El Mundo紙における連語の自動抽出

著者	宮本 正美
雑誌名	神戸外大論叢
巻	49
号	2
ページ	3-27
発行年	1998-09-30
URL	http://id.nii.ac.jp/1085/00001534/

El Mundo 紙における連語の自動抽出

宮　本　正　美

1. はじめに

この数年スペイン語の世界でもコンピュータ可読性のテキストが大量に入手できるようになる[(1)]と同時に，マイクロプロセッサの性能の飛躍的な向上[(2)]で，誰もが大量のスペイン語テキストを短時間に処理することが可能になってきている。スペイン語の研究や教育において，この大量のテキストの処理が期待されるものの１つに，辞書や語彙集の編集作成などの語彙の分野がある。大量のテキストから単一の語形の頻度を数え，高頻度の語形を抽出することは簡単であるが，テキスト中に頻繁に現われ意味のまとまりをもつ単語の集まりである「連語」を効率的に収集することは容易ではない。本稿では，スペインの日刊紙 *El Mundo* を対象に，この連語一般と，さらに動詞を含んだ連語をできる限り自動的に抽出することを試みる。あわせて自然言語処理の分野で連語抽出のために提案されている「仕事量基準」[(3)]の手法の有効性を検討する。

2. *El Mundo* 紙からの連語の抽出

「連語」には，a partir de のような慣用表現，primer ministro のような定形表現，さらには novela diario のような複合語などが含まれる。これらの連語をテキストから抽出するには，２語，３語，...，Ｎ語という語連続をテキストから順次切り出して，それを頻度順に並べたものを人がチェックするという方法が考えられる。例えば，García Lorca の詩 *Llanto por*

Ignacio Sánchez Mejías の冒頭の部分: A las cinco de la tarde. Eran las cinco en punto de la tarde. Un niño trajo la blanca sábana a las cinco de la tarde. Una espuerta de cal ya prevenida a las cinco de la tarde. Lo demás era muerte y sólo muerte a las cinco de la tarde. というテキストから2語ずつ，3語ずつ切り出して頻度順に並べると，次のリストが得られる：

5 de la	5 de la tarde
5 la tarde	4 a las cinco
5 las cinco	4 cinco de la
4 a las	4 las cinco de
4 cinco de	1 la blanca sábana
1 blanca sábana	1 cinco en punto
1 cinco en	1 de cal ya
1 de cal	1 demás era muerte
1 demás era	1 en punto de
以下略。	以下略。

これらの単語列のうち連語とみなせるものは，(de) la tarde, (a) las cinco, (la) blanca sábana などであろうか。

この節では，*El Mundo* の1995年1月2日から31日までのほぼ1ヶ月分のテキストから，2，3，4，5連語候補を抽出してみる。テキストのサイズは延べ語形数約210万，異なり語形数約7万9千である。3連語の場合を例に，その抽出手順と過程を以下に述べる。

a. テキストの先頭から，1語ずつずらしながら3単語の単語列を頻度を数えながら取り出す。その際次の操作を行った[4]:

1）文字列はすべて小文字化した。

2）最後の単語（ここでは第3単語）の末尾に記号が付く場合，この記号1つは削除した。

3）単語列の最終単語以外（ここでは，第1，第2単語）の末尾にピリオド，疑問符，感嘆符が付く場合，この単語列はスキップした。

4）複合形の動詞（例えば，he cantado）を単語列が含む場合は，1単語増やして（ここでは4単語を）単語列とした。

En los últimos 75 años Japón ha sufrido una veintena de grandes terremotos, que han provocado miles de víctimas mortales. La cronología de estos grandes movimientos sísmicos es ésta:（以下略）を例にとれば，まず，En los últimos の3単語を取り出して，小文字化して en los últimos の単語列として頻度1で記憶する。de grandes terremotos, は末尾のコンマを削除して de grandes terremotos で記憶する。que han provocado miles はこの4単語を単語列とみなす。複合形を含む場合に1語余分に単語列に組み入れるのは，後に説明するように，動詞は不定詞に戻した上で単語列の頻度を集計するからで，que han provocado miles は最終的には que provocar miles の3単語列として集計することになる。terremotos, que han はこの3単語のまま記憶するが，文末から次の文にかけての，víctimas mortales. La と，mortales. La cronología の3単語列は間にピリオドを含むのでスキップする。

b. このようにして，異なり数167159行，つまり，約16万7千行から成る3単語列のリストが得られる。出現頻度の上位20位は次の通り。

同様にして得られた2，4，5単語列のリストの上位20位を挙げる。

3単語列(左端の数字は頻度数):	2単語列：
3867 edicion : a	20780 de la
2786 tipo : doc	10463 en el
956 millones de pesetas	9888 de los
831 tipo : opi	8654 en la
830 uno de los	7268 a la
804 en el que	5406 que se
681 en la que	4837 de las
527 seccion : nacional	4167 que el
526 pagina : 7	4113 : a
511 pagina : 6	3893 a los
505 una de las	3447 lo que

493 texto : el	3011 por el
467 pagina : 1	2880 que la
467 pagina : 2	2786 : doc
455 titulo : el	2774 de un
444 seccion : opinion	2733 que no
427 presidente de la	2691 de que
424 a pesar de	2679 y el
422 de los gal	2645 por la
419 que no se	2632 de su

