ユニバーサル・レキシコンによる名詞複合語生成

野村直之
nomura@hum.cl.nec.co.jp
NEC情報メディア研究所

1 はじめに

複数の概念記号を日本語の名詞句として生成する際に名詞連続で生成するための条件は従来あまり検討されていないようである。デフォルトの生成戦略により名詞と名詞の間に「の」を挟むか、不自然な訳出に気づいた時点で複合名詞句全体を辞書登録するかの対処がなされていると考えられる。本稿では、動詞から派生した名詞のカテゴリーとして、他の普通名詞と名詞連続で生成可能なものを切り出し、複合語全体の辞書登録無しでも名詞連続を生成可能にする切り分け条件を提案する。この結果、従来よりも自然で読みやすい生成文を得られるようになることが期待できる。辞書への採録可否の検討にあたっては、語彙項目が記憶されていることと、文中で語としての資格をもち語として振る舞っていること、との違いに着目する[注1]。

一方、音声合成の分野では、名詞連続句のイントネーションの生成に際して、語彙の組合せに起因する誤りが以前から報告されており（匂坂85, 佐藤88）、言語理論の助けによって1つ1つの語彙ごとに登録するだけにとどまらない系統的な解決法が求められている。たとえば、「輸入業者」では、第2の音節でアクセントが上行し末尾の音節で下降するという、通常の1語の名詞のイントネーションとなるが、「輸入再開」では、構成語の境界部における短いポーズが入り、且つ、境界位置から再び上降アクセントが入るという事実が知られている。本稿では、2つの名詞が連続したときの、これら2種に大別されるイントネーションの型が、語彙的な複合語と文法的な複合語という2種類の複合語の違い (Kageyama93, Nomura94) に各々対応するという分析を行うことによって、再現性の高い辞書コーディングとその利用法を提案する。

2 複合語の韻律と統語現象、意味現象

佐藤88によれば、自立語どうしが2つ結合する場合のアクセント結合の基本規則は、次の2つである。（以下では上付き文字を高アクセント部分、他を低アクセント部分とする）

1) 後続語が平板型か尾高型の場合、後続語の第1モーラにアクセント核が来る；

e.g. おんせい（音声）＋ごうせい（合成）→おんせいごうせい

2) 後続語が中高型か頭高型の場合、後続語のアクセントが活かされる；

e.g. スーパー＋コンピュータ →スーパーコンピュータ or スーパーコンピュータ

実際には、2)の場合、上例のように後続語が中高型だった場合、先頭文字「コ」の音高が若干下がって元のアクセントが活かされることもあるが、それが活かされずに複合語の中間部分の音節が全て平板で高アクセントのままとなることのほうが多いようである。仮に後者を事実とし、高低2値のアクセントのみを弁別することにすると、1), 2)の区別は不要となり、自立語どうしのアクセント結合の基本規則としては、「先頭の音節を低く、中間は高く平板となり、最後の音節で低く下降。」の1つだけで近似できる[注2]。これは、3つ以上の語を結合した場合も同様である。

e.g. ようげんはせいめいし（用言派生名詞）

このアクセント結合規則には相当数の例外がある。以下の用言派生名詞では、角括弧内に示した前接の結合相手との間に僅かなポーズが入り、各構成語が元のアクセント型を保持する。

e.g. [演奏会]中止：えんそうかい(pause)ちゅうし、　（以下同様）
[委員会]開催、[志願者]増加、[私品]持ち込み、[新入社員]募集、
[オリエント文化]発祥、[ダニ]発生、[データ処理]可能

[注1] 例えば「上九一色村宗教問題解決住民対策会議第一回会合議事録」のような複合語は文中で1語の名詞として振る舞うが全体を辞書登録すべきとは考え難い。

[注2] 同じ現象が英語の形容詞相当句にもみられる (e.g. Duo2300c, the-hard-to-find-notebook computer)。ここから、高低アクセントの平板化がその範囲がひとまとまりの単位であることを示す普遍的な手段となっている可能性が推察される。

