数し，W_1 のほうが抽象度が高い単語が多いか比較する．
(4) W_1 に存在する単語のうち W_2 に存在しない単語を調べ，抽象度の分布を比較する．
(5) W_2 に存在する単語のうち W_1 に存在しない単語を調べ，抽象度の分布を比較する．

3. 試行

3.1 対象ドキュメントと辞書

同一ソフトウェアで 2 フェーズのドキュメントが公開されている次を対象ドキュメントとした．

- ドキュメント A 特定検診システム 基本設計書 [4]，詳細設計書 [5] の組
- ドキュメント B 災害者支援システム 要件定義書 [3] と基本設計書 [3] の組

対象ドキュメントの語数，ページ数を表 1 に示す．

抽象度を定義する辞書 Dic として日本語語彙大系 [2] を用いた．日本語語彙大系には収録されている名詞 141857 語には 1〜12 段の抽象度が定義されている．1 段がもっとも抽象的，12 段がもっとも具体的である．段数ごとの単語数には偏りがある．日本語語彙大系の各々の段に含まれる単語の一例を表 2 に示す．本評価では対象とする単語を名詞とした．

表 1 対象ドキュメントの概要

対象プロジェクト	ドキュメントの種類	語数	ページ数
特定健診システム	基本設計書	1302	17
特定健診システム	詳細設計書	36588	206
災害者支援システム	要件定義書	4167	11
災害者支援システム	基本設計書	20056	86

表 2 日本語語彙大系段別収録単語例

段数	単語例
1 段	いずれ，これ，何
2 段	具体，個体，主体
3 段	事，現象，主体
4 段	進捗，データ，組織
5 段	選択肢，数値，目安
6 段	程度，条件，順番
7 段	取得，必須，実装
8 段	四捨五入，日本語，暗号
9 段	交付，上書き，定義
10 段	禁止，メンテナンス，押下
11 段	採取，男女，同士
12 段	発行，レイアウト，掲載

3.2 結果

3.2.1 ドキュメントに含まれる単語

形態素解析ツールとして mecab を用いて，ドキュメント A, B の名詞を取り出し，日本語語彙大系で出現頻度を調べた．語数，日本語語彙大系での定義された名詞の語数をまとめたものを表 1 に示す．

3.2.2 単語の抽象度の分布

ドキュメント A の抽象度の分布を求めた結果を図 2 から図 7 に示す．ドキュメント B の抽象度の分布を求めた結果を図 9 から図 8 に示す．図の横軸は抽象度を示している．

図 1 はドキュメント A において多義語でない名詞のみの分布を示している．縦軸はドキュメント A 全体に対する単語数の割合を示している．基本設計と詳細設計の間で，9 段目では 20%以上の差がある．逆に 10 段目では 10%以上の差をつけて基本設計書のほうが出現頻度が大きかった．

図 2，図 3，図 4 はそれぞれ，多義語の抽象度の平均，多義語の抽象度の中央値，多義語の抽象度の最小値としたときの分布を示している．縦軸はドキュメント A 全体に対する単語数の割合である．

図 2 の 8 段目，9 段目の出現頻度が基本設計書よりも詳細設計書のほうが大きく，逆に 5 段目，6 段目，7 段目については基本設計書の方が出現頻度大きくなっている．最頻値についても，基本設計書については 7 段目，詳細設計書については 8 段目となっている．基本設計書よりも詳細設計書のほうが具体的な単語が増えているといえる．

図 2，図 3 の分布は類似している．多義語数が 2 の単語

表 3 対象ドキュメント内の単語数

		特定健診システム		災害者支援システム	
		基本設計書	詳細設計書	要件定義書	基本設計書
名詞総数（重複計上）		827	19994	2546	12307
名詞総数（重複除外）		237	1387	389	628
辞書に収録のない語	語数（重複計上）	21	3913	11	92
	割合（重複計上）	2.54%	19.57%	0.43%	0.75%
	語数（重複除外）	12	472	28	43
	割合（重複除外）	5.06%	34.03%	7.20%	6.85%