4単語列：	5単語列：
284 texto : el mundo	124 consejo general del poder judicial
197 el presidente de la	101 de la seguridad del estado
175 los medios de comunicación	92 de la comunidad de madrid
174 en el caso de	70 de los tipos de interés
169 en el que se	68 desde el punto de vista
167 de la audiencia nacional	60 de estado para la seguridad
161 especial para el mundo	57 el consejo general del poder
157 a lo largo de	57 en el caso de que
157 la mayoría de los	56 ex director general de la
153 a la hora de	56 josé amedo y míchel domínguez
152 la comunidad de madrid	56 ministro de justicia e interior
142 de la guardia civil	55 el ex secretario de estado
142 en la que se	54 secretario de estado para la
133 el hecho de que	53 del consejo general del poder
133 los tipos de interés	52 la trama de los gal
124 consejo general del poder	49 presidente del gobierno, felipe gonzález
124 general del poder judicial	46 de los medios de comunicación
120 texto : madrid.- el	45 josé maría alvarez del manzano
117 de amedo y domínguez	43 ex secretario de estado para
116 a pesar de que	42 director general de la seguridad

対象としたテキストが新聞記事のCD-ROMファイルなので，edicion : a, tipo : doc, seccion : nacional, pagina : 7のような本来の印刷された新聞記事には現れないインデックス的な単語（記号）列を一部含む。これらのリストをここに挙げている実頻度順に，上からチェックしながら1000でも2000で

も連語とみなせそうな単語列を拾いあげていくのも1つの方法ではあるが，本稿ではできるだけ連語の可能性の高い単語列が上位にくるように，これらの異なった語数の単語列のリストをそれぞれ関連づけて処理してみたい。

2，3，4，5単語列の頻度順位20位だけを比較するだけでも，単語列が長いほど私たちが連語と考えたくなる単語列の比率が高くなることが分かる[5]。これは，次節で述べる「仕事量基準」の手法が仮定する「長くて頻度の多い単語列ほど連語の可能性が大きい」という考えを裏付けている。しかし，de la 20780, en el 10463, a la 7268, por el 3011といった単語列は，前置詞と冠詞の組合せとして，また，en el que 804, en la que 681などは，おそらく前置詞と関係詞の組合わせとして語学的には興味のあるデータであるかも知れないが，連語を抽出するという目的から外れるであろう。つまり，2単語列リストの上位に見られるような，機能語どうしの組合わせなどは，語学的見地からの前処理で連語の候補リストから排除しておくべきである。

c. さしあたり，次の1）から5）の（語学的）制約にあてはまる単語列は，前処理として，単語列のリストから削除する。なお，冠詞，前置詞，接続詞，関係詞，非主格人称代名詞（＝弱形人称代名詞），所有形容詞前置形，指示形容詞に含まれる単語をここでは，非連語的単語と呼ぶことにする[6]

1）非連語的単語の2連続で終わる単語列。この場合，noも非連語的単語列に加える。ただし，de queで終わる単語列は削除しない[7]。

2）非連語的単語の3連続で始まる単語列[8]。

3）que＋非連語的単語1つで始まる単語列。

4）接続詞で終わる単語列。

5）”’~^()＜＞以外の記号を含む単語列。（ただし，数字に付く,. は除く）

d. もう1つの前処理として，リスト中の動詞の変化形を不定詞に戻して，動詞を含む同じ単語列は集約し，頻度数を数えなおす[9]。名詞句表現に対して相対的に頻度の低い動詞は各活用形ごとに分散しているので，どういう

動詞句表現の頻度が高いかを知るためには，不定詞に集約する必要がある。

a. から d. の前処理を施して，2，3，4，5単語列リストに現われた単語列を実頻度数順に並べると次のようになる。

2単語列：	3単語列：
2691 de que	956 millones de pesetas
2042 que ser	424 a pesar de
1790 ser el	422 de los gal
1771 ir a	370 * tratarse de
1717 no ser	354 en los últimos
1540 el mundo	351 por lo que
1521 millones de	348 el presidente de
1412 decir que	328 a través de
1394 ser la	323 el presidente del
1391 ser un	309 amedo y domínguez
1298 tener que	303 * convertirse en
1250 que tener	298 por su parte
1240 que estar	292 presidente del gobierno
1187 el presidente	291 a partir de
1172 todos los	288 la mayoría de
1170 uno de	280 * irse a
1129 el gobierno	278 el caso de
1122 más de	274 la posibilidad de
1112 del gobierno	260 en el caso
1100 de pesetas	258 banco de españa

4単語列：	5単語列：
175 los medios de comunicación	124 consejo general del poder judicial
174 en el caso de	101 de la seguridad del estado
167 de la audiencia nacional	92 de la comunidad de madrid
161 especial para el mundo	70 de los tipos de interés
157 a lo largo de	68 desde el punto de vista
153 a la hora de	60 de estado para la seguridad
152 la comunidad de madrid	57 el consejo general del poder
142 de la guardia civil	57 en el caso de que
133 el hecho de que	56 josé amedo y míchel domínguez
133 los tipos de interés	56 ministro de justicia e interior
124 consejo general del poder	55 el ex secretario de estado

