コーパスに基づくがん用語集合の作成と評価

中川 晋一[†,††,†††]・内山 将夫[†]・三角 真[†,†††]・島津 明[††]・酒井 善則[†††]

がん患者に対する情報提供の適正化のため，がん情報処理を可能にする言語基盤であるがん用語辞書を，医師による人手で作成した．権威あるコーパスとして国立がんセンターのウェブ文書を用い，延べ約 2 万 6 千語を収集し，用語候補の集合 Cc（Cancer Terms Candidate：語彙数 10199 語）を得た．10 種のがん説明用コンテンツを対象とした Cc の用語の再現率はそれぞれ約 95%以上であった．次に一般語やがん医学用語との関係と用語集としての整合性から用語選択基準（T1：がんそのものを指す，T2：がんを想起させる用語，T3：T2 の関連語，T4：がんに関連しない語のうち，T3 までを採用する）を作成し，Cc に対して適用，93.7%が基準に合致し 690 語を削除，9509 語をがん用語 C として選択した．選択基準に従って作成した試験用ワードセットを医師に示すことで，用語選択基準を評価した．その結果，T1 と (T2, T3, T4) の 2 つに分割した場合と (T1, T2), (T3, T4) 分割した場合で一致係数 κ が約 0.6，T1, T2, (T3, T4) の 3 つに分割した場合は約 0.5 であり，選択基準を明示せずに単に用語選択を行った場合の κ 値 0.4 に比べて高値であったことから，本研究で提案するがんとの関連性に基づいた用語選択法の妥当性が示された．さらに，既存の専門用語選択アルゴリズムにより得られた用語集合 (HN) と本研究で得られた用語集合 (C) を比較したところ，HN での再現性は 80%以上と高値だが，精度は約 60%であり，本研究のような人手による用語選択の必要性が示された．以上のことから，専門性の高い，がんに関するような用語集合を作成する場合，本研究で行った，信頼性の高いコーパスを用い，専門家の語感を信用して，中心的概念からの距離感を考慮した用語選択を行うことにより，少人数でも妥当性の高い専門用語集合の作成が可能であることが示された．

キーワード：がん用語，専門用語辞書，医学用語

Establishment of Corpus-based Cancer Specific Term Set and its Characteristics

Shin-ichi Nakagawa[†,††,†††], Masao Utiyama[†], Makoto Misumi[†], Akira Shimazu[††] and Yoshinori Sakai[†††]

For providing the appropriate cancer information to patients, we made the Corpus-based Cancer Term Set as the basic linguistic infrastructure for analyzing cancer contents. The specific terms of cancer was carried out by the qualified medical doctors by cutting out each word using the whole web contents of the National Cancer Center

† 独立行政法人情報通信研究機構, National Institute of Information and Communications Technology
†† 北陸先端科学技術大学院大学情報科学科, School of Information Science, Japan Advanced Institute of Science and Technology
††† 東京工業大学大学院理工学研究科, Graduate School of Science and Engineering, Tokyo Institute of Technology

as the authorized corpus. Out of over 26,000 words that were carried out, 10,199 terms were finally collected as the Cancer Terms Candidate (Cc.) This term set covers 96.5–99.5% of 10 different kinds of cancer content, which is enough for analysis. Considering the contrast between this cancer word set and other word set, such as general words, general medical words and proper nouns, the Cc was investigated based on selection standards. As a result, 93.7% terms of Cc was selected into the new word set "C." Secondly, based on the relationship between general terms and cancer/medical terminology, as well as on the consistency of the glossary, the selection criteria (T1: Cancer itself, T2: Terms directly related to cancer, T3: Terms related to both T1 and T2, and T4: Terms of unclear relations to cancer) were proposed. As they were adapted to Cc, 93.7% met the criteria, 690 words were removed, and 9,509 were selected as the C word in terms of cancer. These terms were selected according to the criteria to create the word set for doctors to test, which indicates that the criteria for selection were indirectly evaluated. As a result, in two cases where the word set was split into T1 and (T2, T3, T4,) and where it was split into (T1, T2) and (T3, T4), coefficient of contingency, "κ," was 0.6. And in case where into the word set was split into T1, T2, (T3, T4) was 0.5. And in case where into the word set was split into T1, T2, (T3, T4) was 0.5. These "κ" values were higher than in the different test; making the simple question "Cancer word or not." Thus, the selection and classification of T1 and T2 terminology is plausible. Furthermore, the comparison analysis of detected words were performed for original several cancer corpus using HN : (auto-specific-word-selecting algorithm (Gen-Sen-Web)) and C. As the result, the recall rate of HN for C was around 80%, however the precision rate of HN for C was around 60%. Thus, these automatic word selecting methods are useful for evaluation of consistency for C. However, the reducing the ignore words selection must be required for those systems. Therefore, it was suggested that this method enabled us to create a low-cost, feasible cancer-specific term set. Thus, the selection and classification of T1 and T2 terminology is plausible. Therefore, it was suggested that this method enabled us to create a low-cost, feasible cancer-specific term set.

Key Words: *Cancer Words, Special Word dictionary, Medical Words*

1 はじめに

がんの患者や家族にとって，がんに関する情報（以下，「がん情報」と呼ぶ）を知ることは非常に重要である．そのための情報源として，専門的で高価な医学書に比べて，ウェブ上で提供されているがん情報は，容易に入手可能であり，広く用いられるようになってきている (山室 2000; 野口 2000)．これら Web で公開されているがん情報は，良質で根拠に基づいたものばかりではなく，悪質な商用誘導まで存在する (Wendy A. Weiger，坪野 2004; Humphrey and Miller 1987)．このような多量のがんに関する文書の中からその文書が何を述べているかの情報を抽出

し，良質ながん情報を選別し取得されるがん情報の質を向上させることが求められている．このように，がんに関する文章について，自然言語処理を適用することにより，がんに関して有用な結果を得るための情報処理を，本稿では，がん情報処理と呼ぶ．

がん情報処理のためには，がんに関する用語（以下，がん用語と呼ぶ）の網羅的なリスト，すなわち，網羅的ながん用語集合が必要である．なぜなら，もし，網羅的ながん用語集合が存在すれば，それを利用することにより，がんに関する文書の形態素解析や情報検索等のがん情報処理の精度が向上することが期待できるからである．しかし，現状では，内科学や循環器学等の分野の用語集合は，それぞれの関連学会により作成されているが，がん用語集合は存在しない．そのため，本研究では，がん用語集合を作成するとともに，がんだけでなく，がんとは別の分野における用語集合の作成にも適用できるような，用語集合作成法を提案することを目標とする．

高度ながん情報処理の例としては，「胃がん」や「肺がん」などの単純な検索語から検索エンジンを用いて得られたコンテンツが，一体，どのような意味を含んでいるのかを推定することなどが想定できる．そのような処理のためには，「胃がん」や「肺がん」などのがんの病名だけをがん用語としていたのでは不十分である．少なくとも，「肝転移」や「進行度」のようながんに限定的に用いられる語から，「レントゲン写真」や「検診」のように，がんだけに用いられるわけではないが関連すると思われる語もがん用語とする必要がある．なぜなら，「胃がん」や「肺がん」で検索した文書は，既に，「胃がん」や「肺がん」に関係することは明らかであるから，そこから更に詳細な情報を獲得するには，「胃がん」や「肺がん」よりも，もっと詳細な用語を利用する必要があるからである．

このように，がん情報処理のためには，「胃がん」や「肺がん」等のがんに関する中核的な用語だけでなく，がんに関連する用語や周辺的な用語も網羅的に採用すべきである．ただし，「網羅的」といっても，がんとの関連度が低すぎる語をがん用語集合に加えるのは，望ましくない．そこで，病名などの中核的意味を示す用語から一定以内の関連の強さにある用語のみから，がん用語集合を作成し，それ以外の語に関してはがんとの関連性が低いと考える．

このような関連の強さに基づくがん用語集合を作成するためには，まず，「がん」という疾患の性質を考慮する必要がある．「がん」は図 1 のように，胃がん，肺がんをはじめとする複数の疾患群（50 個以上の疾患）の総称であると同時に，他の疾患とも関わりがある．例えば，図 1 の下部分に示したタバコは，肺がんの直接のリスク要因であることが知られているが，それだけでなく，動脈硬化を引き起こし，心筋梗塞や脳梗塞などの成人病を起こす危険因子としても知られている．ただし，タバコによって引き起こされる動脈硬化が原因で起こる心筋梗塞や脳梗塞は，直接肺がんとは関係しない．そのため，「タバコ」はがんに関連するが，「心筋梗塞」や「脳梗塞」はがんに関連しない．

また，図 1 の上部分に示した肝障害に関連する疾患と密接に関係する「肝がん（肝臓がん）」

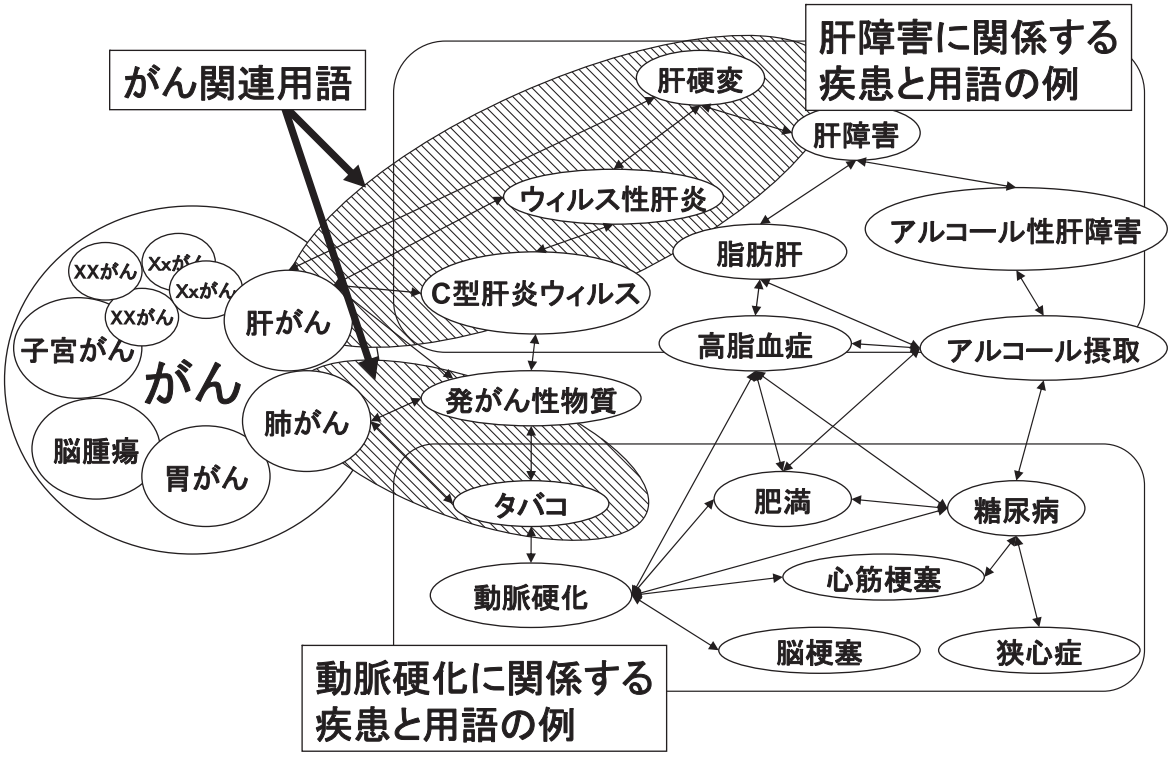

図 1 がんとがんに関する疾患の関係の例

は，肝硬変やウィルス性肝炎から直接発病する場合もある．そのため，肝硬変やウィルス性肝炎は，がんではないが，がんに関連する疾患であり，これらの内容が記述されているコンテンツは，がん関連用語を含む可能性が高い．そのため，肝がんに関連する用語候補を得るためには，図 1 の上の斜線で示した部分である「がん関連用語」を収集する必要がある．つまり，肝がんに直接関係する用語だけではなく，肝硬変やウィルス性肝炎などの関連する疾患に関係する用語であっても，肝がんに間接的に関係する用語は含める必要がある．(「がんに関係する」ということの定義については，3 節で詳述する.）さらに，がんにおける用語の範囲は，それぞれのがんにより異なるためアプリオリな定義を行うことは困難である．そのため，内省により用語集合を作成するのではなく，実際に存在するコーパスから用語を収集することが望ましい．

がん用語の一部は，例えば「リンパ節」や「転移」のように，一般用語辞書（例えば ChaSen 用の ipadic ver. 2.7.0）や，医学用シソーラスである MeSH(National Library of Medicine 2006) にも含まれている．しかし，これらに含まれるがん用語には，がんに関する用語であるとの説明がないため，これらの用語からがん用語を自動的に選択することはできない．また，がんに関するテキストから，専門用語抽出アルゴリズム (Nakagawa 2000; 佐藤，佐々木 2003) を利用

して，がん用語の候補を抽出することも考えられるが，我々の予備実験および 4.5 節の実験によると，このような候補には，がん用語以外のものも大量に含まれる．そのため，既存の一般用語辞書や専門用語抽出アルゴリズムを利用して用語候補を抽出したとしても，妥当な用語集合にするためには，人手によるがん用語の選別が不可欠である．この選別における問題は，選別の妥当性を確保することである．さらに，選別の対象であるがん用語の候補集合が，なるべく多くのがん用語を網羅していることを保証する必要もある．

がんに限らず，ある分野の用語集合の網羅性と妥当性を保証するためには，内科学や循環器学等の医学の各分野における用語集合について 2 節で示すように，学会単位で多大な人手と時間を費やして作成することが考えられる．しかし，これには多大なコストがかかる．そこで本研究では，相対的に低コストで，網羅的で妥当ながん用語集合を作成するために，まず，国立がんセンターの Web サイト (国立がんセンター http://www.ncc.go.jp/index.html) のコンテンツをコーパスとして，がん用語の語感を持つ医師に候補語彙を切り出させ，がん用語候補集合 (Cc: Cancer Term Candidates) を網羅的に作成する．この国立がんセンターのコンテンツは，同センターががんに関するわが国の最高権威の診療機関であること，50 種類以上のわが国の国民の罹患する可能性のあるほぼ全てのがんに関する記述があることから，がん用語に関する信頼性と網羅性が確保できると考える．なお，国立がんセンターの Web サイトのコンテンツの信頼性に関して 3.1 節，がん用語の切り出しの一貫性に関して 3.2 節でそれぞれ検討する．

このように本研究では，用語集合の切り出し元とするコーパスの医学的内容の信頼性と，記述されている内容の網羅性は十分と仮定して，用語候補集合 (Cc：Cancer Term Candidates) を作成する．最初の切り出しの段階では，医師の語感に基づいて，用語候補をできるだけ網羅的に広く収集することによって，初期段階における用語の漏れを防ぐ．次に，これら用語候補の特徴から，がん用語の選択基準を作成し，この基準に基づいて，Cc からがん用語集合 (C: Cancer Terms) を抽出する．最後に，他の医師に選択基準を説明し，評価用の用語候補を分類してもらうことにより，選択基準の妥当性を評価する．ここで，この選択基準は，上で述べたように，病名などの中核的意味を示す用語から一定以内の関連の強さにある用語のみを選ぶための基準である．