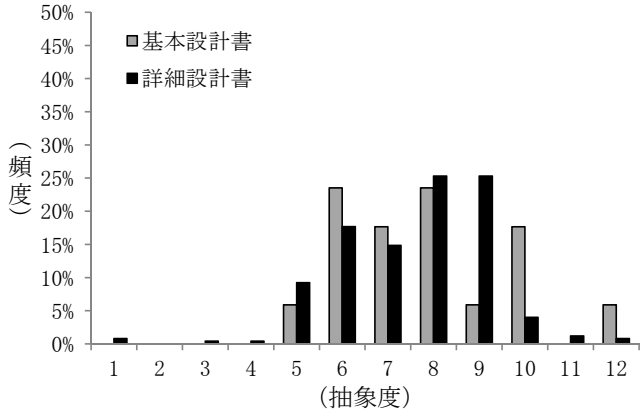

図 1 ドキュメント A（多義語なし・重複なし）

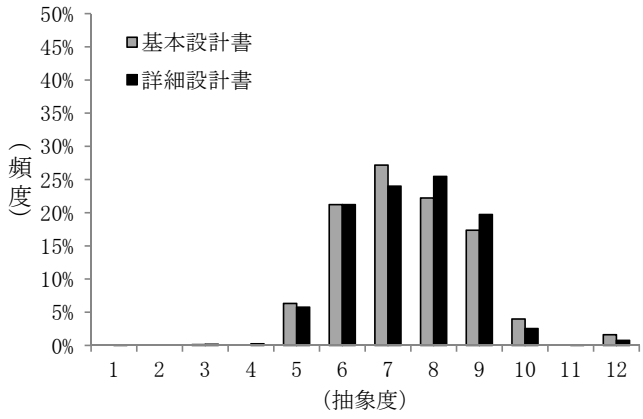

図 2 ドキュメント A（平均値・重複あり）

が多いこと，多義語の間で抽象度のばらつきも少なかったからと考えられる．図 4 では，少し傾向が異なるが，多義語の場合の抽象度の 3 つの求め方に大きな差はない．

図 5，図 6，図 7 はそれぞれ，多義語の抽象度の平均，多義語の抽象度の中央値，多義語の抽象度の最小値としたときの分布を示している．縦軸はドキュメント A に出現する全体に対する単語数の割合である．単語の重複を除いた場合も単語の重複を除かない場合と同様の傾向であった．