124 general del poder judicial	53 del consejo general del poder
117 de amedo y domínguez	52 la trama de los gal
116 a pesar de que	46 de los medios de comunicación
116 del banco de españa	45 josé maría alvarez del manzano
116 el banco de españa	44 no tener nada que ver
116 la seguridad del estado	43 ex secretario de estado para
115 el presidente del gobierno	42 director general de la seguridad
109 de los fondos reservados	42 el secuestro de segundo marey
108 una moción de censura	41 subida de los tipos de

3連語の＊ tratarse deや＊ irse aの＊印は，元のテキストでは，se trata de, se va a, se fueron a, me voy aなどのように，＊印のところに再帰代名詞があったことを示している。前処理のお陰で，それぞれのリストの上位に連語（と考えたいもの）がかなり並ぶようになったのが分かる。特に，動詞を活用形から不定詞に集約した結果，2，3単語列にはかなりの動詞が集まり，5単語列ではno tener nada que verという連語を上位で抽出している。しかし，2，3，4，5単語列の出現頻度の差が大きく，このまま4つのリストをまとめて頻度順に並べなおしたのでは，単語数の多い単語列の連語はどうしても過小評価されてしまう。そこで，実頻度数を何らかの基準で評価しなおすことが考えられる。

3. 仕事量基準

自然言語処理の分野では，連語をテキストから自動的に抽出する方法が最近いくつか提案されている。[10] 自動的に抽出するために何らかの基準となる評価関数を用いて，ある語連続を連語とみなすか否かの決定をしている。ここでは，私たちがスペイン語の文法分析において重要な尺度と考えている，対象（あるいはその構成要素）の出現頻度と長さ（語数）に基づく[11]「仕事量基準」という関数を用いた手法を利用してみる。

北他（1993）は，コーパステキスト中に$|\alpha|$個の単語から成る単語列αが$n(\alpha)$回出現するとき，次のような関数$K(\alpha)$を仕事量基準として定義

している:

$K(\alpha) = (|\alpha| - 1) \times n(\alpha)$

$|\alpha|$ ... 単語列 α の長さ（単語数）

$n(\alpha)$ ... 単語列 α のコーパス中での出現回数

コーパス中の $|\alpha|$ 個の単語から成る単語列 α を1つずつ処理すれば，α に比例した仕事量が必要であるが，α を1つのまとまった表現として処理すれば，仕事量は1で済む。つまり，$(|\alpha| - 1)$ 分の削減になる。$K(\alpha)$ の値の大きな単語列を連語として抽出するというのは，長くて頻度の多い単語列ほど連語の可能性が大きいという仮定に基づいているが，これは，前節で確かめたように，妥当な考えだと言える。しかし，ある単語列は他の単語列の部分列のことが少なくないので，単純に頻度数で比較することができない。例えば，a partir, a partir de, a partir de ahora の3つの単語列の出現頻度はそれぞれ，378, 291, 53であったが，それぞれの単語列の仕事量基準は

K(a partir)=(2−1)×378=378

K(a partir de)=(3−1)×291=582

K(a partir de ahora)=(4−1)×53=159

となり，a partir の仕事量基準は a partir de ahora より大きくなる。しかし，a partir が参照された378回の内，a partir de の部分単語列として参照された回数が291回あるので，純粋に a partir が参照されたのは（378-291）回ということになる。同様に，a partir de が純粋に参照されたのは（291-53）回である。したがって，それぞれの仕事量基準を再計算すると

K(a partir)=(2−1)×(378−291)=87

K(a partir de)=(3−1)×(291−53)=476

K(a partir de ahora)=(4−1)×53=159

このように，部分単語列に対して再計算をすることによって，比較的均整

のとれた値が算出されると考えられる。しかし，この再計算を部分単語列の関係にあるすべての単語列に対して行えば，計算量が爆発的に増大するので，実際には，次の2つの操作で近似的に行っている。

1．再計算は同じ単語系列中の隣り合った単語列どうしに限る：a partir は，長さが2つ異なる a partir de ahora ではなく，1つだけ異なる a partir de とのみ行う。

2．再計算は1度に限る：a partir は a partir de（頻度数291）以外にも a partir del（84）などの部分単語列であるが，再計算は，一番頻度の高い a partir de との間で1度だけ行う。同様に，a partir de の再計算は，a partir de ahora (53), a partir de hoy (20), などの部分単語列だが，a partir de ahora との間の1度に限っている。

そこで，この仕事量基準を利用して先ほど提示したリストの実頻度数を評価しなおしてみよう。評価しなおすためには，部分単語列の関係にある単語列をまとめなければならない。実頻度数の仕事量基準への再評価にあわせて，単語列（＋評価数値）をシソーラス風のリストに作りなおす。ここでは，2，3，4，5単語列のリストのそれぞれ上位500位で作ってみた。結果は次のようになる：