なお，2 節で示すが，わが国では医学のうち内科学や循環器学に関する用語集は存在するが，がん用語集はなく，本研究で作成するがん用語集は，それ自体が新規である．さらに，本研究では，がんだけでなく，他の分野の用語集合の作成にも適用できるような用語集合作成法を提案することを目標とする．なお，関連して，コーパスに基づいて辞書を作成したものとして COBUILD の辞書等があるが，医学用語をコーパスに基づいて収集し評価した例はない．

2 従来研究

まず，既存の公開されている医学用語集を分類し，本研究で作成するがん用語集合との相違点について述べる．次に，内科学と循環器学を例として，これらの用語選択について述べ，がん用語集合の要件について検討する．

2.1 既存の医学用語集

表 1 に現在公開されている医学用語集の例を示す．これら用語集を，公開されている内容に基づいて，作成者，対象領域，公開方法，見出し語数，各用語に対する説明の有無，用語選択の基準公開の別，主な用途の 7 つの観点から分類した．これより，これらの用語集は和英の用語の統一を目的とした対訳辞書（対訳共有）と，概念の共有を目的とした事典の形式（知識共有）をとるものがあることが分かる．

これらのうち最大のものは，1 の日本内科学会が編集した用語集だが，この用語集は学会誌などでの学術用語としての用語を統一するために編纂されたものであり，英語—日本語の対訳辞書である．内科学は医学全般の疾患を網羅的に扱う分野であり，がん用語もこの用語に含まれると考えられるが，各用語の用例と用語別の出典が明示されておらず，がん用語かどうかを機械的に判定することはできない．同様に 2, 4, 5, 6, 7, 8, 9, 10, 11, 12 の用語集には用語の説明と出典がないため，収録されている用語の選択基準が不明である．13 の国立がんセンターの公開する「がんに関する用語集」は，患者や家族に対して，医療関係者が行う用語の理解を助けることを目的にまとめられたものであるが，項目数は約 220 であり，本研究の目的とする，網

表 1 公開されている医学用語集の例

番号	名 称	作成者	対象領域	公開方法	見出し語数	説明	基準	用途
1	内科学用語集 [1)]	日本内科学会	内科学	書籍	35,000 (EN-JP)	なし	公開	対訳共有
2	循環器学用語集 [2)]	日本循環器病学会	循環器	書籍	5,631 語 (En) 525 語（略語） 5,885 語 (JP)	なし	公開	対訳共有
3	Dictionary of Cancer Terms[3)]	U.S. National Cancer Institute (U.S.)	がん	Web	5236 (En)	あり	非公開	知識共有
4	救急医学	日本救急医学会	救急医療	Web	300 (JP)	あり	非公開	対訳共有
5	国際保健用語集	日本国際保健研究会	国際保健	Web	88 (JP)	あり	公開	知識共有
6	寄生虫学用語集	日本寄生虫学会	寄生虫学	Web	2900 (En-JP)		公開	対訳共有
7	放射線腫瘍学用語集	日本放射線腫瘍学会	放射線腫瘍学	Web	1713 (JP)+ 266（略語）	なし	公開	対訳共有
8	最新解剖学用語集	個人	解剖学	Web	7580 (JP-EN)	なし	非公開	対訳共有
9	糖尿病用語集 [4)]	日本糖尿病学会	糖尿病	書籍	英和編 6911 語, 和英編 6704 語, 略語 851	あり	公開	知識共有
10	消化器病学用語集	日本消化器病学会	消化器	Web	4826 (JP)	なし	公開	対訳共有
11	行動分析学用語集	研究会（個人）	行動分析学	Web	196 (JP)	なし	非公開	対訳共有
12	心理学用語集	研究会（個人）	心理学	Web	418 (JP)	なし	非公開	対訳共有
13	がんに関する用語集	国立がんセンター	がん	Web	142（用語）， 76（病名など）	あり	非公開	知識共有

羅的な用語集ではない．

これらの用語集と比べ，表 1 の 3 の合衆国の NCI: National Cancer Institute の公開している “Dictionary of Cancer Terms” は，項目数が約 5200 であること，それぞれの用語に対する説明が行われており，説明内容に基づいて用語が選択されていると考えると，本研究の目的とする，網羅性が高くかつ選択基準が明確な用語集に最も近い．しかし，合衆国とわが国では，疾患分布が異なる（例えば，胃がんは合衆国では稀な疾患である）こと，医学的技術（内視鏡技術はわが国が開発した技術である），社会制度（社会保険の制度が全く異なる）など，様々な相違がある．そのため，必要ながん用語集合が異なる可能性が高い．以上のことから，本研究では，わが国のがん情報処理を可能にするがん用語集合を，実際の用例に基づいて作成する．本研究で作成する用語集合の特徴は，表 1 の 7 つの観点からは，(1) 作成者は本研究の著者ら，(2) 対象領域はがん，(3) 公開方法は未定，(4) 見出し語数は日本語約 1 万語，(5) 用語説明はなし，(6) 用語選択基準は公開，(7) 用途はがん情報処理である．

2.2 用語選択基準の例について

表 1 で示した医学用語集の中で，用語選択基準の例として，日本内科学会と日本循環器学会の基準を図 2 に示す．内科学や循環器学は，医学において歴史も古く，膨大な知識の整理の行われた結果教科書として長年にわたって出版され，それぞれ数千語以上の索引が掲載されており，これらが，用語集を作成する上での言語資源として活用されている．以下，内科学，循環器学を例としてそれぞれの選択基準について述べ，がんに関する用語選択基準について述べる．

内科学会用語集は，初版（ドイツ語見出しで約 9400 語）に始まり，平成 5 年に出版された第 4 版（英語見出しで約 27,000 語）を改訂した第 5 版（英語見出しで約 35,000 語）のものである．内科学は全ての医学の基礎的領域であり，約 100 年近い歴史もあることから，歴史的経過で基礎となる前版の用語に学会の選任した委員（第 5 版の場合，平成 6 年に各専門分野を代表する計 14 名の理事・評議員）で組織された内科学用語集改訂委員会が約 5 年をかけて，第 4 版に約 10000 語を追加する形で平成 10 年に第 5 版を出版した．なお，今回の改訂などでの用語の出典は明示されていない．また，用語選択の基準に関しては，「内科学の分野において使用される用語および関係の深い用語」として，専門委員会が選択したものである．

循環器学用語集は平成 20 年 3 月に改訂第 3 版が出版された．この用語集は，第 2 版（5,638 語）に，数冊の教科書の索引語（合計 19,037 語）を加えて基本用語リスト（24,675 語）を実務委員会が作成し，これに採用の可否を延べ 59 人の委員が判定し，14,858 語を選択し，最終案 14,476 語まで合計 3 回の選択を行ったと記述されている．このように循環器学は心筋梗塞から高血圧までの分野があり，それぞれの細分化された領域に専門家が存在するため，学会としてコンセンサスを形成するために，長い時間と多大な労力を必要としたと考えられる．これも内科学同様，「循環器学の分野において使用される用語およびこれに関連の深い用語を採録」とし

日本内科学会の内科学用語集での採録基準(用語採録の原則)

1. 内科学の分野において使用される用語およびこれと関係の深い用語を収録
2. 訳語はできるだけ各関連学会の用語集に従うようにしたが、一部内科学会独自の見解によった。
3. 固有名詞，人名のついた用語は内科学の教科書にみられるものだけを採録した。
4. 解剖学用語、薬品名の採録は前版にならい最小限度にとどめた。
5. 形容詞の採録は最小限にした。動詞は採録しなかった。
6. 新たに「内科で用いられる略語」を収録した。略語の選択にあたっては乱用を避け，国際的に通用することを基準とした。

循環器学用語集の採録基準

1. 循環器学の分野において使用される用語およびこれに関連の深い用語を採録することを原則とした。使用頻度が多くても，一般的日常用語，一般的医学用語は除外した。
2. 他の関連学会の用語集を出来るだけ尊重したが，循環器学の立場から若干修正を加えたものもあった。
3. 欧語としては，英語(米式綴り)を原則としたが，日常慣用されるラテン語，フランス語なども加えた。
4. 英和編，略語編，和英編とし，見出し語 各々，5,631語，525語，5,885語とした。
5. 名詞を主としたが，形容詞，分詞，過去分詞，動名詞等の形容語も最小限に採録した。
6. 旧称，古語は，原則として採録しなかった。
7. 英語の用語に対して，和訳はなるべく一語にするように努めたが，広く使われているものは止むを得ず複数語を採用した。

図 2 網羅的医学用語集の選択基準

ており，採録の用語範囲に関する明確な選択基準は示されていない．

2.3 がん用語集合の要件

これら内科学や循環器学などに比べ，がんは医学の比較的新しい領域である．この疾患群は，他の医学領域においても重要な疾患であり，各科別に専門家がそれぞれの専門領域のがんを診断治療してきた．そのため，「がん」という専門領域が認識されだしたのは新しい．日本臨床腫瘍学会が専門医試験を開始したのは平成 17 年からである．さらに基準となる索引を含む教科書も数少ない．そのため内科学や循環器学のように，教科書の索引をもとに，基本的な用語集合を作成するなどの方法で作成することは困難である．

さらに，これら他の疾患の用語集の用語選択の基準から以下のことがわかる．

- 用語集の特徴として
  - 主に研究者間での知識共有が目的である
  - 各用語の説明は与えられていない
- 用語選択の基準として
  - それぞれの学問分野に関係が深いと考えられる用語が収集されている
  - 用語の選定方法は，内科学や循環器学という大きな学問領域の中で，細分化され専門化された各領域の専門家が素案を作成し，全体をまとめている
  - 人体の部位を示す解剖学用語など，たとえその分野で多用される用語であっても，他の医学領域の用語や一般語は削除されている

これらに対して，本研究で作成するがん用語集合（C）は，がん情報処理を可能にするために，がんに関連する用語であれば，他の医学領域や一般語であっても，実際のコーパスに従って採用する必要がある．例えば，「大動脈周囲リンパ節」の場合，「大動脈」は解剖学用語，「周囲」は一般語，「リンパ節」も解剖学用語であり，この「大動脈周囲リンパ節」は解剖学用語である．ところが，「大動脈周囲リンパ節への転移」は，がん患者の状態を知るために重要ながん用語である．このような例に対応するために「大動脈周囲リンパ節」もがん関連用語として採用する必要がある．そこで，本研究では，従来研究とは異なり，解剖学用語や一般用語など範囲を限定せず，コーパスに出現するがん関連用語を網羅的に採用する．

また，従来の用語集では，用語の選定にあたって，専門家が素案となる用語候補集合を作成している．この部分は，本研究では，国立がんセンターのテキストから専門家ががんに関する用語を切り出すことに相当する．国立がんセンターのテキストを利用することにより，本研究においても，従来の用語集と同様に，がんに関係の深い用語の収集が可能になると思われる．なお，本研究で作成する用語集も，用語の説明は行わない．

以上のことから本研究では，国立がんセンターのテキストに出現し，かつ，専門家が，「がんに関係する」という語感によって選択した用語（名詞句）をがん用語の候補とする．その上で，各候補について，実際の用例を検討し，用語選択の基準を作成し，がん用語集合を求める．

3 がん用語候補集合 (Cc) の作成

本研究では，図 3 に示すように，がん情報に関する質の高いコーパスを選定し，これを対象として，網羅的ながん用語候補集合 Cc を作成する．得られた用語候補の意味と整合性から，用語選択基準 (SC：Selection Criteria) を作成する．このとき，Cc の抽出の一貫性が十分であるかどうかを元のコーパスを用いて調べる．Cc（がん用語候補集合）を C（がん用語集合）と Dc（削除用語集合）に分ける．C と Dc から Wt（評価用ワードセット）を作成，これを第三者に

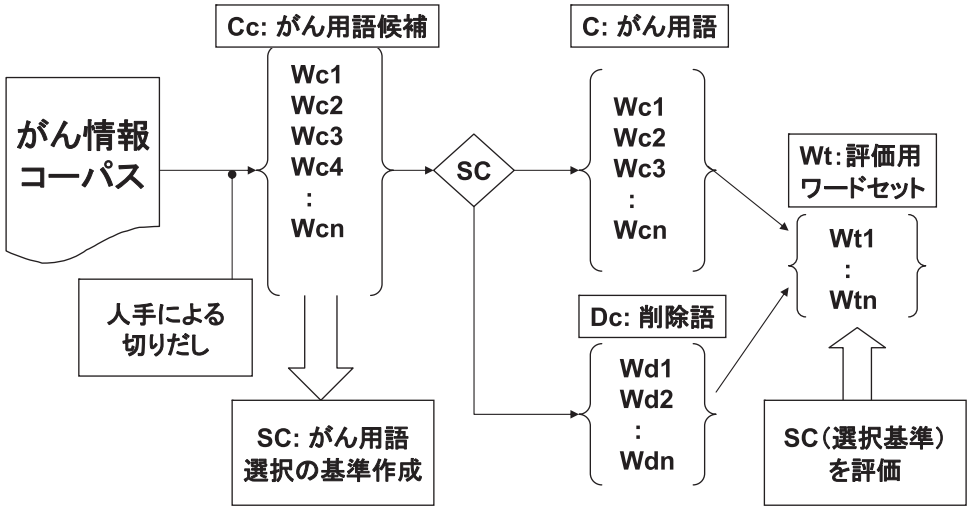

図 3 本研究でのがん用語の収集と評価の過程

提示し，SC の妥当性を検討する．以上により，一貫性をもって抽出された妥当ながん用語集合を作成できると考える．以下，元コーパスの選定，がん用語候補集合 (Cc) の作成，Cc に含まれた用語の実コーパス中での特徴，がん用語の選択基準について述べる．

3.1 コンテンツの選定

1 節で述べたように，本研究では，国立がんセンターの Web サイトのコンテンツを対象として，網羅的な用語候補集合 Cc を作成する．同センターは，わが国の医療情報提供に関しては，最も歴史が長く (Nakagawa, Kimura, Itokawa, Kasahara, Sato, and Kimura 1995)，その医療レベルは国内最高であるといわれている．同センターの「各種がんの説明のページ」には，患者や家族を対象として（1）病名，（2）概説，（3）症状，（4）診断，（5）病期分類，（6）治療方法，（7）期待される予後，（8）その他の説明，などの内容が記述されており，病名別に患者や家族がそれぞれの疾患の知識を系統的に得られるように工夫されている．また，当コンテンツは同センターが情報提供を開始した当初から，複数の専門家からなる Peer-review を導入し，コンテンツ内容に関する検討を行っている．データ量はテキストデータとして約 15 メガバイト，コンテンツ総量は 150 メガバイト（2006 年 10 月に大幅改訂後，疾患数が 53 から 59 に増加）であり，わが国で最大の情報提供を行っている．

この他に，がんに関する情報源として，世界的に有名なものとして，合衆国の National Cancer Institute (NCI) の PDQ(National Cancer Institute (NCI) 2007) がある．しかし，PDQ は白人などの欧米人に対しての記述であり，わが国の状況について記述したものではないので，わが国におけるがんの状況に対応した用語辞書を作成したいという目的には適さない．わが国でも