上例と同一の用言派生名詞が後接する場合でも、上例のように前接名詞が対象格に該当しない場合、アクセント結合の基本規則のほうに準拠したアクセント型となることがある。

e.g. [大量]発生：たいりょうはっせい

同じ語の組合せであっても、第3の語が連接し意味的にその第3の語と先に結合する場合には、元々アクセント結合の基本規則に準拠していたのが、準拠しなくなる場合もある。

e.g. データ処理*の際には*、：でーたしょり

例外データ処理*の際には*、：れいがいでーた(pause)しょり

上述のアクセント結合の基本規則に準拠する複合語とそれに違反する複合語とでは、統語現象、意味現象の違いもみられる。統語面では、1) 尊敬語の接頭辞などの別形態素・別語の挿入不可能／可能、2)サ変語尾の「する」の後接が可能／不可能、等に際だった違いがみられる。

e.g. 1) * 役員会ご決議は、覆せるとは限らない。
ok そのおっしゃっていた演奏会ご開催の折りにはぜひご連絡下さい。
2) ok ダニが大量発生するのは夏だ。
* ダニ発生する。 * 委員会開催する。 * 志願者増加したのはそのためだ。

「データ処理」と「例外データ処理」の対立についても同様の現象が観察される。

e.g. 1) * データ高速処理。 cf. ok 高速データ処理。
ok 例外データ高速処理。
2) ok データ処理する。
* 例外データ高速処理すると速度が落ちる。[注3]

構成語間の意味関係に着目すると、基本規則に準拠する複合語が、様々な特別なニュアンスを生じているのに対して、準拠しない複合語の場合は、構成語間に格助詞「の」を補った場合と同一の意味関係、格関係となっていることがわかる。

e.g. 「計算機処理」： 「計算機を用いた方式で処理すること。計算機的、計算機風の処理。」という意味となり、「方法」のような新たな概念を伴っている。これに対し、「計算機の処理」は、計算機が処理の対象(Object)または動作主(Agent)となる解釈しか存在しない[注4]。
「データ処理」：「対象を一律にデータとみなして一様な処理をする」というくらいの意味となり「計算機処理」とはまた別の概念を伴っている。
「例外データ処理*の際には*」：れいがいでーた(pause)しょり
：間に「の」を挿入して「例外データ*の*処理*の際には*」としたものと全く同じ対象格の関係。
「ロンドン出張」：間に「への」を挿入して「ロンドンへの出張」としたものと全く同じ着点格の関係。「ロンドン的、ロンドン風の出張」のような修飾句的な意味関係は成立しない。

「の」（あるいは「への」「での」等）を挿入した場合と意味関係、格関係が同じということは、2つの構成語が統語的に互いに独立性が高いためだと考えられる。とすると、意味関係、格関係の特徴も、別形態素の挿入が可能であること、複合語全体に活用語尾を付けるのが困難なこと、という前述の統語現象と合わせて、構成語間の統語的独立性の証左とみなすことができる。前接可能な名詞のヴァリエーションを考えても、アクセント結合の基本規則に準拠する複合語では相手がかなり制限されるが[注5]、「開催」、「増大」などの基本規則に準拠しない複合語では、元の動詞における格スロットの選択制約を満たす限り新造語を含むあらゆる名詞と複合しそうである。この事実も構成語間の統語的独立性の証左とみなすことができよう。

本節の結論としては、用言派生名詞が後接する複合語（以下N-V複合語と称す）では、アクセント結合の型の違いと、構成語間の統語的独立性の高さとの間に強い依存関係があること（アクセント結合の基本規則に準拠する複合語とそうでない複合語とを比較すると後者が統語的独立性が高いこと）が指摘できる。

注3 「例外データ処理」で1つの専門術語とみなし、アクセントの基本結合規則に準拠している場合には、「例外データ処理する：れいがいでーたしょりする」が可能。

注4 「処理」をモノ概念と解釈できる文脈では所有格の関係も成立可能かもしれない。

注5 例えば、データ処理の類語として「？材料処理」、「？入力処理」などは*耳慣れない*感じがするのに対し、「開催」、「増大」などは意味制約が守られる限り自由な組合せが可能である。

3 理論的背景　～語彙的な複合語と文法的な複合語

本節では、前節でみた２種のN-V複合語の生成、成立の過程の概略を分析する。もし複合語が本来的にすべて辞書登録されるべきものであるならば、この分析の工学的意義は薄いかもしれない。しかし、２種のうち片方は、複合する相手の名詞が新造語を含むオープンなカテゴリーであり、複合語全体を網羅的に辞書登録が事実上不可能であるため、両者の弁別は工学的に切実な必要性をもつと考えられる。