2480 de que
- 422 después de que
- 330 antes de que
- 38 hecho de que
  - 348 el hecho de que
    - 68 en el hecho de que
    - 48 por el hecho de que
- 12 pesar de que
  - 348 a pesar de que
- 2 posibilidad de que
  - 279 la posibilidad de que
    - 44 ante la posibilidad de que
- 60 caso de que
  - 9 el caso de que

228 en el caso de que
72 en caso de que
50 convencido de que
123 estar convencido de que
1861 que ser
362 lo que ser
262 que ser el
238 que ser un
234 que ser la
174 tener que ser
160 que ser una
1659 ser el
262 que ser el
178 no ser el
56 ser el caso
72 ser el caso de
56 como ser el caso de
18 como ser el caso
56 como ser el caso de
1520 ir a
446 que ir a
84 que ir a ser
66 lo que ir a
268 no ir a
87 que no ir a
266 ir a ser
84 que ir a ser
1485 no ser
464 que no ser
186 no ser un
178 no ser el
146 no ser la
1344 el mundo
318 todo el mundo
111 en todo el mundo
108 de todo el mundo
242 a el mundo
150 a el mundo que
84 ayer a el mundo que
36 manifestar a el mundo que

24 ayer a el mundo
84 ayer a el mundo que
24 para el mundo
483 especial para el mundo
216 en el mundo
66 en el mundo de
42 el mundo que
150 a el mundo que
84 ayer a el mundo que
36 manifestar a el mundo que
156 de el mundo
565 millones de
1790 millones de pesetas
183 millones de pesetas en
111 millones de pesetas a
87 millones de pesetas de
42 200 millones de pesetas
56 de 200 millones de pesetas
84 millones de pesetas que
81 millones de pesetas para
78 mil millones de pesetas
60 millones de pesetas por
222 millones de dólares
1412 decir que
中略
下位の行が多い単語列を１つ：
255 general del
4 consejo general del
0 consejo general del poder
496 consejo general del poder judicial
228 el consejo general del poder
212 del consejo general del poder
48 al consejo general del poder
0 el consejo general del
228 el consejo general del poder
6 del consejo general del
212 del consejo general del poder
48 presidente del consejo general del
48 vicepresidente del consejo general del
0 general del poder

0 consejo general del poder
496 consejo general del poder judicial
228 el consejo general del poder
212 del consejo general del poder
48 al consejo general del poder
0 general del poder judicial
496 consejo general del poder judicial
120 general del poder judicial cgpj
72 secretario general del
84 el secretario general del
56 el secretario general del pp
60 secretario general del partido
16 general del estado
120 fiscal general del estado
88 el fiscal general del estado
60 al fiscal general del estado

中略

動詞の部分を1つ：

202 tener en
192 tener en cuenta
120 tener en cuenta que
44 que tener en cuenta que
12 que tener en cuenta
64 haber que tener en cuenta
44 que tener en cuenta que
60 tener en cuenta la
122 que tener en
12 que tener en cuenta
64 haber que tener en cuenta
44 que tener en cuenta que

以下略。

この種のシソーラス風リストは連語の文法構造の類似したものを下位分類していることにもなるので，語学的資料としての価値もある。実頻度数を仕事量基準で評価しなおした数値順に連語（候補）を並べなおしてみると，上位80位は次のようになる：

2480 de que
1861 que ser
1790 millones de pesetas
1659 ser el
1520 ir a
1485 no ser
1412 decir que
1344 el mundo
1277 ser la
1272 ser un
1082 tener que
1075 que estar
1034 que tener
1033 uno de
1018 el gobierno
1013 más de
999 ser que
965 todos los
950 ser una
902 sin embargo
863 de madrid
845 después de
839 el presidente
833 el pasado
820 del gobierno
808 no poder
771 llegar a
734 que ir
730 de los gal

725 que poder
722 creer que
712 ya que
705 no tener
703 volver a
702 el juez
700 haber que
676 tratar de
657 estar en
656 * tratarse de
654 el presidente de
645 * encontrarse
619 asegurar que
616 a pesar de
612 en españa
606 * convertirse en
603 que hacer
596 por su parte
591 * poderse
585 del estado
569 pero no
565 millones de
562 la primera
557 afirmar que
556 parte de
547 * hacerse
544 del pp
542 en los últimos
539 más que
533 * producirse
533 antes de
519 no haber
517 poder ser
508 del psoe
502 felipe gonzález
502 la mayoría de
499 del partido
496 * irse a
496 consejo general del poder judicial
485 el ex
483 especial para el mundo
480 por otra parte
476 a partir de
471 ser de
468 tener un
467 todas las
464 que no ser
463 ayer en
459 ser los

656 a través de　　456 la vida

実頻度数順では重なり合わなかった上位20位までの4つの単語列が，仕事量基準によって評価しなおした数値順では，2単語列と3単語列はかなり重なってきている。20位までの3単語列の中でも特に連語的な millones de pesetas, * tratarse de, a través de, a pesar de, * convertirse en, por su parte, la mayoría de, por otra parte, がいずれもリストに入ってきていることは，この再評価の成果であろう。しかし，4，5単語列が especial para el mundo, consejo general del poder judicial 以外1つも入ってこないというのは，単語数が増えることへの重みづけがまだ不十分であるか，計算量がふえても近似値計算の対象をもう少し増やすのが望ましいことを示しているのではないだろうか。