これを和訳し提供しているサイトも存在する ((財）先端医療振興財団 2009)．本サイトはがん情報としては有用だが，NCI の全てを翻訳しているわけではないので，量の点で国立がんセンターに及ばない．

また，医師に対する専門知識を提供する民間の有料コンテンツ (山口，北原 2004) は，標準的な医療の指針を与える知識ベースとして広く医師に使われている．しかし，医師をはじめとする医療関係者に対して，専門用語のみで，診断方法や治療方針を説明するものであり，ウェブ上で提供されているような一般人を対象とした情報提供内容をも対象としたい今回の目的には適さない．以上のことから，国立がんセンター (www.ncc.go.jp) を信頼性の高いコンテンツとして選択する．

3.2 がん用語候補集合の作成

がん用語候補集合の作成について述べる．まず，2006 年 6 月に同センターが情報提供を行っているコンテンツのうち，各種がんに関して説明を行っている 53 疾患分「各種がんの説明のページ」（約 1.5 メガバイト）を，それぞれの疾患の説明別にテキストファイルを作成した．それぞれの疾患別のテキストファイルに対して，医師免許を持ち，臨床経験のある専門家である本稿の第一著者が用語の切り出しを行った．なお，わが国の医師は 6 年間の専門教育を受け，内科や外科を問わず全分野から出題される数百問からなる国家試験に合格していることから，一定の語感を共有しており，その用語選択はある程度の代表性を持っていると考える．ここで，用語切り出しは，「がんに関連する用語」として認識する語を幅広く網羅するように，名詞句を中心として網羅的に切り出し，がん用語候補集合 Cc1 を作成した．得られた Cc1 の異なり語数は 3313 語であった．なお，これらコンテンツは，1 ページあたり約 2000 字から 15000 文字，ファイルサイズとして 10 K（5000 文字）から 30 K バイト（15000 文字）であり，切り出しに要した時間は 10 K バイトあたり約 30〜45 分であった．また，切り出し語数は 1 疾患あたり 150〜350 語であった．

Cc1 作成時の異なり語数の成長曲線 (growth curve) を図 4 に示す．図 4 の横軸が，抽出を行った「肺がん」「胃がん」などの疾患の数，縦軸がそれぞれの疾患のコンテンツから抽出した用語を足し合わせた時の異なり語数である．これによると疾患数の増加により，その疾患における固有の用語が増加分となるが，疾患数 30 個前後で増加率は鈍化しており，50 個前後でゆるやかになっている．これは，がんに関する共通語彙の存在によると思われる．例えば，「CT 検査」や「手術」「化学療法」などは，多くのがんで使用される語である．これに比べて疾患固有のものは多くないため新規語の増加率が鈍化すると考えられる．本図からも，53 疾患の約半数の 30 疾患前後で約 80%の用語が出現している．これにより，用語の切り出しの対象とするコーパスの網羅性（本研究の場合，コーパスの説明しているがん情報の内容）を大きくすることによって，がんに関連する用語を十分網羅的に収集することができると考えた．

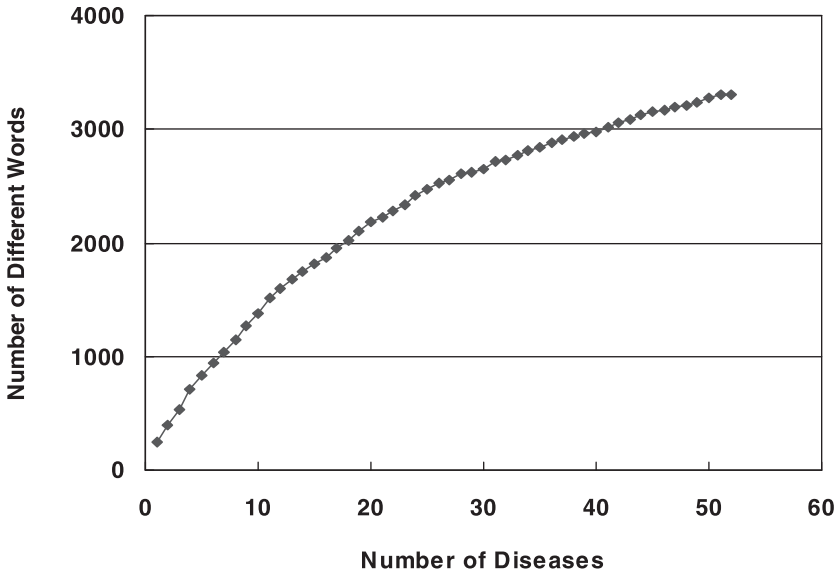

図 4 Cc1 作成時の疾患数別 Growth Curve

なお，この国立がんセンターのコンテンツは，2006 年 10 月に行われた大幅改訂に伴い，疾患数も追加され，53 から 59 となった．そのため，より網羅性を高くすることと関連用語も含めた用語収集を目的とし，疾患別だけではなく全ページ（データ量は合計約 250 メガバイト，テキストとして容量約 15 メガバイト）から，延べ 29,500 語を切り出し，用語候補集合 Cc2（9,451 語）を得た．このようにして得られた用語候補集合 Cc1, Cc2 の和をとり，がん用語候補集合 Cc（Cancer term Candidates, 10199 語）を得た．

3.3 Cc に含まれた用語の実コーパス中での特徴

ここでは，Cc に含まれる用語の特徴や性質を明らかにするために，実際のコーパスからどのような用語が収集されたか，それらはどのような特徴を持ち，どのようにがんと関連性があるのかについて述べる．

3.3.1 実コーパスからの Cc への用語収集の例

本研究で用いた実コーパスの例（図 5 の文例 1）から，専門家（本稿の第一著者）が，がんに関連すると思われる用語候補（Cc に含まれる語）を，網羅的に選んだ例を図 5 の下線により示す．

文例 1 の文章全体は膨大であり，肺がんの（1）病名，（2）概説，（3）症状，（4）診断，（5）病期分類，（6）治療方法，（7）期待される予後，（8）その他の説明が含まれる．ここでは，以下の文例 1-1, 1-2, 1-3 の 3 つの部分を用いて説明する．

文例 1-1: （1）病名と（2）概説を行っている部分

文例1

═══ http://ganjoho.ncc.go.jp/pub/med_info/cancer/010202.html

文例1－1

肺がん はいがん　　掲載日:1995/11/06 更新日:2006/10/01
1．肺がんとは　1）肺の構造と働き 肺は呼吸器系の重要な臓器であり、心臓、気管、食道などからなる縦隔（じゅうかく）という部分を挟んで胸の中に左右 2 つあり、左肺、右肺と呼ばれています。右肺は葉と呼ばれる 3 つの部分からなり（上葉、中葉、下葉）、左肺は右肺よりわずかに小さく上葉と下葉に分かれています。肺は身体の中に酸素を取り入れ、二酸化炭素を排出します。空気は口と鼻から咽頭・喉頭を経て気管を通り、気管支と呼ばれる左右の管に分かれ左右の肺に入ります。気管支は肺の中で細気管支と呼ばれるより細い管に分枝し、木の枝のように肺内に広がり、末端は酸素と二酸化炭素を交換する肺胞と呼ばれる部屋となっています。

：
（中略）
：

文例1－2
4）肺がんの組織分類 肺がんは、小細胞がんと非小細胞がんの２つの型に大きく分類されます。　非小細胞肺がんは、さらに腺がん、扁平上皮がん、大細胞がん、腺扁平上皮がんなどの組織型に分類されます。肺がんの発生しやすい部位、進行形式と速度、症状などの臨床像は多彩ですが、これも多くの異なる組織型があるためです。腺がんは、我が国で最も発生頻度が高く、男性の肺がんの 40%、女性の肺がんの 70%以上を占めています。通常の胸部のレントゲン写真で発見されやすい「肺野型」と呼ばれる肺の末梢に発生するのがほとんどです。肺がんの中でも他の組織型に比べ臨床像は多彩で、進行の速いものから進行の遅いものまでいろいろあります。次に多い扁平上皮がんは、男性の肺がんの 40%、女性の肺がんの 15%を占めています。気管支が肺に入った近くに発生する肺門型と呼ばれるがんの頻度が、腺がんに比べて高くなります。大細胞がんは、一般に増殖が速く、肺がんと診断された時には大きながんであることが多くみられます。　小細胞がんは肺がんの約 15～20%を占め、増殖が速く、脳・リンパ節・肝臓・副腎・骨などに転移しやすい悪性度の高いがんです。

：
（中略）
：

文例1－3
また、受動喫煙によって、肺がんのリスクが高くなるという科学的根拠は十分あると評価され、受動喫煙がない者に対し、20～30%程度高くなると推計されています。　その他、アスベスト、シリカ、砒素（ひそ）、クロム、コールタール、放射線、ディーゼル排ガスなどの職業や一般環境での曝露（ばくろ）、さらに、石炭ストーブの燃焼や不純物の混ざった植物油の高温調理により生じる煙（中国の一部地域）、ラドンなどによる室内環境汚染も、肺がんのリスク要因とする根拠は十分とされています。

（後略）

図 5　がん情報提供コンテンツからの用語候補切り出し例（肺がん）

文例 1-2: 肺がんの組織分類 （種類による分類）

文例 1-3: 肺がんの原因に関して説明している部分

本例では，網羅的用語収集のため，「肺がん」などの明らかにがんに関係する名詞句だけではなく，「肺がんのリスク」などの名詞句や，「空気」「ラドン」などの一般名詞，内科学会の選択基準では除外されている「肺」「気管」など解剖学用語も選んでいる．結果，文例 1-1 から 25 語，1-2 から 33 語，1-3 から 21 語の合計 79 語が切り出された．

3.3.2 Cc に含まれた用語候補の種類

得られた用語を理解しやすくするために，表 2 に各用語候補の重複を除き，どのような医学上の概念に関連するかを（医師免許をもつ第 1 著者が）想起したものを「種別」として付記したものを示す．これら用語の「種別」を付加することは，理解を助けるためや，用語間の整合

表 2 文例 1 からがん用語候補として切り出された語と種別

用語候補	種別	用語候補	種別	用語候補	種別
がん	病名	肺	解剖	進行形式	臨床
肺がん	病名	肺の構造	解剖	レントゲン写真	臨床
小細胞がん	病名	咽頭	解剖	臨床像	臨床
腺がん	病名	右肺	解剖	症状	臨床
腺扁平上皮がん	病名	葉	解剖	診断	臨床
大細胞がん	病名	上葉	解剖	進行	臨床
非小細胞がん	病名	中葉	解剖	増殖	臨床
非小細胞肺がん	病名	下葉	解剖	転移	臨床
扁平上皮がん	病名	左肺	解剖	発生頻度	臨床
アスベスト	一般	喉頭	解剖	部位	臨床
クロム	一般	気管	解剖	肺門型	臨床
コールタール	一般	気管支	解剖	肺野型	臨床
シリカ	一般	細気管支	解剖	受動喫煙	臨床
ディーゼル排ガス	一般	肺胞	解剖	肺がんのリスク	臨床
ラドン	一般	縦隔	解剖	科学的根拠	臨床
煙	一般	食道	解剖	推計	臨床
空気	一般	心臓	解剖	室内環境汚染	臨床
酸素	一般	臓器	解剖	肺がんのリスク要因	臨床
植物油の高温調理	一般	リンパ節	解剖	根拠は十分	臨床
石炭ストーブの燃焼	一般	肝臓	解剖	呼吸器系	臨床
二酸化炭素	一般	胸部	解剖	肺内	臨床
曝露	一般	骨	解剖	末端	臨床
不純物	一般	脳	解剖	組織型	検査
放射線	一般	副腎	解剖	悪性度	検査
砒素	一般	肺の末梢	解剖	組織分類	検査

性を調整する場合に有用である．例えば，「病名」に分類される語はがん用語の中心的な概念を示すことが多い．これに対して「一般語」は，がんに関係する用語ばかりとは限らないなどである．これらの種別を Cc の用語全てに付加することは，膨大な労力を必要とし，異なる専門家間のコンセンサスを得ることが困難であるため今後の課題とするが，用語が一般的な医学的知識の中でどのような範疇に入るかを大まかに理解することが可能になること，各種別内での用語の比較が可能になり，整合性をとりやすくなることから利便性が高い．

表 2 に出現した語では，「肺がん」は病名（病名）であり，「肺」，「肺の構造」，「気管支」は人体の特定の場所を示す解剖学用語（解剖）である．また，「呼吸器系」，「肺内」は臨床医学で使用される用語（臨床）であり，「酸素」，「空気」は一般語（一般）である．これらから，表 2 の Cc に含まれる語は，これら 5 つの種別の想起が可能であり，9 個の病名，25 個の解剖用語，22 個の臨床用語，3 つの検査用語，16 個の一般語に分類された．

3.3.3 Cc に含まれる各語のがんとの関連性分類の必要性

Cc に含まれる語とがんとの関連性に関して検討する．前節で述べたように，表 2 中の「病名」は一様にがんを示すと考えられるのに対し，「一般語」はがんに関係する用語ばかりとは限らない．このように Cc は，がんと関連性が明確な「肺がん」のような語から，「空気」，「酸素」などのがんとの関連性を想起することが難しい語までを含んでいる．

このように Cc に含まれる各語は，それぞれ，がんとの固有の関連性を示す．「関連性」とは，その語が，がんを起点として考えた時，「がんそのものを示す強い関連性を持つ用語」（表 2 の語では「肺がん」「扁平上皮がん」）を起点として，「転移」や「悪性度」などの「がんを想起させる用語」から，「末端」「推計」「酸素」など関連を想起しにくい語を「関連しない語」とする場合における関連の強さを示す順序尺度である．この尺度を用いることによって，本研究では，がん用語集合を，従来研究のように各語が集合に「含まれる」か「含まれない」かの二者択一ではなく，各語に想起される関連性の強弱を示すタグを付加しておくことによって，より広い範囲の用語を含めることが可能になる．また，このような「関連性の強弱」という主観評価で用語を分類する場合，つけられたタグを第三者による評価などで客観化する必要があるため，本研究では，4.3 節で複数医師により，関連性の強弱の一致の度合いを確かめる．

以下では，この関連性の強弱を表現するために，「ホップ」という概念を導入する．本稿では，がんに関連する語が，後掲図 7 のようなネットワーク構造となっていると想定する．そして，このネットワーク上において「がんそのものをさし示す用語」から離れるほど，がんとの関連性が弱くなると考えている．このとき，このネットワークにおけるリンクを，「がんそのものをさし示す用語」から 1 段階離れるごとに，関連性が 1 段階弱くなると想定し，その「がんそのものをさし示す用語」からのネットワーク上における距離を，インターネットとの類推から「ホップ」と呼ぶことにする．つまり，0 ホップ目が，「がんそのものを示す用語」であり，そこ

からリンクを 1 つたどるごとに，1 ホップずつがんとの関連性が弱くなると考える．

3.4 がんとの関連性による用語分類と選択基準

前節で述べた「がんとの関連性」（ホップ）によって用語を分類することにより，各用語の中心用語から関連用語という概念的な距離感が表現可能になる．ただし，これら関連性の分類には医学的知識が必要だが，医学的知識のみでは，一貫性を保っての用語分類は困難である．そこで，以下で述べる「がん用語の選択基準」に基づいて Cc の各語を分類する．