図 8 はドキュメント B において多義語でない名詞のみの分布を示している．縦軸はドキュメント B 全体に対する単語数の割合を示している．

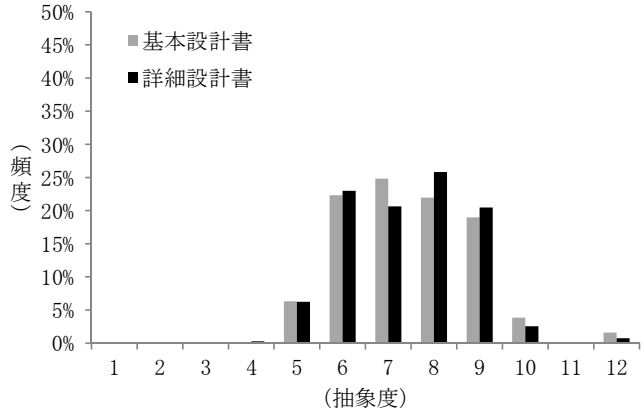

図 3 ドキュメント A（中央値・重複あり）

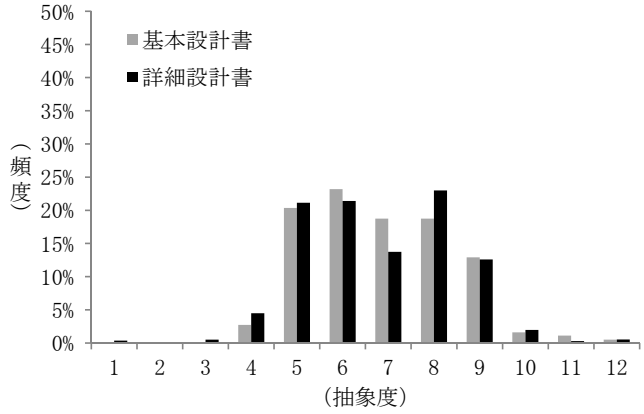

図 4 ドキュメント A（最小値・重複あり）

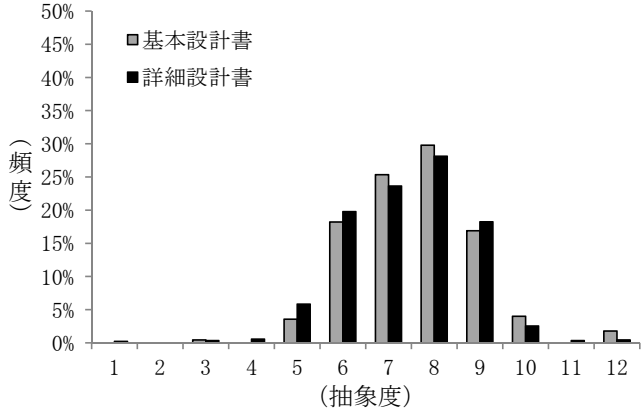

図 5 ドキュメント A（平均値・重複なし）

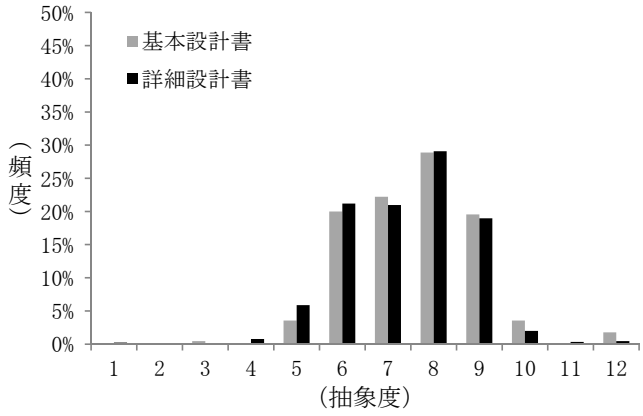

図 6 ドキュメント A（中央値・重複なし）

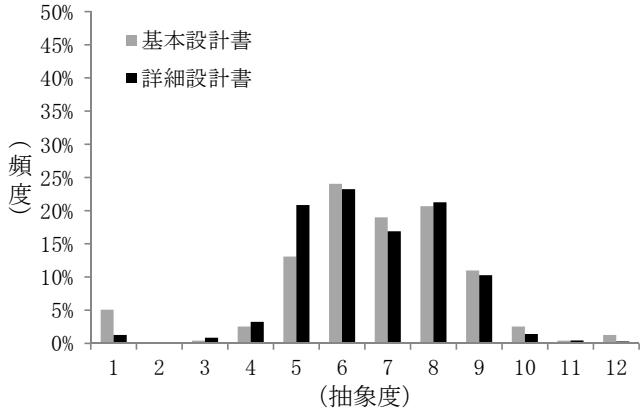

図 7 ドキュメント A（最小値・重複なし）

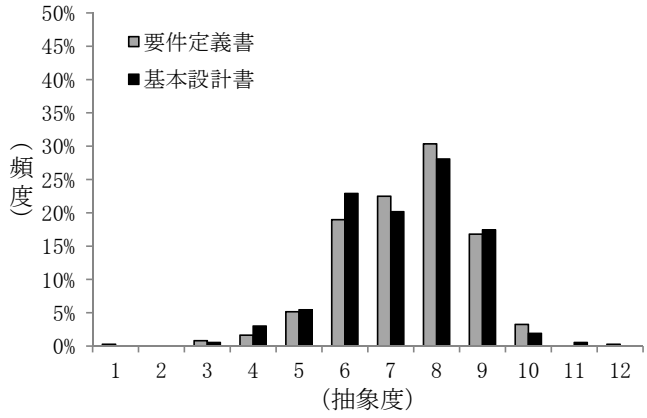

図 8 ドキュメント B（多義語なし・重複なし）

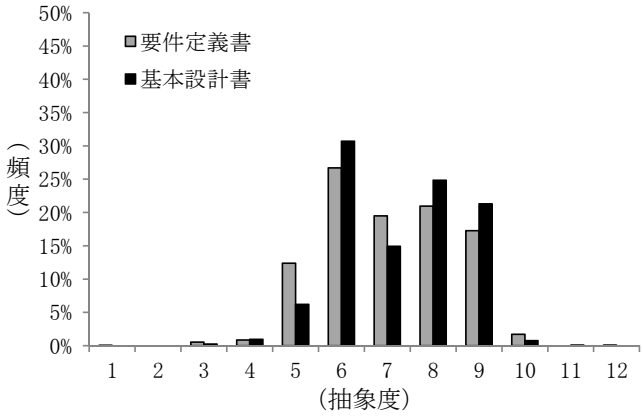

図 9 ドキュメント B（平均値・重複あり）

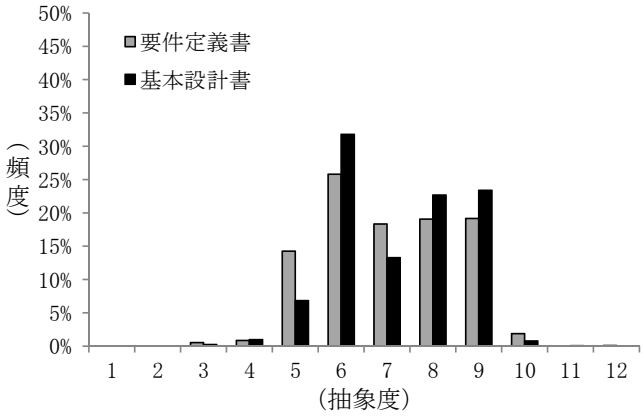