ここでは，近似値計算の2。で「再計算を1度に限っている」のを改め，直接下位（つまり，語数の差が1）の単語列すべてと再計算をしてみよう：

$$K(\alpha) = (|\alpha| - 1) \times (n(\alpha) - \sum_{i=1}^{m} n([\alpha, i]))$$

$|\alpha|$ ……単語列 α の長さ（単語数）

$n(\alpha)$ ……単語列 α のコーパス中での出現回数

$[\alpha, i]$ ……単語列 α の直接下位の単語列群の頻度順 i 番目の単語列

m ……指定した頻度上位，例えば，上位500の連語に含まれる範囲での i の最大値

先の例で言えば，a partir の再計算は，a partir de (291) 以外に a partir del (84) とも，また，a partir de の再計算は，a partir de ahora (53) だけでなく a partir de hoy (20) とも行うことを意味している（500位まででは，a partir, a partir de の直接下位の単語列は2つずつである)。

K(a partir)＝(2－1)×(378－291－84)＝3

K(a partir de)＝(3－1)×(291－53－20)＝436

K(a partir de ahora)＝(4－1)×53＝159

この計算法を採用すると上位80位は次のようになる：

2163 de que
1602 millones de pesetas
1520 ir a
1504 ser el
1412 decir que
1344 el mundo
1327 que ser
1230 no ser
1179 ser un
1133 ser la
1082 tener que
1034 que tener
999 ser que
959 que estar
950 ser una
902 sin embargo
863 de madrid
858 el gobierno
845 después de
839 el presidente
833 el pasado
820 del gobierno
790 más de
771 llegar a
734 que ir
730 de los gal
722 creer que
712 ya que
705 no poder
704 todos los
703 volver a
702 el juez
700 haber que
686 uno de
644 que poder
625 no tener
619 asegurar que
612 en españa
606 * convertirse en
603 que hacer
596 por su parte
591 * poderse
585 del estado
574 a pesar de
569 pero no
565 millones de
557 afirmar que
547 * hacerse
544 del pp
539 más que
533 * producirse
533 antes de
519 no haber
508 del psoe
502 felipe gonzález
499 del partido
496 * irse a
496 consejo general del poder judicial
485 el ex
483 especial para el mundo
480 por otra parte
471 ser de
468 tener un
467 todas las
464 que no ser
463 ayer en
459 ser los
456 no * poderse

676 tratar de	448 años de
657 estar en	446 general de
656 * tratarse de	443 de ser
656 a través de	442 la mayoría de
654 el presidente de	440 por parte de
645 * encontrarse	436 a partir de

再計算の結果，主に上位の2単語列の数値が少し低くなった以外は大きな変化は見られない。このリストに新たに入った単語列は，no * poderse, años de, general de, de ser, por parte de で，落ちたのは，la primera, parte de, en los últimos, poder ser, la vida の5つだけである。

そこで，単語列の長さに対する重みづけをおもいきって，$(|\alpha|-1)^2$ で2乗する操作もあわせて行うことにする：

$$K(\alpha) = (|\alpha|-1)^2 \times (n(\alpha) - \sum_{i=1}^{m} n([\alpha, i]))$$

$$K(\text{a partir}) = (2-1)^2 \times (378-291-84) = 3$$

$$K(\text{a partir de}) = (3-1)^2 \times (291-53-20) = 872$$

$$K(\text{a partir de ahora}) = (4-1)^2 \times 53 = 477$$

この操作で次のリストを得る：

3024 millones de pesetas	912 no * poderse
2163 de que	902 sin embargo
1984 consejo general del poder judicial	
	884 la mayoría de
1616 de la seguridad del estado	880 el ex secretario de estado
1520 ir a	880 por parte de
1504 ser el	872 a partir de
1472 de la comunidad de madrid	863 de madrid
1460 de los gal	858 el gobierno
1449 especial para el mundo	848 del consejo general del poder
1412 decir que	846 de la guardia civil
1344 el mundo	845 después de
1327 que ser	844 después de que
1312 * tratarse de	839 el presidente
1312 a través de	837 el secretario general de
1308 el presidente de	837 la posibilidad de que
1305 a la hora de	833 el pasado

1230 no ser
1212 * convertirse en
1192 por su parte
1179 ser un
1170 a lo largo de
1148 a pesar de
1133 ser la
1120 de los tipos de interés
1088 desde el punto de vista
1082 tener que
1053 en el caso de
1044 a pesar de que
1034 que tener
999 ser que
992 * irse a
960 de estado para la seguridad
960 por otra parte
959 que estar
950 ser una
936 el hecho de que
936 el presidente del gobierno
928 que no ser
912 el consejo general del poder
912 en el caso de que
832 el presidente del
832 la trama de los gal
828 de todos los
820 del gobierno
810 de los fondos reservados
804 que ir a
790 más de
771 llegar a
768 amedo y domínguez
765 de la unión europea
765 el banco de españa
756 que tener que
748 de la sociedad
747 en los últimos años
736 de los medios de comunicación
736 una serie de
734 que ir
724 * encontrarse en
724 lo que ser
722 creer que
712 ya que
708 presidente del gobierno
705 no poder
704 josé maría aznar