まず，Cc の各語を，がんとの関連性の強いもの（T1）から弱いもの（T4）までの 4 段階に分類する．これらは，T1（0 ホップ目）：がんそのものをさし示す語，T2（1 ホップ目）：がんを想起させる語，T3（2 ホップ目）：がんを想起させる語に関連する語，T4（3 ホップ以上）：がんとの直接の関連を説明しにくい語である．これらに基づいて，用語選択基準を図 6 のように

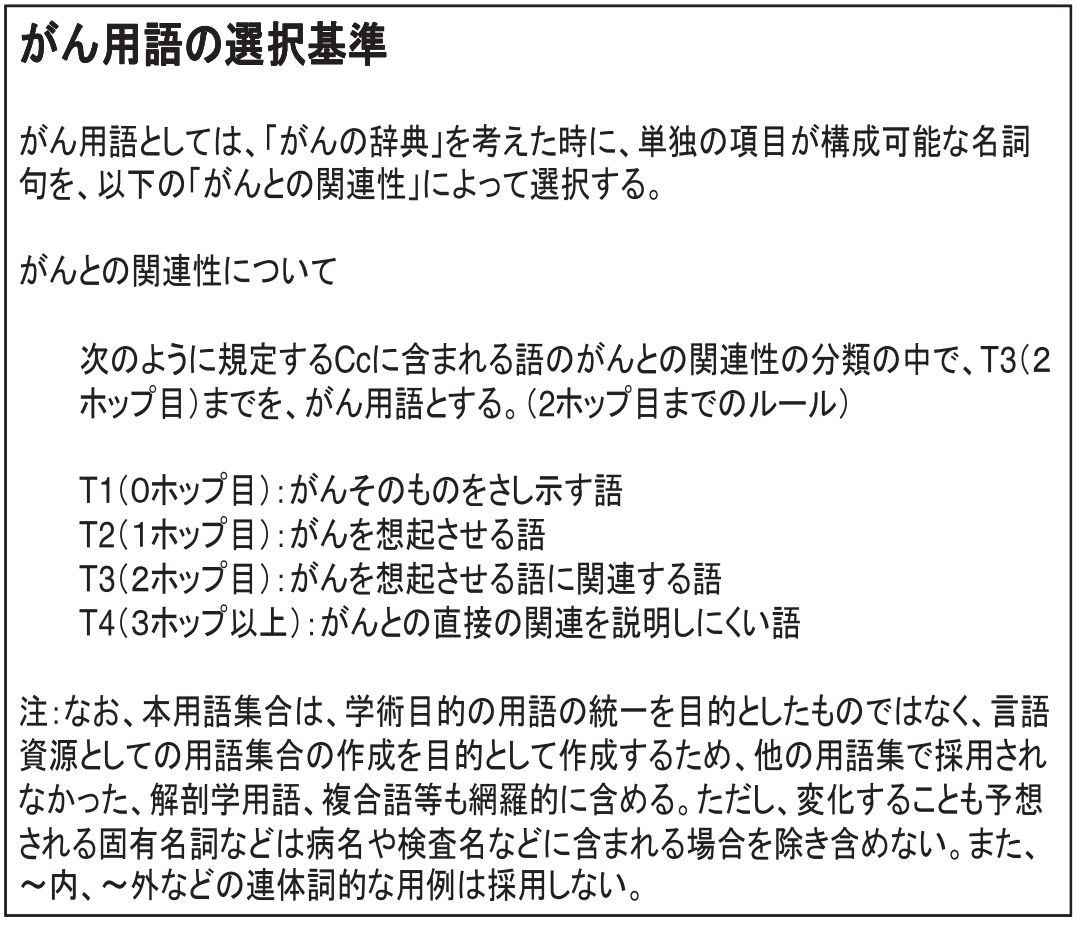

図 6 がん用語の選択基準

決める．そして，T1, T2, T3, T4 の用語候補の中で，「2 ホップ目までのルール」（単に「2 ホップルール」とも呼ぶ）により，T1, T2, T3 をがん用語とする．

本基準によって表 2 を T1 から T4 に分類した結果を表 3 に示す．文例 1 に出現した病名はすべてがんの病名であったため T1 に，また，解剖学用語は 3.4.3 節で示すように T3 に，臨床用語のうち「がん」という語句を含む語は T1, がんを想起させる「転移」や「進行」は T2 に，さらに，一般語の「空気」や「二酸化炭素」は T4 に分類できることがわかる．以下，T1 から T4 の各分類について説明する．

3.4.1 T1（0 ホップ目）：がんそのものをさし示す語

表 2 の用語のうち，「肺がん」「扁平上皮がん」など「がん」を含む「病名」は，文脈なしにがんに関することが明らかな語である．また，臨床用語である「肺がんのリスク」や「肺がんの

表 3 文例 1 から得た Cc 各用語のがんとの関連性による分類

用語候補	種類	関連	用語候補	種類	関連	用語候補	種類	関連
がん	病名	T1	肺がんのリスク	臨床	T1	リンパ節	解剖	T3
肺がん	病名	T1	肺がんのリスク要因	臨床	T1	咽頭	解剖	T3
小細胞がん	病名	T1	進行形式	臨床	T2	右肺	解剖	T3
腺がん	病名	T1	レントゲン写真	臨床	T2	下葉	解剖	T3
腺扁平上皮がん	病名	T1	進行	臨床	T2	肝臓	解剖	T3
大細胞がん	病名	T1	増殖	臨床	T2	気管	解剖	T3
非小細胞がん	病名	T1	転移	臨床	T2	気管支	解剖	T3
非小細胞肺がん	病名	T1	肺門型	臨床	T2	胸部	解剖	T3
扁平上皮がん	病名	T1	肺野型	臨床	T2	喉頭	解剖	T3
アスベスト	一般	T2	受動喫煙	臨床	T2	骨	解剖	T3
クロム	一般	T2	室内環境汚染	臨床	T2	左肺	解剖	T3
コールタール	一般	T2	臨床像	臨床	T3	細気管支	解剖	T3
シリカ	一般	T2	症状	臨床	T3	縦隔	解剖	T3
ディーゼル排ガス	一般	T2	診断	臨床	T3	上葉	解剖	T3
ラドン	一般	T2	発生頻度	臨床	T3	食道	解剖	T3
曝露	一般	T2	部位	臨床	T3	心臓	解剖	T3
放射線	一般	T2	科学的根拠	臨床	T3	臓器	解剖	T3
砒素	一般	T2	推計	臨床	T4	中葉	解剖	T3
煙	一般	T4	根拠は十分	臨床	T4	脳	解剖	T3
空気	一般	T4	呼吸器系	臨床	T4	肺	解剖	T3
酸素	一般	T4	肺内	臨床	T4	肺の構造	解剖	T3
植物油の高温調理	一般	T4	末端	臨床	T4	肺の末梢	解剖	T3
石炭ストーブの燃焼	一般	T4	悪性度	検査	T1	肺胞	解剖	T3
二酸化炭素	一般	T4	組織型	検査	T2	副腎	解剖	T3
不純物	一般	T4	組織分類	検査	T2	葉	解剖	T3

リスク要因」のように明示的に「がん」という句を含む複合語も，単なる一般語である「リスク」に「肺がんの」と限定されることによって，「病名」と同程度に，文脈なしにがんに関することが明らかな語となる．これらの，病名を含み，文脈なしにがんに関することが明らかな用語のことを本研究では，「T1：がんそのものをさし示す語」と呼ぶ．

このほかに，文例 1 には出現していないが，Cc に含まれる語である「白血病」，「リンパ腫」，「ボーエン病」，「リヒター症候群」のように「がん」を含まないがんの病名もある．また，「ATLL 細胞」，「骨髄腫細胞」，「経尿道的膀胱腫瘍切除術」等の語は，それぞれ「ATLL（急性 T 細胞性白血病）」，「髄膜腫」，「膀胱腫瘍」というがんの病名を含んでおり，これらも「がんそのものをさし示す語」である．

3.4.2 T2（1 ホップ目）：がんを想起させる語

肺がんの悪性度や性質を示す「転移」「悪性度」「進行形式」や，肺がんの最も古い診断方法の一つである「レントゲン写真」は，文中にこれらの用語が出現した場合，その文の意味が肺がんに関係することを連想させる語である．同様に，「肺門型」や「肺野型」はレントゲン写真に関する所見の記述で，肺がんの形状を示す語である．「受動喫煙」や「アスベスト」も肺がんの原因として重要であることが知られている．これら用語は T1 の，がんそのものをさし示す語ではないが，文中に出現した場合，その文脈ががんの意味を含むことが多い語である．このような語のことを，「T2：がんを想起させる語」と呼ぶ．

このほかに，「寛解」「寛解導入不応」「急性転化」「自家造血幹細胞移植」は，白血病にしか用いられない．また，「周囲臓器浸潤」，「遠隔転移」，「全身再発」などの，がんの状態を示す語も他の疾患で用いられることはほとんどない．抗がん剤として使用される「ブレオマイシン」「5-FU」などの薬剤名，がん特有の治療法である「自家骨移植」「ミニ移植」「温熱療法」も同様である．これらも「がんを想起させる語」である．

3.4.3 T3（2 ホップ目）：がんを想起させる語に関連する語

表 2 の「肺」，「咽頭」などの解剖学用語は，人体の特定の場所を示す語であり，解剖学者によって一義に定義されている名詞句である．2 章で述べたように，従来研究において，これらの解剖学用語は，用語の重複を避けるために，内科学や循環器病学の用語集には含まれなかった．しかし，がん情報処理の場合，2.3 で述べたように，解剖学用語の一部も網羅的にがん用語に含める必要がある．

これら解剖学用語は，例えば図 7 のように，「肺がん」から「転移」という 1 ホップ目の関連語を介して，「肺」や「リンパ節」を連想することが可能である．また，これらの連想関係は方向性を持つ．つまり，「肺がん」から「転移」という 1 ホップ目の関連語を介して，「肺」や「リンパ節」を連想することは可能であるが，これと逆方向の連想，すなわち，「肺」から「肺がん」

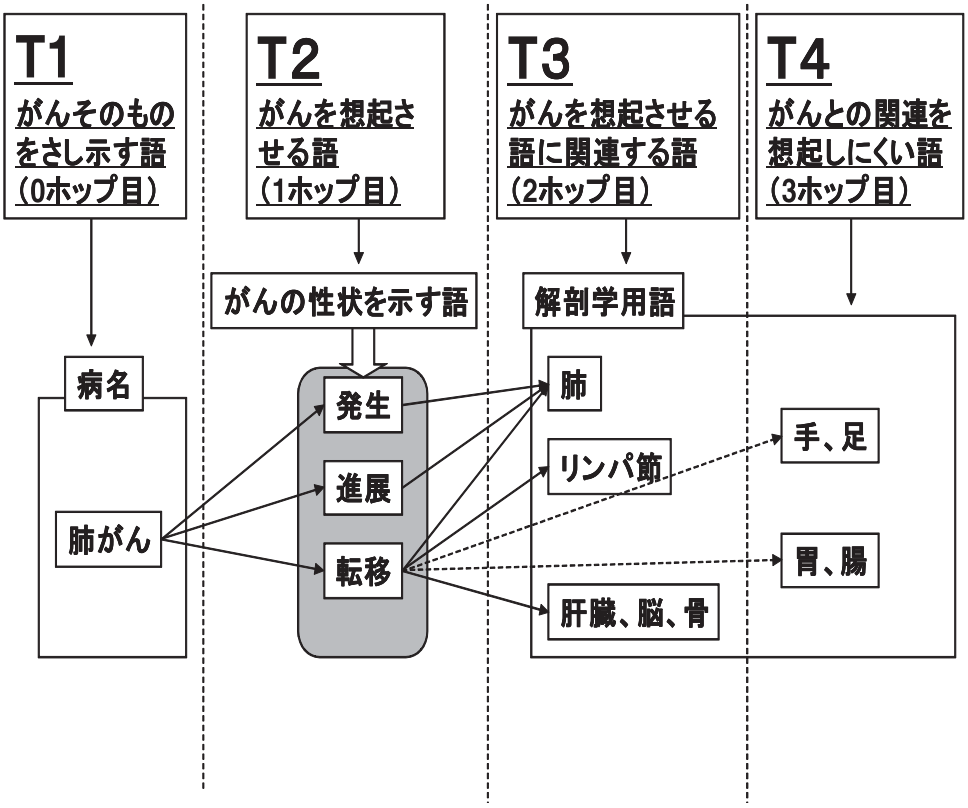

図 7　肺がんを起点とした場合の各種用語の関係例

あるいは「リンパ節」から「肺がん」を連想することは難しい．

このように，「肺がん」と「肺」の連想関係は，「発生」「進展」「転移」など，がんの性状を示す語の仲介によって可能であり，このような解剖学用語の一部はがん用語とすべきである．このように，何らかの中間的用語を介してがんに関連する用語も，がん用語と考えることができる．そこで，これらを「T3（2 ホップ目）：がんを想起させる語に関連する語」と呼ぶ．

なお，肺がんから見る場合，胃や腸は，「肺がんの胃や腸に対する転移はきわめてまれである」という医学的知識により，「転移」を仲介しても連想できない．そのため，「肺がん」を起点とした場合には，「胃」や「腸」は T3 ではなく，次節の T4（がんとの関連を想起しにくい語）に分類される．しかし，胃や腸は，肺がんとの直接の関係はないが，胃がんや大腸がんでは，胃がん―発生―胃，大腸がん―発生―大腸のような想起順で T3 に分類される．このように，解剖学用語は，図 8 に示すように起点とするがんによって，T3 になる場合もあれば，T4 になる場合もある．