図 10 ドキュメント B（中央値・重複あり）

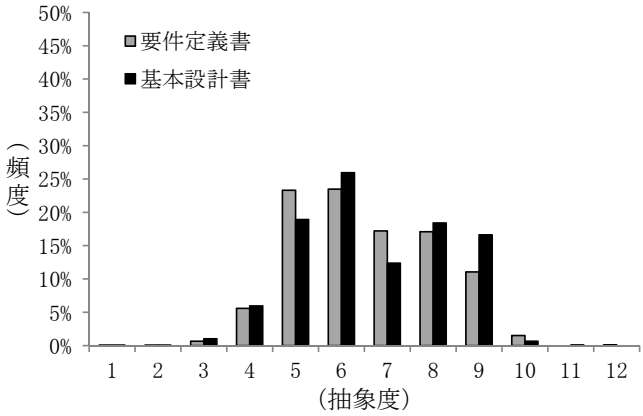

図 11 ドキュメント B（最小値・重複あり）

図 9，図 10，図 11 はそれぞれ，多義語の抽象度の平均，多義語の抽象度の中央値，多義語の抽象度の最小値としたときの分布を示している．縦軸はドキュメント B 全体に対する単語数の割合である．

図 9，図 10，は分布が類似していた．図 11 はドキュメント A と同様に全体的に抽象度が大きい語の出現頻度が大きかった．

図 12，図 13，図 14 はそれぞれ，多義語の抽象度の平均，多義語の抽象度の中央値，多義語の抽象度の最小値としたときの分布を示している．縦軸はドキュメント B に出現する全体に対する単語数の割合である．図 12，図 13 は図 8 と分布が類似していた．

3.2.3 一方のドキュメントにしか存在しない単語の抽象度

図 15，図 16，図 17，図 18 はドキュメント A において基本設計書のみ，詳細設計書のみに出現する単語の抽象度の分布を表している．図 15 は多義語の名詞を取り除いた分布，図 16 は多義語の名詞の抽象度を平均値として求めたもの，図 17 は多義語の名詞の抽象度を中央値として求めたもの，図 18 は多義語の名詞の抽象度を最小値として