大胆に，長さの重みづけをすると，a lo largo de, desde el punto de vista, a pesar de que, el hecho de que, el presidente del gobierno, en el caso de que, el ex secretario de estado, después de que, la posibilidad de que, la trama de los gal, el banco de españa, en los últimos años, una serie de, * encontrarse en, presidente del gobierno など4，5単語を中心に連語とみなしたい単語列が入ってくる一方，落ちるものに特に連語的なもの（volver a, haber que, * encontrarse, asegurar que, en españa, afirmar que, * producirse, no haber を連語的とみるかどうか）は少ないと言える。このことから，実頻度数を再評価する際には，長さ（単語数）の重みづけを十分にすることがよいと考えられる。

4. 動詞連語の抽出

第2節でも述べたように，相対的に出現頻度の低い動詞の現われる連語を抽出するには，特別の工夫が必要である。動詞の活用形リストを参照しながら，先のリストの単語列中の動詞活用形に文法データを付け，このデータを参照しながら活用形を不定詞にもどして，単語列とその頻度数を集約し，頻度順に並べなおす。具体的には，以下の手順で抽出した：

a. 動詞（不定詞）のリストを作る（不規則動詞の場合はその活用パターンを示す番号を付ける）。今回は7430個の動詞（不定詞）のリストを作成した。

b. 語尾の後ろに，法，時制，人称/数の文法データを付けた，-ar，-er，-irの各動詞の規則活用リストを作る。このリストの構成は以下の通り（-ar動詞の場合）：

ar(0.inf.0)
ando(0.ger.0)
ado(0.pp.0)
o(i.p.1s)
as(i.p.2s)
a(i.p.3s)(m.p.2s)
amos(i.p.1p)(i.ind.1p)
áis(i.p.2p)
an(i.p.3p)
以下略。

c. a.の不定詞リスト（と不規則動詞の場合はその番号）とb.の規則活用のリストを利用して，複合形も含めて全活用形が得られるようなスクリプトを-ar，-er，-ir動詞ごとに書き，不定詞リストの7430個の動詞の全活用形（文法データ付き）リストを作る。ただし，過去分詞形は削除した。-ar動詞のリストは644946語形，-er動詞は44344語形，-ir動詞は48373語形になった。

d. 動詞の活用形は名詞その他の品詞語と同形のものが少なくないので，他の品詞語（複数形も含む）と重複する活用形のリストと照合して，c.の全

活用形リストからそのような活用形は削除する。[12]

e. この活用形リストの文法データを単語列リストの該当する単語に付与し，動詞を含む行を抜き出して，非連語的単語列を削除する。

この段階で，例えば3単語列リストは次のようになっている：

46 se limitó(limitar.i.ind.3s) a
18 se limita(limitar.i.p.3s)(limitar.m.p.2s) a
5 se han limitado(limitar.i.per.3p) a
2 me limito(limitar.i.p.1s) a
など。

f. 文法データを利用して（複合形も含めて）活用形を不定詞に戻す。なお，その際再帰動詞と判断できれば（上の場合の se limitó などは直前の se と，もし se le limita なら，その前の se とデータ中の 3s の 3 から，me limito なら me とデータ中の 1s から再帰動詞であることが分かる），不定詞に se を付け，再帰代名詞のあった場所に＊印を残す。この操作も動詞句表現の頻度を高めるためである[13]：

46 ＊ limitarse a
18 ＊ limitarse a
5 ＊ limitarse a
2 ＊ limitarse a
など。

g. 不定詞に戻された単語列を含めてリスト全体をマージして頻度数を数えなおす：

99 ＊ limitarse a

以上，a. から g. までの処理で得られた2，3，4，5語の動詞連語（候補）のリストの上位20位は次の通り：

2連語：	3連語：
2042 que ser	424 a pesar de
1790 ser el	370 ＊ tratarse de
1771 ir a	303 ＊ convertirse en
1717 no ser	291 a partir de
1412 decir que	280 ＊ irse a

1394 ser la	251 que ir a
1391 ser un	232 que no ser
1298 tener que	228 no * poderse
1250 que tener	216 que tener que
1240 que estar	185 del poder judicial
1087 ser que	181 * encontrarse en
1030 ser una	181 lo que ser
985 que ir	180 llevar a cabo
935 no poder	171 que haber que
871 haber que	165 que estar en
845 llegar a	163 no ir a
837 no tener	161 ir a ser
826 * encontrarse	147 tener en cuenta
822 estar en	137 ser uno de
819 * poderse	132 que no tener