胃がんや大腸がんから見た場合，明らかに胃や腸は発生部位である．このように，図 8 に示

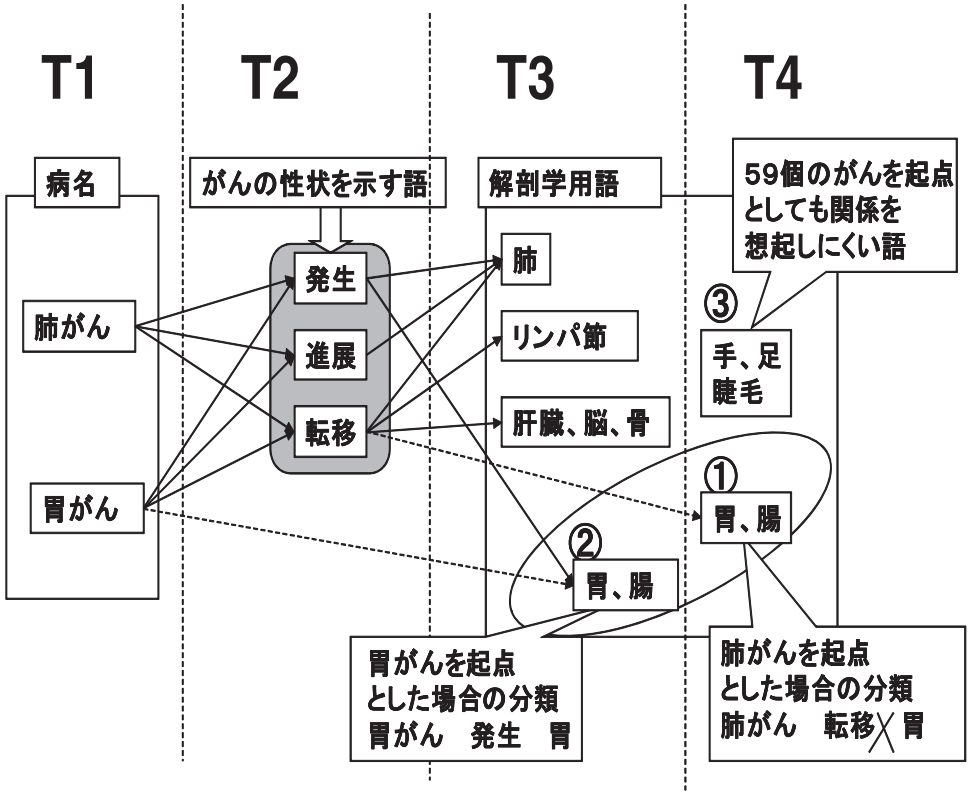

図 8 肺がんと胃がんを起点とした場合の「胃」の分類例

すように，①の肺がん—転移—胃が連想しにくいため，胃や腸は，肺がんからは T4 に分類されるが，②のように胃がんを起点とした場合，胃がん—発生—胃という想起順になるため，T3 に分類できる．この場合には，「胃」の分類は，肺がんを起点とした T4 ではなく，胃がんを起点とした場合の T3 に分類することとする．一般に，ある用語候補に対して，T1, T2, T3, T4 のうち，複数の分類が考えられるときには，最も関連性の強いものに分類する．

このように用語選択を行った場合，図 8 の③に示すように，元コーパスに出現した肺や胃以外の解剖学用語のうちで，例えば手，足，睫毛などは，本研究の元コーパスの対象とする 59 個のがんとの 2 ホップ以内として分類されないので，T4 となり，がん用語集合からは除外することが可能となる．このように，解剖学用語は基礎医学用語であるため，大きく医学用語という範疇でがんと無関係とは言えないが，すべての解剖学用語ががんと関係があるとは言えないため，それを適切に反映するために，なんらかのがんと 2 ホップ以内に関係する解剖学用語のみを T3 として分類することが有効であると考えた．

3.4.4 T4（3 ホップ目）：がんとの関連を想起しにくい語

以上の用語に対して，一般語である「酸素」や「二酸化炭素」は，肺の機能である呼吸機能に関連する語であって，肺がんには直接関係しない．これらは文例 1 の，「肺は身体の中に酸素を取り入れ，二酸化炭素を排出します.」という文で出現している．これは図 7 の解剖学用語である「肺」の機能を，肺がんと独立して説明している文章である．以上より，文例 1 を根拠とする限り，これらの「酸素」や「二酸化炭素」は，肺がんを起点とする場合，T3 である「肺」の関連語となり，「肺がん」の関連語とはいえない．以上より，これらの用語を「T4：がんとの関連を想起しにくい語」とする．

3.5 本研究で行った用語分類と用語選択基準について

以上のように，本研究では，まず，信頼できるがん情報が記述されているコーパスから，専門家の語感に基づいてがん用語の候補 Cc を抽出した．つぎに，得られた Cc の各語を理解しやすくすることや用語間の整合性を調整するために，各語の用例から「病名」「解剖学用語」「臨床用語」「一般語」「検査用語」などの種別を作成した．また，各語をがんとの関連性を想起しつつ分類した結果，T1（がんそのものをさし示す語），T2（がんを想起させる語），T3（がんを想起させる語に関連する語），T4（がんとの関連性を想起しにくい語）の 4 つに分類できた．これら「種別」とがんとの関連性を参考に，がん用語として，T1 から T3 までをがん用語 C として選択するという選択基準を規定した．これにより，各がん用語には，T1 から T3 までの分類がつけられているので，その語ががんとの関連性が強い語かどうかの情報を得ることができるので有用である．

ただし，3.4.3 節であげた肺がんと胃がんに対する胃の例のように，起点とする（0 ホップ目として）想起する語が異なれば，その語の分類が T3 か T4 かという揺らぎが必然的に生じることが予想される．そのため，本研究ではこれら用語については，がん用語集合の網羅性を確保するために，最も強い連関を示す分類に分類した．なお，それぞれの用語に対して，例えば胃なら，肺がんからは T4, 胃がんからは T3 のような情報を付加することも有用であると思われるが，これについては今後の課題である．

4 がん用語集合の特性評価

本節では，用語候補集合 Cc（10,199 語）を対象として，がん用語抽出の一貫性とがん用語選択基準の妥当性を検討する．

4.1 Cc の用語抽出の一貫性の検討

本節では，作成した用語候補集合 Cc が，抽出対象のテキスト中のがん用語候補を一貫性をもって抽出されたかを検討する．そのために，3 節で人手により用語候補抽出をした人と同一の人物（本稿の第一著者，抽出者と呼ぶ）が，約 1 年後に，3 節と同一のコーパスから，主要ながんのうち 10 疾患（ALL（急性リンパ急性白血病），腎細胞がん，膵がん，卵巣がん，肺がん，肝臓がん，グリオーマ，胃がん，大腸がん，乳がん）を対象として，3 節と同様に人手で用語候補を抽出した．次に，得られた用語候補と Cc を比較し，Cc が，2 回目に抽出した用語候補を（後で定義する）再現率高く抽出しているかを調べた．もし，この再現率が高ければ，Cc は，抽出対象のテキストから，抽出したい用語候補を一貫して抽出していると考えることができる．なお，用語候補抽出の一貫性を調べるためには，同一人物ではなく，他の人物による用語候補抽出の結果と比較することが望ましい．しかし，本稿の第一著者と同様に医師免許を持ち，臨床経験のある専門家で，かつ，用語候補の抽出に協力してくれる専門家を見つけることが我々にはできなかったため，同一人物による抽出の一貫性を調べた．

4.1.1 用語抽出の一貫性に関する数量的な検討

結果を表 4 に示す．表 4 の①は，それぞれの疾患別に抽出された用語候補集合の語数である．②は，それぞれの疾患別の用語候補集合で，Cc に含まれている用語候補の数である．これらより，③に示したように，見かけの再現率が算出される．「見かけ」とは，今回新たに人手で切り出した結果も，がん用語の網羅性を高くすることを意図していたため，今回抽出した用語候補すべてが，がん用語として適切かどうかは不明である．そのため，Cc にカバーされていない用

表 4 Cc（10,199 語）の用語抽出の一貫性に関する集計結果

病名	①専門家が手で抽出した語数	②：①のうち Cc に含まれた語数	③：みかけの再現率：②/①	④：①の中で Cc にない語数	⑤：④中で採用すべき語数	⑥：④中で不要だった語数	⑦：①のうち採用すべきだった語数：②＋⑤	⑧：真の再現率：②/⑦
ALL	368	291	0.79	77	12	65	303	0.96
Relancell Cancer	208	173	0.83	35	1	34	174	0.99
Pancreas Cancer	228	187	0.82	41	7	34	194	0.96
Ovaryan Cancer	280	224	0.80	56	5	51	229	0.98
Lung Cancer	475	396	0.83	79	22	57	418	0.95
Liver Cancer	296	256	0.86	40	10	30	266	0.96
Glioma	236	204	0.86	32	5	27	209	0.98
Gastric Cancer	541	437	0.81	104	19	85	456	0.96
Colon Cancer	517	432	0.84	85	7	78	439	0.98
Breast Cancer	346	302	0.87	44	10	34	312	0.97
10 疾患の平均値	350	290	0.83	59	10	50	300	0.97

ALL: 急性リンパ球性白血病，Renalcell Cancer: 腎細胞がん，Panceras Cancer: すい臓がん，Ovaryan Cancer: 卵巣がん，Lung Cancer: 肺がん，Liver Cancer: 肝がん，Glioma: グリオーマ（神経膠重），Colon Cancer: 大腸がん，Breast Cancer: 乳がん

語候補は，実際には，がん用語でない可能性もある．

そこで，用語の検討のため，抽出者に自然言語処理の研究者 1 名（第 2 著者）を加え，計 2 名で，抽出された用語候補の中で Cc に含まれていない用語候補を選別した（④）．それら用語候補の中で 3.4 節の基準に相当するかどうかによって真に必要な用語数⑤と，選択されたが不必要であった用語候補数⑥を求めた．これより，それぞれのテキストから抽出されるべきだった用語数⑦を求め，②を分子として求めた真の再現率が⑧である．Cc の再現率⑧は 0.94 から 0.995 であった．これらのことから，Cc は，元コーパス中の用語を十分に網羅していると考えられた．すなわち，Cc と本節での抽出結果とを比較した結果，Cc は，本節で抽出された用語候補を十分に網羅しているといえる．これより，Cc は，抽出対象のテキストから，抽出したい用語候補を一貫して抽出していると考えることができる．なお，表 5 には，表 4 の⑤として，Cc に含まれてはいなかったが，本検討によって採用すべきと判断した用語の例を示した．

4.1.2 再検討を必要とした各語に関する検討

表 4 の⑤（Cc に含まれなかったが，再度抽出時に採用すべきと判断した語）は，10 疾患全体で 98 語（1 疾患あたり約 10 語）であった．これらの例を表 5 に示す．また，これら 98 語を採用すべきと判断した 4 つの理由（R1 から R4）について以下に説明する．

4.1.2.1 R1（2 ホップルール）

3.4 節図 6 の「がん用語の選択基準」は，Cc の用語候補を整理する過程で確立されたものである．一方，3.2 節で述べた Cc の抽出時には，この選択基準は存在しなかったため，専門家が「がんに関連する用語」として認識する語を幅広く網羅するように Cc を作成した．つまり，「Cc の抽出基準」と「がん用語の選択基準」は異なるものである．

そのため，Cc の抽出時には，がんに関連しないと判断されたため切り出されなかった用語候

表 5 表 4 の⑤（Cc になかったが再検討時採用すべきと思われた）語数と例

理由	個数	用語例
R1	25	致命的，粒子線，焼灼，橋，減圧，手縫い法，術前療法，症状，腸液，閉鎖術，茎，日系移民，反対側
R2	69	1 年生存率，ステージ 1，病期 IB，病変の拡がり，病変の広さ，肺がん罹患者，肺門型の肺がん，肝障害度 B，漿膜への浸潤，進行速度，初診時白血球数，患者さんの権利，確立された治療法，著明な体重減少，消化管の再建，足のつけ根，体重の減少，膵がん罹患者
R3	3	lymphoblastic lymphoma，卵巣がん検診，噴門部がん
R4	1	上皮がん
計	98	

R1: 用語範囲，R2: 複合語，R3: 見落とし，R4: 表記のゆれ

補が，がん用語の選択基準に基づいて再検討をした結果として，用語として採用した方が良いと判断されるものがありうる．それらが表 5 の R1 として示されている．

例えば，「焼灼」の意味は「焼くこと」であるため，Cc 抽出時には一般医学用語と考えて切り出さなかったが，これは，肝臓がんや転移性の肝臓がんに限定して用いられるため，肝臓がんに対して T2 である．また，「橋」は，単に脳の一部を示す解剖学用語であると思われたため，Cc 抽出時には切り出されなかったが，神経膠腫が多発する部分であるため T3 と考えられる．また，「日系移民」も Cc 抽出時には一般名詞と考えたが，大腸がんの発症原因である食生活の欧米化と関係するので T3 に入れるべき語である．

このように，Cc の作成段階では「がんに関係するかどうか」という基準で切り出したため，がん用語選択基準である 2 ホップルールに照らすと用語であっても，Cc には採用されないものがあった．ただし，このような用語は 25 語と少数であるため，本研究のアプローチ，すなわち，まず専門家が「がんに関連する用語」として認識する語を幅広く網羅するように Cc を作成し，次に，そこから，がん用語選択基準に従って，がん用語を選択するというアプローチは有効であるといえる．

4.1.2.2 R2（文法）：切り出し時のゆれ

Cc 作成における用語候補の切り出し時において，一つの名詞句に対して複数の用語候補が考えられるときには，一つの用語候補のみを選択して切り出したが，その選択に揺れが生じた場合である．例えば，「1 年生存率」は，Cc の抽出段階では「生存率」を用語候補として選択したが，これは，「がんを発症後 1 年生存する率」の意味であり重症のがんで多用される用語であるため，本節では採用した．「ステージ 1」についても，Cc 抽出時には，がんの病期を示す「ステージ」を選択していたが，これは，「軽症のがん」の意味を示す用語であるため，本節では採用した．「消化管の再建」も，Cc 抽出時には「消化管」と「再建」に分かれて抽出されていたが，これは，「消化管の再建」という一つの単位で，胃がんの主な手術である胃切除術に関連する用語として利用されるものであるので，本節では採用した．このように，Cc の抽出において，一つの名詞句に対して複数の用語候補があるときに，どの名詞句を切り出すのかについては，後の見直しで採用すべき語もあることが分かった．このような用語は 69 語であった．

4.1.2.3 その他：R3（見落とし），R4（表記のゆれ）

以上の語の他に，明らかにがんに関連する語である「lymphoblastic lymphoma」（病名），「卵巣がん検診」「噴門部がん」（これら 2 語は文節中にがんを含んでいる）が Cc に含まれていなかった．これらは切り出し時の見落としによるものであると思われた．また，「上皮がん」は，通常「上皮内がん」や，「移行上皮がん」などが一般的であり，医学的には一般的ではないため Cc 抽出時には採用しなかったが，本例はコーパス側における「上皮内がん」の表記のゆれであ

ると思われるため，本節では採用すべきと考えた．

4.2 用語の削除と，がん用語集合 C の作成

3.2 節で得た用語候補の集合 Cc から妥当ながん用語集合 C を得るために，前節同様，医師免許を保有する有資格者 1 名と自然言語処理の研究者 1 名（第 1 著者と第 2 著者）が，がん用語の選択基準に基づき，がん用語とすべき用語の範囲と選択基準の整合性を整理しつつ，1 語 1 語を音読し，必要に応じて用例を参照して，がん用語かどうかを判断した．

この判断の結果，Cc のうち 9509 語をがん用語として採用し，690 語を除外した．表 6 には，除外した理由別の語数を示す．表 6 における「誤用」とは，元コーパス中で，「全脳照射」とすべきところを「全能照射」としていたなど，明らかに用語の使用が間違っている場合であり，「ミス」とは，用語候補の切り出しにあたって，用語の一部のみしか抽出しないなど，切り出しに失敗した例である．以下では，その他の理由である「固有名詞」「文法」「2 ホップ」について説明する．

4.2.1 「固有名詞」：固有名詞の削除例

3.4 節の「がん用語の選択基準」では，「固有名詞」はがん用語に含めないと述べた．固有名詞として削除例に挙げた「LSG」は Lymphoma Study Group（悪性リンパ腫研究会：わが国の悪性リンパ腫の治療法を共同研究として行っている団体），「ASCO」は “American Society of Clinical Oncology”（アメリカ臨床がん学会）ならびに「アメリカがん協会」は，がんの学会や研究会である．これらは，がんに関連する文書の中でも頻繁に出現する．しかし，「LSG」など研究会名称は，今後変更されることも予想される．また，「ASCO」や「アメリカがん協会」の呼称については，ASCO, American Society of Clinical Oncology, ASCO (American Society of Clinical Oncology), ASCO（アメリカ臨床がん学会），アメリカ臨床がん学会（協会）など，表記のゆれもある．このような，団体や研究グループなどの固有名詞は，「がんの辞典」を考える場合に項目を作成可能ではあるが，同じ用語であることの同定が専門家でも困難であることも多いため，がん用語とすることは難しい．