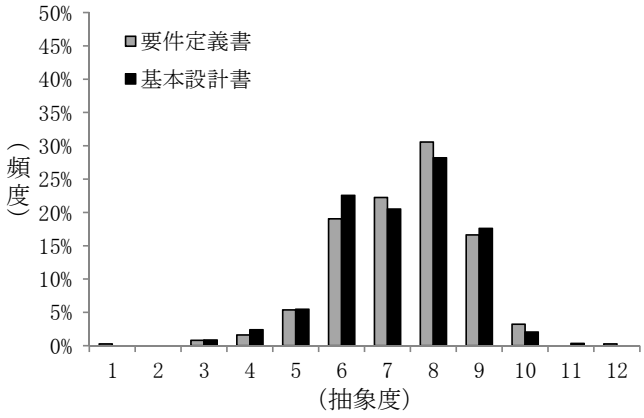

図 12 ドキュメント B（平均値・重複なし）

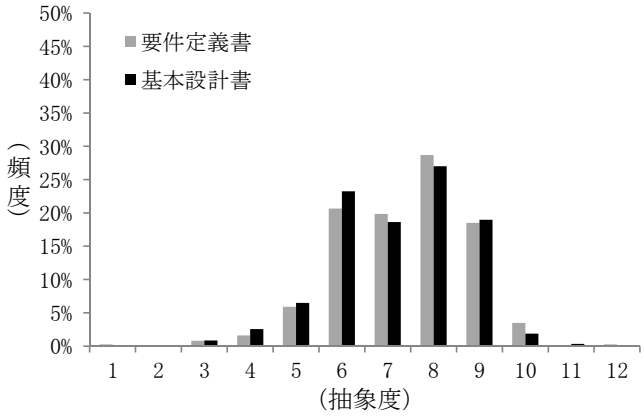

図 13 ドキュメント B（中央値・重複なし）

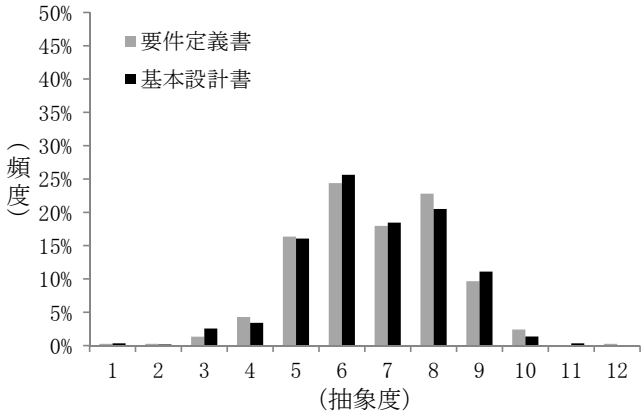

図 14 ドキュメント B（最小値・重複なし）

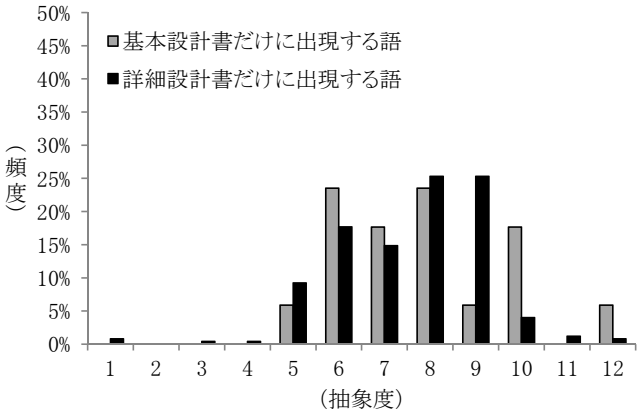

図 15 ドキュメント A（多義語除外・片方のみに現れる語・重複なし）

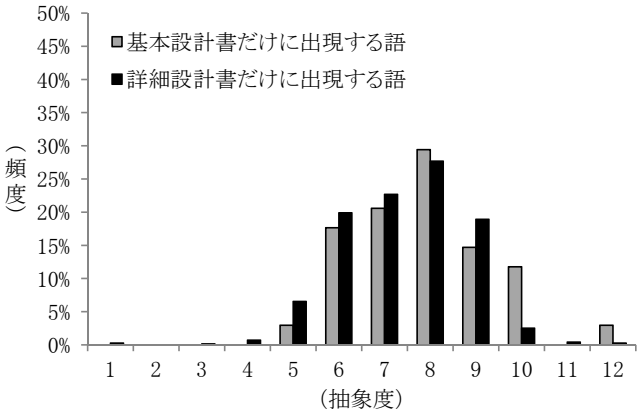

図 16 ドキュメント A（平均値・片方のみに現れる語・重複なし）

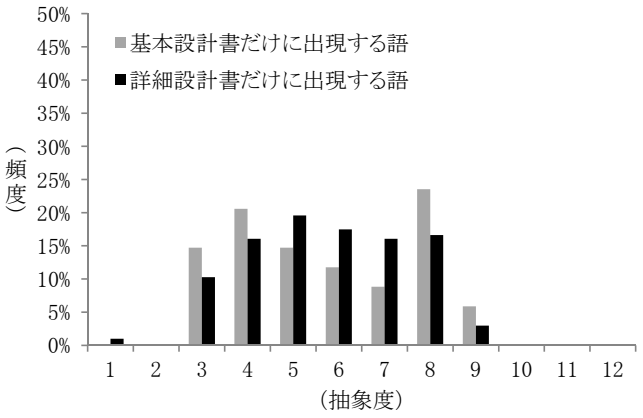

図 17 ドキュメント A（中央値・片方のみに現れる語・重複なし）

求めたもの，である．

図 16，図 17，図 18 で分布が異なった．図 17 は抽象度 5, 6, 7 において詳細設計書だけに出現する語の出現頻度が大きくなった．

図 19 は，基本設計書，詳細設計書のどちらか一方にしか現われない単語の語義数の分布を示している．基本設計書のみに出現する単語の方が語義が一意に定まるものが多かった．

図 20，図 21，図 22，図 23 はドキュメント B において要件定義書，基本設計書のみに出現する単語の抽象度の分布を表している．図 23 は多義語の名詞を取り除いた分布，図 20 は多義語の名詞の抽象度を平均値として求めたもの，図 21 は多義語の名詞の抽象度を中央値として求めたもの，図 22 は多義語の名詞の抽象度を最小値として求めたもの，である．図 20，図 21 の分布は類似していた．