4連語：	5連語：
124 consejo general del poder	124 consejo general del poder judicial
124 general del poder judicial	57 el consejo general del poder
116 a pesar de que	53 del consejo general del poder
82 estar a punto de	44 no tener nada que ver
59 no tener nada que	37 presentar una moción de censura
57 lo cierto ser que	36 tener nada que ver con
53 a partir de ahora	32 * darse cuenta de que
53 tener nada que ver	30 general del poder judicial cgpj
51 tener en cuenta que	24 ser la primera vez que
49 nada que ver con	23 titular del juzgado de instrucción
46 tener que ver con	22 dar la impresión de que
45 * darse cuenta de	17 si * tenerse en cuenta
44 * llevarse a cabo	16 haber que tener en cuenta
44 llegar a un acuerdo	16 maquinación para alterar el precio
42 no * tratarse de	16 para alterar el precio de
41 estar convencido de que	15 lo que pasar ser que
41 presentar una moción de	15 no caber duda de que
40 la verdad ser que	15 que tener que ver con
39 que no * poderse	14 como ser el caso de
38 ser el caso de	14 estar a la espera de

del poder judicial などの，poder, titular del juzgado de instrucción の titular などを除けば，連語と考えたい動詞句表現の単語列がかなり抽出されていると言えるだろう。

最後に，これら４つの動詞連語（候補）リストの各上位500位までを，前節最後で提案した手法で実頻度数を評価しなおした数値順に上位80位を並べてまとめてみる：

1984 consejo general del poder judicial
1861 que ser
1659 ser el
1520 ir a
1485 no ser
1283 decir que
1277 ser la
1272 ser un
1232 a pesar de
1212 * convertirse en
1156 * tratarse de
1082 tener que
1075 que estar
1034 que tener
1008 a pesar de que
964 * irse a
952 a partir de
936 ser que
912 el consejo general del poder
912 no * poderse
892 que ir a
864 ser una
848 del consejo general del poder
808 no poder
756 que tener que
734 que ir
725 que poder
724 * encontrarse en
722 creer que
720 que no ser
715 llegar a
592 llevar a cabo
592 presentar una moción de censura
585 estar a punto de
584 que estar en
576 tener nada que ver con
552 * poderse
548 * encontrarse
545 asegurar que
532 ir a ser
519 no haber
517 poder ser
512 * darse cuenta de que
512 * referirse a
508 afirmar que
508 ser uno de
480 general del poder judicial cgpj
476 que no poder
472 que no tener
470 * hacerse
464 no ir a
464 que ser el
460 que no haber
456 * producirse en
448 lo que hacer
441 lo cierto ser que
438 * producirse
436 que ser la
436 ser de
431 que haber
423 a partir de ahora
423 ser los

705 no tener	409 tener un
704 no tener nada que ver	404 que ser un
700 haber que	396 * llevarse a cabo
657 estar en	389 de ser
657 volver a	388 lo que estar
636 que haber que	384 ser la primera vez que
626 tratar de	384 tener en cuenta
612 lo que ser	381 considerar que
603 que hacer	378 no * tratarse de

4，5単語列からも，no tener nada que ver, presentar una moción de censura, estar a punto de, tener nada que ver con, * darse cuenta de que, lo cierto ser que, a partir de ahora, * llevarse a cabo, ser la primera vez que, no * tratarse de など連語とみなしたい単語列が拾い上げられていることからも，仕事量基準の計算の際に1．直接下位のすべての単語列と再計算を行い，2．長さ（単語数）に対する重みづけを（単語数-1）^{2} によって十分にすれば，2から5語の異なる単語数の連語（候補）をかなり正確に抽出できることが分かる。

5. 結び

従来，主に編纂者の語学的経験や直観によっていた，重要語彙，高頻度連語の抽出が，ここに提示したような手法によれば，かなり客観的にしかもローコストで自動的に実現できることが明らかになった。また，仕事量基準による実頻度数の再計算による手法も，スペイン語の場合適当な前処理と長さ（単語数）への重みづけを十分に行えば，かなり有効であることも分かった。今回は連語一般と動詞連語の抽出を試みたが，形容詞連語，関係詞連語などを抽出することで，それぞれの品詞語の研究にも十分寄与することが予想される。あるいは名詞データも利用して，現代スペイン語の特徴の1つである novela diario のような複合語[14]の抽出も容易に実現できる。今後の課題としては，データ付与処理，活用形の不定詞への変換処理の部分での同形異品詞

語，同形異義語に個別的な対応をすることで，連語抽出の精度をより高めることが考えられる。

注：

（1） 宮本（1997：77）参照。

（2） 1971年の750KHzから，1998年の1GHzマイクロプロセッサの発表まで，動作周波数だけを見れば，約1500倍になっている（青木他（1998：693）参照。）

（3） 北他（1993）参照。

（4） 以下，すべての処理はAWK（一部Perl）のスクリプトとシェルスクリプトで行った。

（5） 3単語列の上位には，文字列：文字列という特殊な単語列がたくさん入っているので，20位以下のそれ以外の単語列をさらに挙げると：

414 de lo que
381 de que el
354 en los últimos
351 por lo que
348 el presidente de
328 a través de
323 el presidente del
309 amedo y domínguez
300 lo que se

これらを加えて比較しても，2，3，4，5単語の単語列順に連語（と考えられるもの）の比率が高くなっていくことに変わりはない。

（6） これらの品詞群の単語は，不定冠詞を除けば非強勢語とされている。不定冠詞は，従来強勢語とされている：Academia（1973:2.3.3），Alarcos（1994;79）。un libroが「1冊の本」のときunに強勢がくるとしても，「（ある）本」のときにもunに強勢がくるとはおもえないが。