表 6 選択基準により Cc から除外した用語例と理由

理由	頻度	用語
2 ホップ	n=221	うがい，いらいら，マーガリン，魚類
固有名詞	n=122	LSG, ASCO, ベンス・ジョーンズ，薬物療法部長，がん治療の三本柱
誤用	n=6	全能照射、抹消滅
文法	n=305	転移を認めない，生理以外，膣がんの病期，同種造血幹細胞移植前，中枢側，寛解導入療法中，髄腔内，鎮痛薬使用
ミス	n=36	PPG), MRI, PET, 瘍崩壊症候群

人名であるベンス・ジョーンズは多発性骨髄腫という血液のがんに特有なベンス・ジョーンズ蛋白という物質を発見したことで有名な医師だが，がん用語としては，物質名である「ベンス・ジョーンズ蛋白」は採用するが，単独の人名では採用しない．

「薬物療法部長」などの病院の役職名も，がんに関する記述の中で比較的良く用いられる呼称である．薬物療法部長が薬物治療を行う疾患はほとんどの場合，がんであり，「薬物療法部長の○○氏と面談」などの文が患者のブログなどでも出現することも予想できる．しかし，医師，看護師，薬剤師，患者などの呼称とは異なり，統一された資格者を示すものではない．また，「がん治療の 3 本柱」は，主に国立がんセンターで患者に対して，がんの治療法である手術・抗がん剤を用いる化学療法・放射線療法の 3 つをまとめて言う場合に用いられる語である．これらも，一般的用語ではないと考えられるため，ここでは「固有名詞」として扱う．

これらの，普遍的な名詞句となっていない固有名詞，単純な人名（手術法等に含まれるものはそのつど採用），呼称の一定しない役職名など 122 語を「固有名詞」として削除するのが妥当と考えられる．

4.2.2 「文法」：文法による削除例

Cc 作成では網羅的に複合語を積極的に収集したが，3.4 の「がん用語選択の基準」の文法上の理由（「〜内」，「〜外」などの連体詞的な用例は採用しない）によって，「転移を認めない」「生理以外」，「放射線治療単独」，「同種造血幹細胞移植前」，「寛解導入療法中」などの複合語を削除語とする．

本研究ではこれらを削除語とするが，これらはがん情報処理に有用な場合もある．たとえば，「生理以外」は，婦人科がんの不正性器出血と呼ばれる症候に関連する語で，「生理以外の出血はありますか？」などの用例がある．この場合，「生理以外」を単独のがんに関係する句として認識することによって，この文が，がんに関するものであることを予測することができる．また，「放射線治療単独」は，化学療法や手術を複合して用いていない「単独」という限定を強調する用例で，「放射線治療だけで治療する」という意味を明示する．この場合も，「放射線治療」と「単独」に分割しては「放射線治療だけで治療する」という意味が弱くなる可能性もある．このように，一つ一つの複合語を用例に従って吟味すると，実際の用例で出現した文脈が，がんの意味を含むことを限定できる可能性もあるため，本研究では一旦削除語とするが，今後の検討を可能とするために，削除語の中でも「文法」という理由で削除したことを明記（305 語）し，本研究で公開するがん用語集合の付録として公開する予定である．

4.2.3 「2 ホップ」：2 ホップ目までのルールでの削除例

2 ホップ目までのルールとは，2 ホップよりも関連性の弱い語（T4）を削除する規則である．たとえば，表 6 の「うがい」は，白血病治療などで感染症の危険が増加することを予防する方

法のひとつである．「白血病治療」を起点として用語の連関を想起する場合，「白血病治療」は「白血球減少」を起こし，「感染予防」や「感染症の危険」を高めるから，予防方法として「うがい」を行う．そのため，「白血病治療」（0 ホップ目）—「白血球減少」（1 ホップ目）—「感染予防」（2 ホップ目）—「うがい」（3 ホップ目）という想起順となる．すなわち，「うがい」は T4 に分類される．

しかし，「白血病治療」—「感染予防」が，「白血球減少」のような，橋渡しを行う用語を介さずに，例えば，「白血病治療は感染症予防を行うことが肝要であり，うがいは最も重要だ.」のような用例を想起するのであれば，「白血病治療」（0 ホップ目）「感染予防」（1 ホップ目）「うがい」（2 ホップ目）という想起順になり，この「うがい」は除外語ではなく採用語 T3 となる．

さらに，「白血病」を起点（0 ホップ目）とした場合は，「白血病」（0 ホップ目）—「化学療法」（1 ホップ目）—「白血球減少」（2 ホップ目）—「易感染性」（3 ホップ目）—「うがい」（4 ホップ目）という想起順が考えられ，この場合の「うがい」は白血病から考えて 4 ホップ目となる．すなわち，T4 である．

このように，作成した選択基準の「2 ホップ目まで」のルールは，0 ホップ目にどのような用語を選択するか，あるいは仲介する用語として何を想起するかによって変化する．このような「2 ホップ」よりも関連性が低い単語については，4.3 の第三者医師による評価の項で述べるが，がんとの関連性の判断が人により異なる．そのため，本研究では，「文法」により削除された語と同様に，がん用語集合の付録として T4 に分類された語も公開する予定である．

4.2.4 削除語に関するまとめ

以上より，Cc から抽出した T1, T2, T3 の 9509 語をがん用語集合 C として採用し，残りの 690 語を削除語とした．これらの語数を表 7 に示す．なお，本研究により作成したがん用語集合を公開するにあたって，我々は，がん用語集合 C に加えて，削除語のなかで「文法」と「2 ホップ」に相当する語については，付録として公開する予定である．その理由は，前述のように，これらの削除語については，本研究においては削除語と判断されたが，場合によっては，がん情報処理に有用な場合があると考えるからである．

表 7 選択基準による Cc（10119 語）の分類結果

分類		頻度	累積頻度	パーセント	累積パーセント
C: がん用語	T1: がんそのものをさし示す語	1637	1637	16.1	16.1
	T2: がんを想起させる語	4167	5804	40.9	56.9
	T3: がんを想起させる語に関連する語	3705	9509	36.3	93.2
Dc: 削除語	（T4: 2 ホップルール）	221	9730	2.2	95.4
	（文法上の理由）	305	10035	3.0	98.4
	（国有名詞，ミス，誤用）	164	10199	1.6	100.0

4.3 複数医師による評価

これまでに作成し分類した用語集合は，専門家が作成したものであるが，その人数が 1 人であるので，必ずしも，他の専門家が同意する用語集合であるとは限らない．そのため，本節では，複数医師により，上述の T1, T2, T3, T4 の分類の妥当性を確認する．

評価用データとして，T1, T2, T3, T4 のそれぞれから無作為に約 50 語（合計 197 語）を選んだ．この評価語データを付録 1 に示す説明文書とともに，臨床経験 15 年以上の医師 4 名（大学院講師以上の腫瘍内科医 2 名，脳神経外科指導医 1 名，循環器内科認定医 1 名：以下，C1–C4 と呼ぶ）に提示し，各人のそれぞれの用語に対する T1 から T4 の評価値を得た．

本研究が付加した T1 から T4 の分類と，各医師による分類の比較結果を表 8 に示す．左のカラムにある 1 から 4 のカラム (Cat) は，本研究で各用語に付与した分類値であり，T1 から T4 の分類を示す．これに対して，それぞれの医師 C1 から C4 の別に，各人による分類を T1 から T4 で示し，それぞれの要素をクロス集計した頻度を示した．また頻度の合計を Total として示した．

これより，対角線に近い部分の頻度が高いことがわかる．たとえば，医師 C1 については，本研究で付与した T1 (Cat1) は，T1 もしくは T2 にほとんどが分類されている．また，C2 については，Cat1 は，46 例中の 38 例が T1 に分類されている．このことより，本研究で付与した分類が，他の医師の判定と一致することが多いことがわかる．

さらに，表 8 における分類の一致の度合いを，Cohen's Kappa(Landis and Koch 1977) を用いて，数値化することにより，本研究による分類と各医師による分類との一致の度合いを調べる．

表 8 医師 4 名（C1 から C4）から得た分類結果（単語数 197）

	C1				
Cat	T1	T2	T3	T4	Total
1	20	24	2	0	46
2	0	32	19	3	54
3	0	10	37	3	50
4	0	7	34	6	47
Total	20	73	92	12	197

	C2				
Cat	T1	T2	T3	T4	Total
1	38	7	0	1	46
2	7	17	21	9	54
3	1	4	25	20	50
4	0	3	13	31	47
Total	46	31	59	61	197

	C3				
Cat	T1	T2	T3	T4	Total
1	30	12	4	0	46
2	1	15	20	18	54
3	0	2	10	38	50
4	1	0	6	40	47
Total	32	29	40	96	197

	C4				
Cat	T1	T2	T3	T4	Total
1	22	21	3	0	46
2	3	23	27	1	54
3	0	2	46	2	50
4	0	3	37	7	47
Total	25	49	113	10	197

そのために，表 8 から複数のクロス集計を得て，それらにおける Kappa を調べる．複数のクロス集計を得るときには，T1 と T2 など，隣接するカテゴリを一つのカテゴリとすることを，次に示す分割例 1 から 7 の全ての場合について試した．たとえば，被験者 C1 の分割例 2 と分割例 6 について，本研究の想定に対する Kappa の算出対象とするクロス表について図 9 に示す．

分割例 1：1, 2, 3, 4　　…　表 8 と同分類
分割例 2：1, (2, 3, 4)　　…　1 と，(2, 3, 4) 2 つに分割
分割例 3：1,(2, 3), 4　　…　1, (2, 3), 4 の 3 つに分割
分割例 4：(1, 2), 3, 4　　…　(1, 2), 3, 4 の 3 つに分割
分割例 5：1, 2, (3, 4)　　…　1, 2, (3, 4) の 3 つに分割
分割例 6：(1, 2), (3, 4)　…　(1, 2) と (3, 4) の 2 つに分割
分割例 7：(1, 2, 3), 4　　…　(1, 2, 3) と 4 の 2 つに分割

Cohen's Kappa は，0.4〜0.6 が中等度，0.6 以上で生起反応において強い連関を示すと言われている (青木 2002)．それぞれの分割例別の本値を検討することによって，どの分割が実際の医師の持つ語感に合致するかを調べることができる．結果を表 9 に示す．左カラムにそれぞれの分割例を示し，この分割例別に C1, C2 など各人と，本研究で行った分類をそれぞれの分割例に割り付けなおし，各人の反応との間の Kappa を求め，最右カラムに平均値を示した．これより，

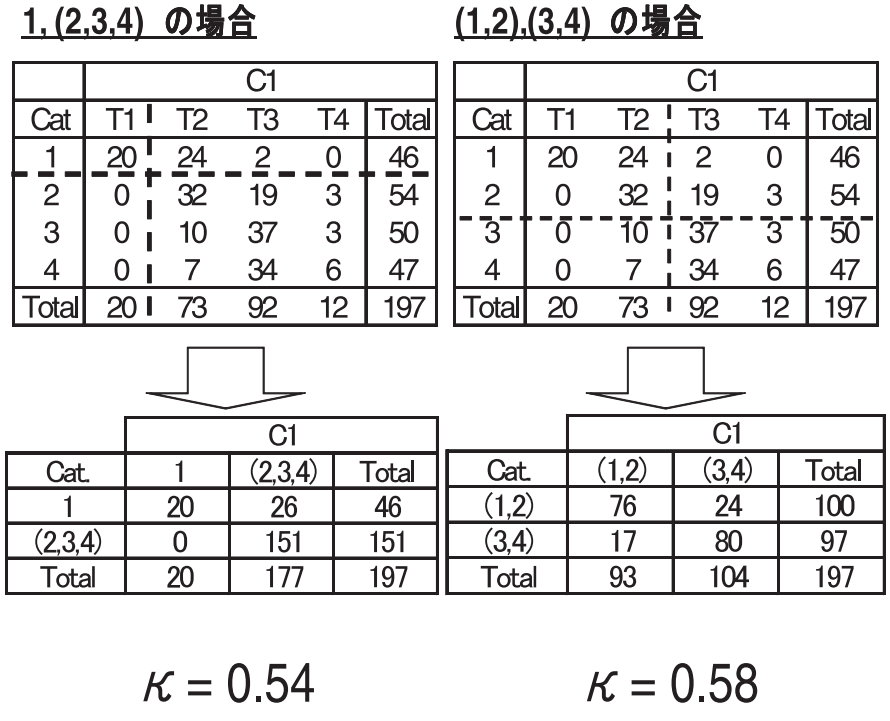

図 9　被験者医師 C1 での分割例 2 と分割例 6 における κ 値の算出対象

表 9 各分割例別の Cohen's Kappa 値

分割法	分割数	Cohen's Kappa				
		C1	C2	C3	C4	平均値
分割例 1　1, 2, 3, 4	4	0.30	0.42	0.31	0.32	0.34
分割例 2　1, (2, 3, 4)	2	0.54	0.77	0.71	0.54	0.64
分割例 3　1, (2, 3), 4	3	0.29	0.51	0.39	0.32	0.38
分割例 4　(1, 2) 3, 4	3	0.36	0.43	0.33	0.41	0.38
分割例 5　1, 2, (3, 4)	3	0.46	0.55	0.49	0.49	0.50
分割例 6　(1, 2), (3, 4)	2	0.58	0.61	0.55	0.64	0.59
分割例 7　(1, 2, 3,) 4	2	0.12	0.42	0.35	0.18	0.27

分割例 1, 3, 4, 7 では，Kappa 値がいずれも 0.4 以下であり，有意な一致とはみなせないが，分割例 2（T1 とそれ以外の 2 分割）と分割例 6（T1, 2 とそれ以外の 2 分割）では 0.6 前後の高い一致であった．また，分割例 5（T1, T2 とそれ以外の 3 分割）でも 0.5 の比較的高値であった．

これらのことから，本研究で想定した，T1（0 ホップ目）：がんそのものを示す語，T2（1 ホップ目）：がんを想起させる語までは，実際の被験者医師の語感に近いことが示された．これにより，本研究で行った一人の医師による用語の切り出しとがんとの関連性を想起した分類であっても，概念の中核となる「がんそのものをさし示す語」から，「がんを想起させる語」として感じるような距離感は，第三者医師にとっても共通する語感であることが示された．

これは，国家資格を持つ専門職である医師は，一旦国家試験の段階で用語の統一が行われていること，臨床の現場では患者の診断や治療などの相談を頻繁に書面でやりとりする機会が多いこと，ほとんどの医師ががん患者を診断治療した経験を持つことなどが主な理由と思われる．これに対し，本研究で T3（2 ホップ目）「がんを想起させる語の関連語」より関連性が低いと想定した語（T4 も含む）に関する分類が医師によって異なるのは，4.2.3 で述べたように T3 や T4 の語は T2 や T1 に対して何らかの関連語を連想できるかどうかによって，距離感に差が生じやすいことなどが理由と思われる．T3, T4 の分類について，複数医師間で一致の高いような基準を研究するのは，今後の検討課題である．