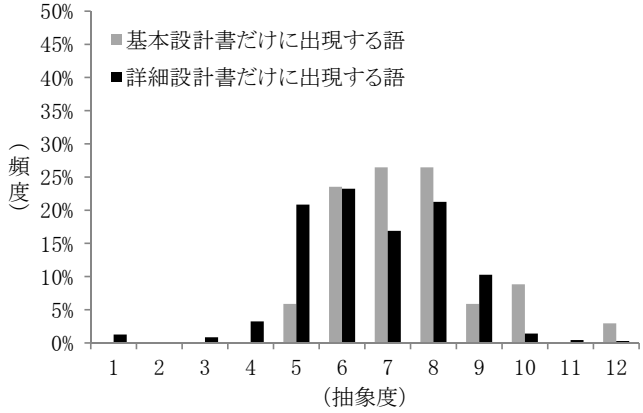

図 18 ドキュメント A（最小値・片方のみに現れる語・重複なし）

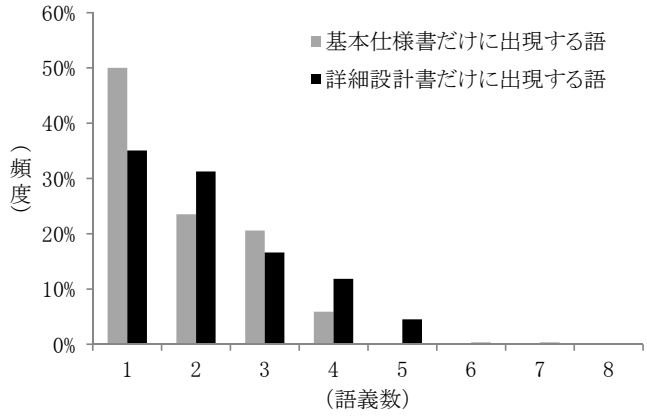

図 19 ドキュメント A（語義数分布・片方のみに現れる語・重複なし）

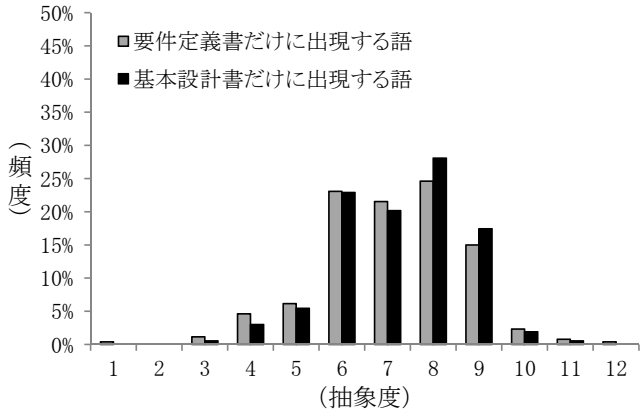

図 20 ドキュメント B（平均値・片方のみに現れる語・重複なし）

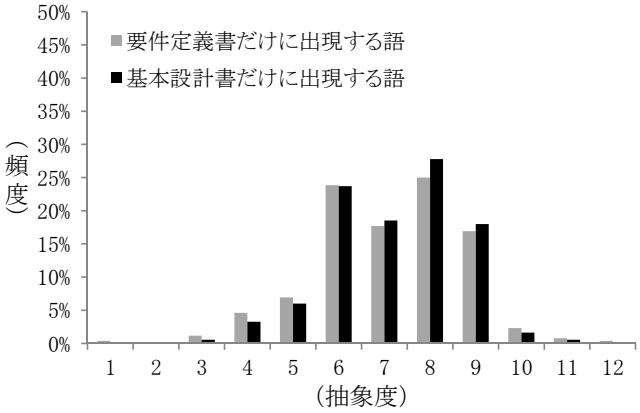

図 21 ドキュメント B（中央値・片方のみに現れる語・重複なし）

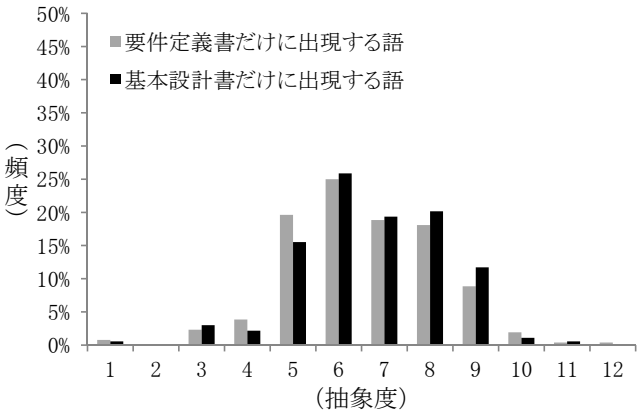

図 22 ドキュメント B（最小値・片方のみに現れる語・重複なし）

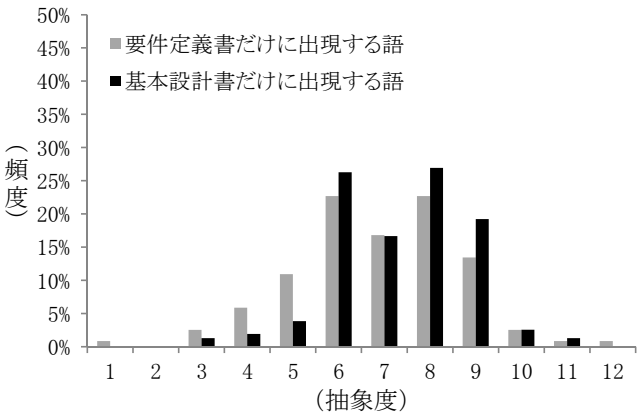

図 23 ドキュメント B（多義語除外・片方のみに現れる語・重複なし）