（7） a pesar de que, después de que, el hecho de que, la posibilidad de que など de que で終わる単語列は構文的に連語とみなしたいものが多いので削除しないことにした。nuestro, vuestro（とその変化形）は一律に所有形容詞前置形として扱う。なお，en contra de, en cuanto a, pese a que, por lo que, por lo cual, tal y como, de vez en cuando を拾い上げることにした。

（8） por lo que respecta を拾い上げた。2，3，4，5単語列の上位1000行のチェックでは，数単語列を例外として拾い上げるだけで済んでいるが，例外が少なくなるように，これらの制約をさらに洗練する必要があるだろう。

（9） 手順は4。節を参照。

（10） Church（1990）は相互情報量（mutual information）という確率関数を用いて単語の結びつきの強度の大きい単語対を求め，Smadja（1993）は単語間の強度に加えて，2単語間に何語現れるかという単語間の距離なども考慮して連語を自動的に抽出する手法を提案している。

no sólo... sino (también)... のような不連続な連語も含めて自動抽出しようという試みには，尾本／北（1996）や小田／北（1997）などがある。語彙の（半）自動抽出の最近の動向を概説するOoi（1998:75-80）は，語学的手法と統計的手法の併用を効果ありとしている。

(11) 例えば，宮本（1997）は，名詞句中のスペイン語形容詞の位置が，名詞と形容詞の長さ（音節数）とアクセント位置，さらに形容詞の頻度数という基準でかなり説明できることを明らかにしている。

(12) ただし，abra, beba, cabe, cedo, coma, crece, crea, debe, di, dije, empiezo, enseña, eras, haya, haz, manda, miente, ordenando, pasa, pienso, pongo, quedo, quita, reclame, resultas, recibí, referí, sacas, saluda, suba, tarda, ten, toca toma, trata, vale, ve など動詞としての頻度が明らかに高いと思われる語形については削除していない。しかし，me caso の caso, yo como の como, para el coche の para などは文法データが削除される。el caso, un caso でなければ，caso には（casar.i.p.1s）を付けておくといった個別的な処理が必要になる。

(13) さらに述べれば，empecemos (empecer.i.p.1p) (empezar.s.p.1p)，vende (vender.i.p.3s) (vender.m.p.2s) (vendar.s.p.1s) (vendar.s.p.3s)，vengas (venir.s.p.2s) (vengar.i.p.2s)，ve (ir.m.p.2s) (ver.i.p.3s) のように，同形異義の活用形にはデータがいくつか付いている。活用形から不定詞への変換処理が現在のところ，左端のデータしか照合しないので，同形異義の活用形リストを利用して，これらの例のように，明らかに一方の不定詞の動詞の頻度が高い場合は，データを入れ替える処理も組み込んでいる：empecemos (empezar.s.p.1p) (empecer.i.p.1p)，vende (vender.i.p.3s) (vender.m.p.2s) (vendar.s.p.1s) (vendar.s.p.3s)，vengas (venir.s.p.2s) (vengar.i.p.2s)，ve (ver.i.p.3s) (ir.m.p.2s)。しかし，fue (ir.i.ind.3s) (ser.i.ind.3s) のような，いずれの頻度も高いものはそのままにせざるをえない。

(14) 宮本（1996:176-177）参照。

資　料：

分析対象テキスト

El Mundo 1995, Primer Semestre, Unidad Editorial, S.A., CD-ROM, 1997.

参考文献

Academia Española, Real (1973): *Esbozo de una nueva gramática de la lengua española,* Espasa-Calpe, S.A., 1973.

Alarcos Llorach, Emilio (1994): *Gramática de la lengua española,* Espasa-Calpe, S.A., 1994.

青木直明，Hofstee, H.Peter, Dhong, Sang (1998): "GHz マイクロプロセッサ", 情報処理, Vol.39, No.7, pp.693-698。

Church, Kenneth Ward (1990): "Word Association Norms, Mutual Information, and Lexicography", *Computational Linguistics,* Vol.16, No.1, pp.22-29.

北研二，小倉健太郎，森元逞，矢野米雄（1993）: "仕事量基準を用いたコーパスからの定型表現の自動抽出", 情報処理学会論文誌, Vol.34, No.9, pp.1937-1943。

宮本正美（1996）: "近代および現代のスペイン語", 山田他（1996），第4章：125-188。

——（1997）: “ABCにおける形容詞の位置”，神戸外大論叢，Vol.48，No.3，pp.77-98。

尾本貴志，北研二（1998）: “距離反比例型スコアを導入したコロケーションの自動抽出”，情報処理学会自然言語処理研究会報告，112-11，pp.75-82。

小田祐樹，北研二（1997）: “単語の出現位置情報を用いたコーパスからのコロケーションの自動抽出”，情報処理学会自然言語処理研究会報告，121-11，pp.75-82。

Ooi, Vicent B.Y.(1998): *Computer Corpus Lexicography,* Edinburgh University Press, 1998.

Smadja, Frank（1993）:“Retrieving Collocations from Text: Xtract”, *Computational Linguistics,* Vol.19, No.1, pp.143-177.

山田善郎他（1996）: スペインの言語，同朋舎出版，1996。