これらの結果は，本研究が 3.4 で規定した，「Cc（10199 語）の用語をがんとの関連性によって T1, T2, T3, T4 と分類し，T1 から T3 までをがん用語 C とする」という選択基準に対して，T1 から T2 までは複数医師間で一致がとれていることを示している．そのため，T1 と T2 については，中心的で妥当ながん用語といえると考える．一方，T3 と T4 については，かならずしも複数医師間での一致はないが，これらの用語候補は，まず，(1) 国立がんセンターのテキストに含まれているということ，(2) 専門家が吟味した語であることから，がん用語集合 C（T1 から T3）およびその付録としての削除語のリスト（T4）として公開する価値があると考えた．

4.4 がん用語の選択基準の必要性に関する検討

4.3 節の実験により，3.4 節のがん用語の選択基準に従って行った T1, T2, T3, T4 への分類は，第 3 者医師が行った分類と有意に相関することが示された．このことから，本研究がこれまでに行った，網羅的収集による Cc の作成（3.2 節），収集された用語の分類と用語選択基準の設定（3.4 節）と削除語の選別（4.2 節）に関しては，妥当性が示されたと考える．しかし，これだけではがん用語の選択基準として 3.4 節で提案した 2 ホップルールの必要性は示していない可能性がある．すなわち，従来の用語選択基準である，単に，「がんに関係する用語か，そうでない用語か」（以下，「がん用語か否か」と表記する．）による選択でも十分である可能性がある．

この問題について検討するため，用語候補集合 Cc から無作為に 100 語を選択し用語候補集合を作成した．そして，4.3 節の医師 4 名とは別の医師 6 名（D1 から D6：国立がんセンター研究者 2 名，大学医学部教授 2 名，ならびに内科医師 2 名．順不同.）を被験者として，付録 2 に示した依頼文と共に示した用語 100 語を「がん用語か否か」に分類を依頼した．なお，100 語の内訳は，T1 が 9 語，T2 が 24 語，T3 が 39 語，T4 が 28 語である．また，各医師毎に，がん用語として選択した語数は，D1 が 61 語，D2 が 18 語，D3 が 53 語，D4 が 48 語，D5 が 33 語，D6 が 6 語である．なお，被験者 D6 は他の医師 5 名に比べ，がん用語とした語数が顕著に少ないが，除外せずに評価結果とすることとした．

本節での目的は，従来の「がん用語か否か」という選択基準と，提案手法である「図 6 のがん用語選択基準」とを比較することであるので，まず，「がん用語か否か」という選択基準によりがん用語を選択した場合における，6 名の医師（D1–D6）間での κ 値を表 10 に示す．

表 10 より，D1 と D4, D5, D2 と D5, D3 と D4, D5, D4 と D5 はそれぞれ中等度以上の一致を示しており，医師間では相関する場合もあるが，D6 のように，他の医師と有意な相関を示さない例もある．また，D1 から D6 全体としての κ 値の平均値は 0.32 であり，がん用語を極端に少なく選択した D6 を除き D1 から D5 までとした場合の κ 値の平均値は 0.39 であった．

これに対して，提案する選択基準による医師間の一致の度合いをみるために，表 11 に，4.3 節の医師 4 名（C1〜C4）による 197 語の分類結果に対して，T1 から T4 の 4 つを仮に 2 つに分類すると仮定し，P1（T1 と T2, T3, T4），P2（T1, T2 と T3, T4），P3（T1, T2, T3 と T4）

表 10 従来法（がん用語か否か）による用語選択の被験者間の κ 値

	D1	D2	D3	D4	D5
D2	0.04				
D3	0.05	0.28			
D4	0.62	0.38	0.58		
D5	0.44	0.46	0.49	0.57	
D6	0.08	0.26	0.11	0.13	0.23

の各分割例について，4 名の医師間の κ 値を算出した結果を示す．表 10 に比べ，分割例 P1, P2 において各医師間で高い一致を示し，P1, P2, P3 におけるこれら κ 値の平均値はそれぞれ 0.67, 0.69, 0.19 であった．

以上のことから，提示した用語集合を表 10 のように「がん用語か否か」で分類する場合よりも，表 11 に示した本研究の提案する「2 ホップルール」により分類する場合のほうが，医師間の用語選択の一致性が高く，得られた用語集合のコンセンサスを得やすいことが示された．

4.5 専門用語抽出アルゴリズムでの抽出語例とがん用語集合 C の比較

4.2 節で得られたがん用語集合 C は，信頼できるコーパスから専門家が抽出し，その妥当性が複数の医師により確認されたものである．そのため，このがん用語集合 C を用いて，従来開発されてきた専門用語抽出アルゴリズムの性能を評価することが可能である．つまり，用語集合 C は，専門用語抽出アルゴリズムの正解データとして有用であると考える．

そこで，専門用語抽出アルゴリズムの評価の一例として，中川らによって実装されている「言選 Web」(http://gensen.dl.itc.u-tokyo.ac.jp/gensenweb.html) を用いて得られた用語とがん用語集合 C とを比較することにより，用語集合 C の正解データとしての有用性を検討する．

比較の方法としては，3 節と同一のコーパスから，表 12 に示す各疾患を対象として，言選 Web を利用して用語を抽出した．これを HN とする．つぎに，同じコーパスについて，用語集合 C に含まれる用語を抽出し，これを用語集合 Cd とした．なお，このとき，形態素解析器 Mecab を利用し，Mecab の辞書に用語集合 C を加えることにより，C 中の用語が自動的に同定できるようにした．（Mecab の辞書に用語を加えるときには，その品詞とコストを試行錯誤により決定し，C 中の用語がテキストにあるときには，それが解析結果に優先的に出力されるようにした．）

これを元コーパスである国立がんセンターの Web データの中で，肺がん，胃がん，食道がん，

表 11 提案法（2 ホップルール）による用語選択の各被験者間の κ 値

		C1	C2	C3
P1	C2	0.51		
	C3	0.74	0.68	
	C4	0.72	0.61	0.73
P2	C2	0.69		
	C3	0.61	0.71	
	C4	0.66	0.71	0.76
P3	C2	0.11		
	C3	0.13	0.04	
	C4	0.52	0.12	0.22

P1: 1 and (2,3,4), P2:(1,2) and (3,4), P3:(1,2,3) and 4

大腸がん，および乳がんに適用し，HN と Cd に関する諸量を表 12 に示した．表 12 では，それぞれの疾患のコーパスにおける，①|Cd|：C によって得られた語数，②|HN|：HN によって検出された数，③|Cd ∩ HN|：Cd ∩ HN の語数，④|Cd − HN|：Cd と HN に含まれた語を比較して Cd にのみ含まれた語数，⑤|HN − Cd|：HN と Cd を比較して HN にのみ含まれた語数，⑥HN − Cd の再採用語：表 13 に詳細を示すが⑤の語で，Cd に含まれなかったが用語とすべきと思われた語数，⑦|Cd ∩ HN|/|HN|：HN に対する HN ∩ C の語数の比，（③/②, C を正答とした場合の HN の精度），⑧|Cd ∩ HN|/|Cd|：C を正答とした場合の再現率（③/①）を示す．

これより，これら疾患での HN の再現率（⑧）は平均値 0.86 であり，HN の C に対する網羅性は高いと思われる．しかし，精度（⑦）は平均値で 0.52 であり，HN で得られた用語の約半数は人手で選択しなおす必要があることがわかる．また，HN で検出でき，C で検出できなかった語数⑥は数語であり，少数であることも分かる．すなわち，C の網羅性は高いといえる．以上により，HN は用語切り出しに関しては本研究で専門家が行った用語抽出を高い再現率で実現する可能性を示すが，約半数の削除語とすべき語を含んでいることから，用語選択を追加して行う必要のあることを示している．

さらに，HN で得られた語と C の比較を行い，HN で検出されたが C に含まれなかった語を「除外語」として，4.2 節表 6 と同様に分類した結果を表 13 に示す．

表 13 より，HN では切り出しミス，文法（複合語など）は少数であり，大多数は，本研究で規定した 2 ホップルールによる除外理由である．2 ホップルールによる除外例は，肺がんのコーパスでは，扁平，率，利用，要素，要因，葉，用量，有無，ハイリスク，つけ根，タイプなどで

表 12　従来法 (HN) とがん用語辞書で得られた用語 (Cd) の比較

疾患名	① |Cd|	② |HN|	③ |Cd ∩ HN|	④ |Cd − HN|	⑤ |HN − Cd|	⑥{HN − Cd} の再採用語	⑦ |Cd ∩ HN|/|HN|	⑧ |Cd ∩ HN|/|Cd|
肺がん (LC)	440	677	381	68	296	4	0.56	0.87
胃がん (GC)	319	637	275	48	366	8	0.43	0.86
食道がん (EC)	393	669	337	56	329	5	0.50	0.86
大腸がん (CC)	435	689	378	57	311	3	0.55	0.87
乳がん (BC)	318	474	270	48	204	3	0.57	0.85
平均値	381	629.2	328.2	55.4	301.2	4.6	0.52	0.86

表 13　従来法 (HN) によって得られた語の除外理由

理由	肺がん	胃がん	食道がん	大腸がん	乳がん
2 ホップ	218	279	249	227	162
国有名詞	4	6	16	19	4
採用	4	8	5	3	3
切り出しミス	18	20	20	21	19
文法	52	53	39	41	16
合計	296	366	329	311	204

あった．また，胃がんのコーパスでは，老化度，輪切り，量，両方，流れ，率，利用，陽性，有無，役割，門，目的，膜，麻酔，本国在住者，方法，変更，変化，壁，分泌，物質，部分，負担，不十分，病院，評価，表面，標準，範囲，髪の毛，発生，白人，年齢別，年単位などであった．ここで，2 ホップルールによる除外というのは，専門家により，意味的に判断した結果としての除外である．すなわち，HN は，意味的な理由により除外された語を用語候補として抽出したといえるため，この点において，HN には改善の余地があるといえる．これより，今後 HN 法などの自動抽出アルゴリズムの教師データとして，今回作成した C を教師データとすることが有用であると考えられた．

5 考察

本研究で行った，がん用語集合の作成方法についてまとめる．まず，(1) がん用語集合は，がんに関連する用語をできるだけ網羅することが望ましい．ただし，あまりに関連性の小さい用語も含めると，それらが，がん情報処理に悪影響を与えることが考えられるので，好ましくない．(2) そのため，がんとの関連性が一定以上の強さの用語のみを，がん用語集合に含めるべきである．(3) そこで，本研究では，関連性の強さの指標としてホップ数を導入し，「がん用語の選択基準」としての「2 ホップルール」に基づき，がん用語集合を選定した．

次に，本研究と従来の用語集の作成法を比較する．まず，医学領域での多くの用語集の妥当性は，2.2 節で示したように，長い時間と多数の用語選定委員によって担保されている．ただし，これらの用語の選択基準は，それぞれの分野において「使用される用語およびこれに関連の深い用語」というものであり，ある用語について，それがその分野で使用されるかどうかや関連が深いかどうかの判断の基準については，2.2 節で示した用語集においては明示されていない．

一方，本研究では，まず，国立がんセンターの Web テキストに出現した用語候補を Cc 抽出の対象とすることにより，抽出された用語が，医学的内容の信頼性および網羅性を持つことを仮定した．次に，専門家が，このテキストから「がんに関係する」と判断した用語候補を網羅的に抽出することにより，Cc を作成した．この段階までにおける本研究と従来の用語集との違いは，従来の用語集の用語の採取元は明示されていないが，Cc に採用された用語の採取元は明らかである点である．また，本研究においても，従来の用語集と同様「がんに関係する」という主観により，用語候補を選定している．ただし，従来の用語集においては，多くの選定委員が用語選択に関与することにより，用語の妥当性を保証しているのであるが，本研究においては，一人の専門家の語感のみによるものであるため，用語選択の妥当性は，それほど保証されていないと考えられる．(ただし，3.2 節で述べたように，このような語感は，国家試験などにより基本的知識を共有する専門家間では共有されていると考えた．)

そこで，本研究では，がん用語候補集合 Cc から，妥当ながん用語を選定するために，3.4 節で

「がん用語の選択基準」を設定し，それに基づいて，がん用語を選定した．これは，「がんに関係する」という曖昧な判断基準から，Cc の整理を通して，相対的に明確な判断基準である「がん用語の選択基準」としての 2 ホップルールを構築し，それに基づいて，がん用語を選定したといえる．

この 2 ホップルールの概略は以下のものである．まず，がん用語候補を T1（0 ホップ目）：がんそのものをさし示す語，T2（1 ホップ目）：がんを想起させる語，T3（2 ホップ目）：がんを想起させる語に関連する語，T4（3 ホップ目以上）：がんとの関連を想起しにくい語，の 4 つに分類した．そして，2 ホップ目までである，T1, T2, T3 をがん用語として選定するというものである．

この 2 ホップルールに基づいて一人の専門家により選定されたがん用語集合の妥当性を検討するために，4.3 節では，本研究で行った T1 から T4 までの分類と，複数医師の行った分類の一致性を評価し，T1 から T4 の間で T2 までの用語選択の一致性（κ 値 0.5 から 0.6）が示された．これにより，がん用語選択基準である 2 ホップルールの妥当性を示すことができたと考える．

さらに，4.4 節で検討したように，専門知識を有する医師であっても，ある用語が「がん用語か否か」で分類する場合の各人の一致率は，2 ホップルールを明示して段階的に用語分類を行った場合に比べ低値（κ 値 0.3〜0.4）であった．これより，本研究の提案する用語分類法である「2 ホップルール」を与えたほうが，従来の「がん用語か否か」という分類を行う場合よりも，適切に専門用語を選択することが可能であることが示された．

また，4.5 節で示したように，本研究で作成した用語集合（C）を正解データとした場合，自動抽出アルゴリズムによって得られた用語集合（HN）の再現率は約 0.86 を示したが，精度は 0.52 であった．これは，HN のアルゴリズムに改善の余地があることを示すと考える．このことから，本研究で作成した用語集合は，自動抽出アルゴリズムの評価に有用であると考える．

次に，1 節で目標としたように，本研究におけるがん用語の作成方法が，他の分野の用語集合の作成にも適用可能かについて考察する．まず，本研究におけるがん用語の選定方法を一般化すると次のようになる．(1) まず，医学的内容の信頼性と網羅性が高いコーパスを選定する．(2) 次に，そのコーパスから，専門家が，対象の病気に関連すると考える用語候補を網羅的に収集する．(3) 最後に，2 ホップルールに基づいて，対象の病気に関係する用語を選定する．なお，2 ホップルールは，より一般的には，中心的な用語から関連語までについて，T1, T2, T3, T4,… のような関連度の段階を定め，そのある段階までを，用語として認定するというものである．