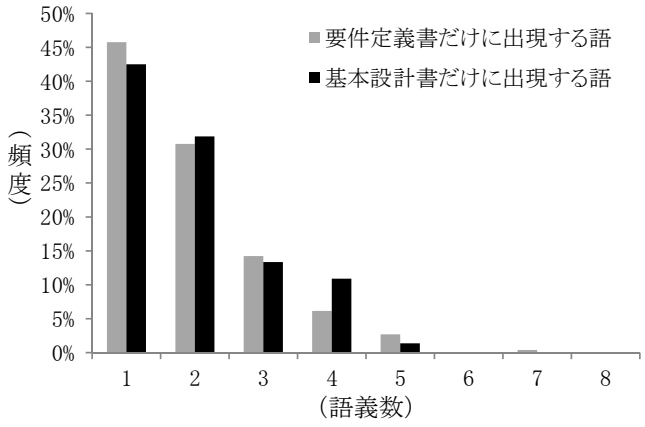

図 24 ドキュメント B（語義数分布・片方のみに現れる語・重複なし）

4. 考察

ドキュメント A の一部において，基本設計書に含まれる語よりも詳細設計書に含まれる語のほうが抽象度が大きい結果となった．多義語がある場合の抽象度の算出方法を 3 つで比較したが，平均値を使うと基本設計書よりも詳細設計書のほうが抽象度が大きな単語の出現頻度が大きくなった．ドキュメント A の単語の抽象度の出現頻度の分布から基本設計書，詳細設計書を識別することはできないと考えられる．その原因の 1 つとして，基本設計書よりも詳細設計書のほうが，辞書で抽象度が定義されていない単語が多かったことが考えられる．辞書で抽象度が定義されていない単語として次のような単語があった．これらは，ソフトウェア開発の観点からみると十分に具体化された単語であるといえ，このような単語の抽象度を補完的辞書で定義すると基本設計書よりも詳細設計書に含まれる単語のほうが抽象度が大きいという結果となる可能性がある．