ここでは、Kageyama93の概略を引用して、両者の違いは２つの構成語が語形成の結果１語となったものであるか、受け身文を作る際の文の生成と同様の統語的な操作の結果生じた名詞連続であるかの違いに帰する分析を行う。前者のモデルによる複合語を語彙的な複合語、後者のモデルによる複合語を統語的な複合語、と呼ぶことにする。

Kageyama93では、ＧＢ理論（Chomsky86）の道具立てを用い、D構造、Ｓ構造を経てLFに至る２段階の統語的派生を認める。この際、動詞は本来、X-bar理論に基づいて高々２つの内項（動作主以外の格要素）を取り込んで格マーカーを従えた動詞句を構成するはずのところを、語彙挿入の起こった底辺の位置に格要素の１つが移動して格マーカーを失う。

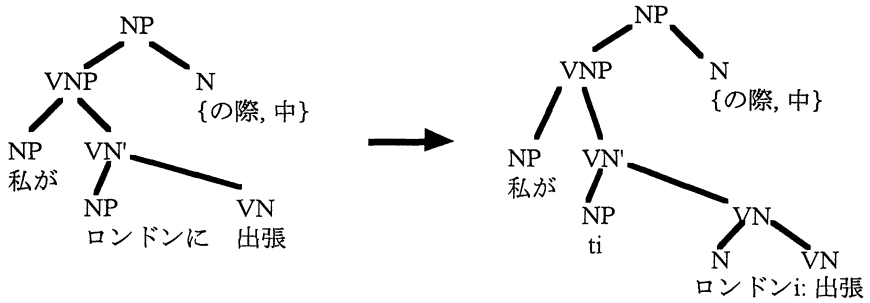

図１　統語的複合語の生成過程の分析例　～ Kageyama93より

図１の統語的派生の過程で、動詞に元々[注6]付加されていた尊敬の概念や、適合する副詞の概念があれば、それらはそのまま残って、「ロンドンご出張」、「志願者大幅増加」のような表層語のならびを形成する。「*ロンドン出張する」が不適格なのは、「する」という時制をともなう活用語尾（屈折辞）が元々動詞に付随していた場合、その段階で語としての単位を閉じてしまい、統語的な操作の結果といえども新たな複合語を構成することがもはやできなくなるからである。一方、「抜け駆け出張[注7]」のような語彙的な複合語は、図１のような派生の過程を経ずに、最初に辞書から語彙挿入される段階で複合語を構成しているため、問題なく「抜け駆け出張する」という形態をとれることになる。

4 辞書のコーディングとその結果の利用

本節では、辞書のコーディングの戦略、「志願者減少」などの佐藤88らの一般則に当てはまらない複合語アクセントを生成する機構、そして、辞書・コーパスに未登録の名詞連続を正しい構文で生成できるようにするための機構を提案する。

4.1 辞書コーディングの際のメリット

ある複合語を辞書登録するか否かの決定には、全体のアクセント型が２種のいずれであるか、構成語間の意味関係・格関係、あるいは尊敬語化の統語テストという複数の、互いに独立な判定条件を利用する。最終結果は、語彙的な複合語か否かの１ビットの情報に還元される。これにより、２節に示した現象面の依存関係を素性評価の制約条件とすることにより、語彙素性のコーディング結果の誤りを機械的に検出することが可能となる。すなわち、少数の優秀なレキシコグラファに頼る

[注6] ここで言う「元々」とは、「D構造の段階で」既に存在していた、との意味。以下同様。

[注7] この例が語彙的複合語であることは、アクセント型からだけでなく、類例として同等の意味関係、様態をとる「？駆け足出張」などが自由に作れないことから判定している。

ばかりでなく、多人数により同時並行してコーディングを進めながら高品質化する手段が得られたことになる。

4.2 複合語アクセントの生成

上記の1ビットの弁別情報を用言派生名詞中にもつ辞書があれば、text to speechにおける2語の複合語の生成は直截的である。未定義語を含むか否かを問わず、複合語区間と推定された形態素列については、全体の基本的な高低アクセントは、第2音節以後、高アクセントで平板に進め、最終音節の1つ手前で下降して低アクセントに落とす。統語的複合語を構成する派生用言が後接した場合だけ、例外として、構成語のアクセント型をそのまま残し、途中に僅かな時間的空隙をはさむ。