このような用語集合の作成法の一般化が，他の病気に対しても適用可能かを実証することは今後の課題であるが，がんという，多数の疾患からなり肺がんや胃がんのような固型がんから，白血病のような肉腫と呼ばれる疾患群までを総称する複雑な概念からなる用語集合の作成が可能であったことから，少なくとも「パーキンソン病」や「アルツハイマー病」などの難病，「糖尿病」や「高血圧」のような生活習慣病など，ほぼ医学の全分野に応用可能と思われる．また，

本作成法の医学以外の分野での応用可能性（例えば，「不安」や「抑圧」などの心理学分野の症候を示す語群）に関しても，今後検討する予定である．

以上，本研究における用語集作成法と従来の用語集の作成法とを比べて，本研究における新規な点は，用語選択の基準について，「がんに関係する」という曖昧な判断基準から，2 ホップルールという相対的に明確な判断基準を構築し，それに基づいて，がん用語を選定したことであると述べた．

次に，本研究で作成したがん用語集を実際のがん情報処理に適用するために必要と考えられる 2 つの拡張について述べる．これらは今後の課題である．

まず，本研究においては，国立がんセンターの Web テキストから抽出した用語のみをがん用語集に採用しているが，その他のテキストから抽出した用語もがん情報処理に有効な場合が考えられる．たとえば，「がんに効果がある」と宣伝している悪質な商品誘導へのページをフィルタリングして，良質ながん情報のページを推薦するためには，本研究で作成した用語集に加えて，悪質なページに特徴的な単語を利用すると効果的と考えられる．また，ブログの検索などの応用においては，検索ユーザが頻繁に入力する単語も追加すると効果的と考えられる．また，4.2.2 節で検討した，「生理以外」等の文法的理由による削除語についても，検索等には有用であると考えられるので，本研究で提案したがん用語集合に加えて使用すると有効であると考える．

次に，本研究においては，がん用語を T1, T2, T3 に分類した．これは，がんとの関連性により，用語をランキングしたと考えることもできる．このランキングは，がん用語集合を選定するためには有効であるが，その他の目的に対して最適とは限らない．たとえば，がんに関係するページを Web 検索するときに，「レントゲン写真」と「急性転化」とは，どちらも T2 であるが，前者ががん以外のページも上位の検索結果に含むのに対して，後者はがん（白血病）のページがほとんどである．つまり，検索結果の適合率という観点からは，「急性転化」の方が良い用語である．

このように，T1, T2, T3 は関連性をベースにした用語の分類であるが，目的が明確な場合には，この分類は最適とはいえない．しかし，本研究により作成したがん用語集合があれば，それを検索のためにランキングする等，目的に応じてがん用語をランキングすることも考えられる．これは，全語彙にランキングを行うよりはるかに容易である．すなわち，本研究で作成した用語集合のランキングを目的別に整備することにより，目的対応の用語集合ができる．したがって，本研究により作成したがん用語集合は有用であると考える．

6 まとめと今後の方針

がん情報処理を補助することを目的として，その言語基盤であるがん用語辞書を，医師免許を持つ専門家が人手で作成した．わが国で生じる可能性のあるほぼ全てである 59 個のがんの説

明用コンテンツを含む国立がんセンターの Web 文書全体（テキストファイルとして約 15 メガバイト）を，がんの情報に関して十分な網羅性を持つ権威あるコーパスとして選択した．これから直接人手で，がん用語として理解可能な用語を網羅的に収集し，10199 語の用語候補集合 Cc (Cancer term Candidates) を得た．

得られた Cc の各語を理解しやすくすることや用語間の整合性を調整するために，各語の用例から「病名」「解剖学用語」「臨床用語」「一般語」「検査用語」などの種別を作成した．また，各語をがんとの関連性を想起しつつ分類した結果，T1（がんそのものをさし示す語），T2（がんを想起させる語），T3（がんを想起させる語に関連する語），T4（がんとの関連性を想起しにくい語）の 4 つに分類できた．これら「種別」とがんとの関連性を参考に，がん用語として，T1 から T3 までをがん用語 C として選択するという選択基準を規定した．

元コーパスに対する Cc の用語候補抽出の一貫性を調べるために，コーパスから 10 個の疾患の説明用ページのテキストファイルを対象として，再度網羅的な用語収集を行い，得られた用語のうち Cc に含まれた語の比率を調べた．その結果，これら 10 個の対象における Cc の再現率は 94%から 99.5%であり，元コーパスに対する Cc の再現性は十分であることが示された．

さらに，選択基準をもとに，T1, T2, T3, T4 の分類を，Cc（10199 語）の全用語に対して人手で行い，T1, T2, T3 に分類される用語を，がん用語集合 C とした．

選択基準の妥当性を検討するために，T1（1637 語），T2（4167 語），T3（3705 語），T4（221 語）の中から約 50 語ずつを無作為に選び，評価用ワードセット（用語数 197 語）を作成し，これを選択基準の説明文とともに医師 4 名に示し，本研究で想定した T1 から T4 の分類と各医師の分類の比較を行った．その結果，T1, T2 までの分割に対する Cohen's Kappa 値は約 0.6 であり，さらに T1, T2 とそれ以外の 3 分割の場合でも 0.5 を示したことから，T1 と T2 までの語彙選択の妥当性が示された．

以上より，本研究で行ったコーパスからの網羅的用語収集と用語選択基準の組み合わせによって，少人数で妥当性のあるがん用語集合を作成することができた．

本研究のような用語集合の作成は，目的とする分野での質の高いコーパスが存在することが重要だが，今後他の医学分野においても同様の手法で，妥当性のある用語集合を作成していくことが可能と思われる．

また，2.1 節であげた National Cancer Institute の Dictionary of Cancer Terms の語彙数 5236 語と，Cc の T1 と T2 の合計語彙数 5804 語が，対象とするがんの種類や社会制度などが異なる 2 つの地域で同規模であることも興味深い．これら語彙の相互比較も今後の検討課題である．

今後，このがん用語集合を用いたがん情報処理の実現にむけて研究を行うことが課題である．なお，本研究で収集したがん情報コーパスならびに，分類タグつきのがん用語集合は，国立がんセンターとの協議の上で公開する予定である．

謝 辞

本研究を行うに当たり用語辞書作成に御協力いただいた，元北陸先端科学技術大学院大学学生 木村俊也氏，国立がんセンター中央病院 若尾文彦医長，同がん対策情報センター 石川ベンジャミン光一室長，滋賀医科大学 藤山佳秀教授，程原佳子教授，八尾武憲博士，近畿大学医学部 西尾和人教授，鹿児島大学大学院医歯学総合研究科 秋葉澄伯教授，高岡医院 高岡篤博士，野洲病院 木築裕彦医師に謝意を表する．論文作成に御協力いただいた，情報通信研究機構主任研究員 門林理恵子博士ならびに村松亜左子氏に謝意を表する．なお，本研究は NICT 運営費交付金（新世代ネットワーク研究センター），平成 19 年度，20 年度厚生労働省がん研究助成金研究総合研究「がん情報ネットワークを利用した総合的がん対策支援の具体的方法に関する研究 」若尾班等の支援を得て行った．関係各位に深謝する．

注

1) 日本内科学会編，内科学用語集（第 5 版）(1998)．医学書院．ISBN: 978-4-2601-3641-9 (4-2601-3641-0).
2) 循環器学用語合同委員会，循環器学用語集（第 3 版）(2008)．丹水社．ISBN: 978-4-9313-4722-9 (4-9313-4722-3).
3) National Cancer Institute (NCI). http://www.cancer.gov/dictionary/
4) 日本糖尿病学会（編集）(2005)．糖尿病学用語集．文光堂．ISBN: 978-4-8306-1363-0 (4-8306-1363-7).

参考文献

青木繁伸 (2002)．http://aoki2.si.gunma-u.ac.jp/lecture/Kappa/kappa.html

Humphrey, S. M. and Miller, N. E. (1987). “Knowledge-based indexing of the medical literature: the Indexing Aid Project.” *J Am Soc Inf Sci*, **38**(3), pp. 184–196

国立がんセンター．http://www.ncc.go.jp/index.html

Landis, J. R. and Koch, G. G. (1977). “The measurement of observer agreement for categorical data” in *Biometrics*. **33**, pp. 159–174.

Nakagawa, H. (2000). “Automatic Term Recognition based on Statistics of Compound Nouns.” *Terminology*, **6**(2), pp. 195–210.

Nakagawa, S., Kimura, M., Itokawa, Y., Kasahara, Y., Sato, T., and Kimura, I. (1995). “Development of Internet-Based Total Health Care Management System with Electronic Mail.”

Journal of Epidemiology, **5**(3), pp. 131–140.

National Cancer Institute (NCI) (2007). PDQ (Cancer Information Physician Data Query from National Cancer Institute). http://www.cancer.gov/cancerinfo/pdq/

National Library of Medicine (2006). *Medical Subject Headings (MeSH) fact sheet*.

野口迪子 (2000)．医学書を探す：基本図書を主として．情報の科学と技術, **50**(11), pp．542–552.

佐藤理史，佐々木靖広 (2003)．ウェブを利用した関連用語の自動収集．情報処理学会研究報告, 2003-NL-153, pp. 57–64.

（財）先端医療振興財団 (2009), PDQ 日本語版, http://cancerinfo.tri-kobe.org/

Wendy A. Weiger（原著），坪野吉孝（翻訳）(2004)．がんの代替療法―有効性と安全性がわかる本．法研.

山口徹，北原光夫（編集）(2004)．今日の治療指針．医学書院.

山室真知子 (2000)．医学情報の患者へのバリアフリー．情報の科学と技術, **50**(3), pp．138–142.

略歴

中川　晋一（正会員）：1988 年滋賀医科大学卒，医師．1996 年京大院（医）終了，博士（医学）．同年国立がんセンター研究所，1998 年郵政省通信総合研究所（現情報通信研究機構），現在，同主任研究員，次世代インターネット技術開発に従事，IT 技術の実社会への応用（情報通信医学）に興味がある．言語処理学会，情報処理学会，日本内科学会等会員

内山　将夫（正会員）：1992 年筑波大学卒業．1997 年同大学院工学研究科修了．博士（工学）．現在，情報通信研究機構主任研究員．言語処理の実際的で学際的な応用に興味がある．言語処理学会，情報処理学会，ACL 等会員.

三角　　真（非会員）：2004 年北陸先端科学技術大学院大学博士前期課程修了．修士（情報科学）．同年，JST 重点支援協力員に採用．2006 年 NICT 技術員に採用．2008 年から東京工業大学博士後期課程在学．情報通信の研究に従事.

島津　　明（正会員）：1973 年九州大学大学院理学研究科修士課程修了．同年，日本電信電話公社武蔵野電気通信研究所入所．1985 年日本電信電話株式会社基礎研究所．1997 年北陸先端科学技術大学院大学情報科学研究科教授．工学博士.

酒井　善則（非会員）：1969 年東京大学工学部電気工学科卒業 1974 年同大学院博士課程修了．工学博士．同年電電公社電気通信研究所入社 1987 年東京工業大学助教授 1990 年同教授，画像情報処理，情報ネットワークの研究に従事 1994 年テレビジョン学会著述賞，1998 年画像電子学会論文賞，2001 電子情報通信学会業績賞

（2008 年 7 月 11 日 受付）
（2008 年 8 月 26 日 再受付）
（2009 年 1 月 15 日 再々受付）
（2009 年 1 月 18 日 採録）

付録 1

以下に示す「各評価担当医師への依頼文」は，4.3 節で述べた評価語データについて，各医師に各語を 4 分類してもらうための説明文である．ここで，本文中で説明した T1, T2, T3, T4 の各分類は，依頼文においては，それぞれ，1, 2, 3, 4 に対応する．なお，本文中における T1, T2, T3, T4 の説明と依頼文中における 1, 2, 3, 4 の説明とは，言葉遣いや用例等が若干異なるが，同一の分類内容を説明している．言葉遣いや用例等が若干異なる理由は，論文執筆時に，下記依頼文の分類内容が，より明確に伝わるように努めたためである．

各評価担当医師への依頼文

用語集合分類のお願い

現在，がんの用語集を作成中ですが，評価を必要としています．全部で約 1 万語あります．用語は全て国立がんセンターの Web から手で抽出したものですが，これを，がんを直接さす用語から関連用語までに分類することを考えています．この用語集合が完成すれば Web から直接がんの概念を含むページを選択することなどが可能になると思います．

次の 4 つに分類することを考えています．

1. がんそのものをさす語（用語の中にがんの概念を含む語）
2. がんを想起させる用語（その用語ががんを想起させる語）
3. 関連用語（がんの関連用語と思われる語）
4. 除外すべき語（がんとは関係がないと思われる語）

つきましては，添付のエクセルのファイルに示しました約 200 語それぞれについて，上の 4 つの番号を振っていただいて返信していただけると大変助かります．分類は，厳密に行っていただくのではなく，このメイルをお読みいただき，用語をごらんになり，第一印象で分類して下さい．

用語分類の説明

1．がんそのものを指す用語

例えば，胃がん，肺がん，乳がん手術のように，がんという用語そのものを含む用語をはじめとして，進行期中悪性度，ATLL 細胞，骨髄腫細胞のようながん自体の病名や病態，あるいは経尿道的膀胱腫瘍切除術のように，がんを用語の中に含んでいるもの．脳腫瘍のような総称や，髄膜腫のような鑑別を要するものも含めてください．

2．がんを想起させる用語

1．に比べて用語の中にがんの概念を含んでいるものではないけれども，文中にその用語が出現することによって内容ががんの意味を表していると思われるような用語です．

例えば，化学療法という用語の場合は，感染症に対する抗生物質を用いると言う意味もありますが，「患者に対して化学療法を行った．」という用例の場合，がんに対する抗がん剤を用いた化学療法という意味に用いられます．放射線療法の場合は殆ど全ての場合，がんに対する治療をさします．腔内照射装置，内視鏡的逆行性胆管造影，乳腺 X 線検査，病理生検組織などの，その用語から，がんを想起させることのできる用語．また，「住民検診」のように，がんの検出を目的としている用語も含めてください．

3．がん用語の関連用語

上の 1，2 ほど，がんの概念に近くはないが関連していると思われる用語です．

例えば，眼底検査という用語の場合，さまざまな腫瘍で脳圧亢進の診断などで用いられますが，次のように考えます．
・脳腫瘍―脳圧亢進―眼底検査
このように考えて，関連する語と考えられる用語を含めてください．
また，「石綿金網」は胸膜中皮腫の原因物質であるアスベストを含有しています．これは，想起順から連想すると，次のように考えられます．

・胸膜中皮腫―アスベスト―石綿金網―金網

このように直接の原因物質に比べ，関係が少し遠いと思われる関連語をこの分類とし，金網は関連語には入れないが，石綿金網までは胸膜中皮腫の関連語として分類します．それに比べて，金網は原因物質を含まないので，関連語ではないと判断します．

4．関係ない用語

上のどの範疇にも入らない，がんとは関係ない用語と思われる用語．また，意味不明と思われる用語もこの範疇に入れていただいて結構です．

付録 2

各評価担当医師への依頼文

用語集合分類のお願い

2 ページ目から 4 ページ目までの表に，単語が合計 100 個書いてあります．これを，「がんに関係する用語か，そうでない用語か」の 2 つに分けてください．
単語を見ていただいて，
がんに関係する語なら○，がんに関係ない語なら×

を右側のカラムに書き入れてください．