- ガイドライン
- アップデート
- ピクセル
- インデント
- 文字種
- レセプトソフト
- メタボリック

ドキュメント B では，要件定義書に含まれる語の抽象度のほうが基本設計書に含まれる語の抽象度よりも小さいという結果は得られなかった．ドキュメント A と同様に，要件定義書よりも基本設計書の方が，抽象度が辞書に定義されていない語が多かった．

5. おわりに

計算機によるドキュメント評価の自動化を目指し，作成時に記述方法等の制約を課すことなく，ドキュメント全体にわたって網羅的な定量化が可能かどうかを評価した．同一ソフトウェアのドキュメントは詳細化に伴い，出現単語が具体化すると考え，ドキュメントに含まれる単語の出現頻度を，辞書で定義された抽象度ごとに調べた．具体的には，同一ソフトウェアの基本設計書と詳細設計書，及び，要件定義書と基本設計書を対として，それらの対の間で，ドキュメントに含まれる単語の抽象度が異なるかを調べた．単語の抽象度を定義する辞書として，単語の抽象度を 1〜12 に分類した日本語語彙大系を用いた．

評価結果から，対象としたドキュメントの基本設計書と詳細設計書を比較すると詳細設計書のほうが具体的な単語を若干多く含む傾向があった．もう一対の対象ドキュメントでは，要件定義書に含まれる単語と基本設計書に含まれる単語の抽象度の違いはみられなかった．基本設計書よりも詳細設計書のほうが，要件定義書よりも基本設計書のほうが辞書に抽象度が定義されていない単語の出現頻度が大きかった．今回の評価では，これらの単語は対象外としたが，個々の単語をみると，ソフトウェア開発の観点で，十分に具体化された単語であった．専用辞書によって抽象度を定義して補完する方法や辞書に抽象度が定義されていない単語を加味することが，ドキュメントの具体化の度合いの定量化に向けた今後の課題である．

謝辞 立命館大学 大西淳先生のご助言に感謝する．本研究は文部科学省科学研究補助費（基盤研究 B:課題番号 23300009）による助成を受けた．

参考文献

[1] L. Franz, Quality Control in Software Documentation Based on Measurement of Text Comprehension and Text Comprehensibility", Information Processing and Management, vol.29, no. 5, pp. 551-568(1993)
[2] 日本語語彙大系　CD-ROM 版，岩波書店 (1999)
[3] 地方公共団体内及び地方公共団体間における「被災者支援システム」を活用した防災・. 災害情報のデータ連携による効果等に関する調査研究, 入手先 ⟨http://www.soumu.go.jp/main_content/000120786.pdf⟩ (2013.02.07).
[4] 日医特定健康診査システム（仮称）基本仕様書，入手先 ⟨http://ftp.orca.med.or.jp/pub/tokutei/doc/tokutei_basic_spec_080314.pdf⟩ (2013.02.07).
[5] 日医特定健康診査システム （仮称） 詳細設計書 - 日本医師会，入手先 ⟨http://ftp.orca.med.or.jp/pub/tokutei/doc/tokutei_detail_080314.pdf⟩ (2013.02.07).
[6] 大杉直樹，並川顕，小橋哲朗，重木昭信，木谷強，山本修一郎: 記述漏れと曖昧な表記の防止を目的とした要件定義書の第三者スコアリングに向けた試み", ソフトウェア品質シンポジウム 2009(2009)
[7] M. Wilson, H. Rosenberg, E. Hyatt, Automated analysis of requirement specications", In Proceedings of the Nineteenth International Conference on Software Engineering, pp. 161-171(1997)
[8] H. Yang, A. Willis, A. Roeck, B. Nuseibeh, Automatic Detection of Nocuous Coordination Ambiguities in Natural Language Requirements." Proceedings of International Conference on Automated Software Engineering pp. 53-62(2010)