このデフォルト戦略に付加する形で、辞書引きされた3語以上の構成語の間でどれが先に結びつくかのコスト計算を行う。たとえば、前述の「例外データ処理の際には」という入力に対しては、「例外データ」、「データ処理」の両方が語彙的複合語として辞書登録され、「例外データ処理」という術語は登録されていない状態で、形容動詞的接頭辞と語彙的複合語の連接よりも、語彙的複合語と他動詞サ変語幹の連接のほうが良いコストとなるような規則を用意する[注8]。

4.3 名詞句生成の概略

1つまたは複数の概念記号に対応して、語彙的複合語の全体が辞書に登録されていた場合、入力された概念木が当該の条件にマッチすることを判定して複合語を生成する。マッチしない場合は連体形の格助詞を補った名詞句を生成する。統語的複合語を生成可能な用言派生名詞が辞書にあった場合、2つの概念の間の格関係、係り先の機能名詞（「中」「際」など）の組み合わせにより名詞連続とするか格助詞を補うかを決める。統語的複合語とした場合、「する」が後接できないので、途中に「ご（御）」を挟み、逆に語彙的複合語では「ご（御）」の挿入が不可、文頭への付加も不自然なので[注9]、尊敬の助動詞「れる／られる」を付加する。

5 おわりに

語彙的な複合語と文法的な複合語の区別を仮定して、複合語全体の辞書登録をせずに名詞連続を生成する条件を提示するとともに、この区別が音声合成のアクセント型の区別および複数の統語条件に対応するため、再現性の高い辞書作りが可能なことを示した。これにより、野村95 におけるユニヴァーサル・レキシコンの2つの利点「直截的でない語彙素性の判断基準を提供」、「辞書登録すべき複合語とそうでない複合語を区別する手段を提供」が、2つの動詞からなる複合語だけでなく、新たに用言派生名詞を含む名詞複合語について示されたことになる。今後は、語彙的複合語の中でも組合せ相手のヴァリエーションが豊富で生産性の高いものの辞書中での生成過程のモデル化や、動作概念をもたない名詞間の結合や、各種の接辞が結合した際のイントネーションの制御、さらに名詞連続生成の条件に対して分析をひろげることにより、知識ベース、生成結果の高精度化と管理性の向上へ結びつけていきたい。

参考文献

Chomsky86: Chomsky, N., "Barriers", MIT Press.

匂坂85, 匂坂芳典、「音声合成のための韻律制御の研究」、早稲田大学理工学研究科博士論文, 1985

佐藤88: 佐藤大和,「辞書情報とアクセント」, 重点領域研究「音声言語」・特定研究「言語情報処理」合同シンポジウム, 1988.12.8

Kageyama93: Kageyama, T., Bunpoo-to Go-keisei (Grammar and Word Formation), Hitsuji-shoboo, 1993

Nomura94: Nomura, N., Jones, D., & Berwick, R., An Architecture for a Universal Lexicon, in Proceedings of COLING94

野村95: ユニバーサルレキシコンの工学的メリット、言語処理学会第1回年次大会予稿集, 1995

[注8] 第1節脚注に示した「上九一色村宗教問題解決／住民対策会議／第一回／会合／議事録」のように構成形態素数の多い例に対しては、用言派生名詞および数量・接辞をブロック分割の手がかりとし、各ブロック内で統語的複合語か語彙的複合語かのアクセント型の違いを生成する。ブロック間には、僅かな時間的空隙をはさむとともに、中間部の低アクセントは、文全体の低アクセント部ほどにはピッチを下げないようにすることにより、全体で大きな1語相当であることを示す。

[注9] たとえば、* 意志ご表示。? ご意志表示。先頭への「ご（御）」の付加が不自然なのは、用言派生名詞でなく前接する名詞のほうを修飾するかにみえる（実態は用言派生名詞にかかる尊敬の接頭辞であるにもかかわらず）ためであろう。