ELUA
ISSN 2171-6692
Núm. 35, 2021, págs. 53-72
https://doi.org/10.14198/ELUA2021.35.3

# PALABRAS MONOCOLOCABLES EN ESPAÑOL Y EN ITALIANO: ¿CÓMO IDENTIFICAR PALABRAS CON UNA COLOCABILIDAD MUY RESTRINGIDA?

## MONOCOLLOCABLE WORDS IN SPANISH AND ITALIAN: HOW TO IDENTIFY WORDS WITH VERY LIMITED COLLOCABILITY?

Petr Čermák
Universidad Carolina, Praga, República Checa
petr.cermak@ff.cuni.cz
https://orcid.org/0000-0002-9298-1701

Zora Obstová
Universidad Carolina, Praga, República Checa
zora.obstova@ff.cuni.cz
https://orcid.org/0000-0002-1678-6947

**Resumen**
El artículo estudia el fenómeno de la colocabilidad extremadamente restringida en español y en italiano, fenómeno poco estudiado hasta ahora. Su objetivo principal es poner a prueba un método automatizado capaz de identificar las llamadas palabras monocolocables (entendidas como palabras cuyo potencial para formar parte de colocaciones es enormemente limitado) en los corpus lingüísticos. Además, ofrece un listado de palabras de este tipo en español e italiano, así como una tipología de las combinaciones

**Abstract**
The article deals with the so far little-explored phenomenon of extremely limited collocability in Spanish and Italian. Its main aim is to verify the effectiveness of an automatic method capable to extract the so-called monocollocable words, i.e. words whose potential to form collocations is very restricted, from corpora. Moreover, it offers a list of monocollocable words in Spanish and Italian, a typology of fixed combination in which they occur (different types of phraseological units/idioms,

**Para citar este artículo:** Čermák, Petr y Obstová, Zora (2021). Palabras monocolocables en español y en italiano: ¿cómo identificar palabras con una colocabilidad muy restringida? *ELUA*, 35: 53-72. https://doi.org/10.14198/ELUA2021.35.3

Recibido: 26/05/2020, Aceptado: 22/10/2020

fijas en las cuales aparecen (diferentes tipos de locuciones, términos) y un comentario de sus propiedades básicas. Todo ello permite comparar el fenómeno de la colocabilidad restringida en los dos idiomas.
PALABRAS CLAVE: colocabilidad restringida; locuciones; términos; palabras monocolocables; lingüística de corpus.

terms) and a commentary on their properties, which allows to compare the phenomenon in both languages.
KEYWORDS: limited collocability; phraseological units; terms; monocollocable words; corpus linguistics.

## 1. A MODO DE INTRODUCCIÓN[1]

La colocabilidad de las palabras es un fenómeno que se estudia mucho en la lingüística actual. Los métodos basados en los corpus lingüísticos nos suministran instrumentos capaces de identificar coaparaciones frecuentes de palabras en una cantidad de datos enorme, lo que permite contrastar el tradicional enfoque introspectivo del investigador con aspectos cuantitativos. Este estudio pretende poner a prueba un método automatizado capaz de identificar palabras con una colocabilidad extremadamente restringida en español y en italiano, esto es, palabras cuyo potencial para combinarse se ve limitado enormemente, a veces incluso de una manera anómala. Además, ofrece un listado de palabras de este tipo en ambos idiomas. Nuestro objetivo primario no es en ningún caso analizar teóricamente las propiedades de estas palabras o de los conjuntos de palabras formadas por ellas, ni entrar en disquisiciones terminológicas: nos limitamos a proponer un procedimiento para identificarlas de manera automatizada, sin intervención de la introspección del investigador. Creemos que un método así les puede ser de gran utilidad a los especialistas, ya que pone a su disposición una cantidad de material lingüístico tan enorme que sería inanalizable por otros métodos.

## 2. COLOCABILIDAD RESTRINGIDA

Palabras como lirondo en español (que forma parte de la construcción mondo y lirondo) o squarciagola en italiano (en la construcción cantare/gridare a squarciagola) aparecen en los diccionarios con el comentario "se usa solo en la loc."/"usato solo nella loc." que remite precisamente a las construcciones mencionadas. Lo que hace el comentario es sencillamente explicitar la colocabilidad extremadamente restringida de estas palabras.

En el presente artículo nos interesa la colocabilidad concebida como una capacidad de las palabras de unirse en el texto con otras palabras[2]. La naturaleza de las palabras con las que se unen y el modo como lo hacen suministran mucha información sobre la palabra estudiada. Esta información resulta también relevante a la hora de definir su significado léxico. La combinabilidad constituye una propiedad básica y una función natural de todos los lexemas. Se supone que se presenta como una escala: un polo abarca palabras con una colocabilidad muy amplia, con pocas restricciones (p.ej.: el adjetivo *grande*), mientras que el otro lo forman palabras con un paradigma colocacional tan restringido que podrían

1 El trabajo fue financiado por el Fondo Europeo de Desarrollo Regional, Proyecto «Creatividad y adaptabilidad como condiciones del éxito de Europa en un mundo interrelacionado» (No. CZ.02.1.01/0.0/0.0/16_019/0000734), y por el proyecto *Progres Q10: La variabilidad del lenguaje a lo largo del tiempo, el espacio y la cultura* (Universidad Carolina).

2 Para un análisis teórico del concepto, véase Wotjak (2012).

denominarse monocolocables (p. ej.: la palabra *lirondo*). Aunque muchas veces se pone en contraste el llamado principio de libre selección (*open-choice principle*) y el principio idiomático (*idiom principle*)[3], queda patente que una colocabilidad ilimitada de unidades léxicas no existe (la colocabilidad tiene una dimensión semántica que hace imposible que aparezcan usos como **aire grande*; es decir, siempre hay restricciones)[4].

Es probable que las palabras con un potencial combinatorio muy reducido existan en todas las lenguas, aunque hay que suponer que cada lengua se comporta de una manera particular (para más detalle, véase Čermák, Čermák, Obstová y Vachková 2016: 7). En inglés suelen denominarse *cranberry words* (en analogía con *cranberry morphemes*, cfr. Bloomfield 1933 y Aronoff 1976), *bound words* o *unique words*; en alemán se emplean los términos *unikale Komponente* o *Unikalia* (cfr. Dobrovoľskij y Piirainen 1994); en italiano se habla de *componenti (vocaboli) a collocazione unica* (cfr. Veland 2005 y 2006) o, recientemente y siguiendo la terminología checa, de *parole monocollocabili* (cfr. Konecny 2018); en español encontramos los términos *palabras diacríticas* (Zuluaga Ospina 1980[5]; Ruiz Gurillo 1997, 1998 y 2001[6]), *palabras idiomáticas* (García-Page Sánchez 1990)[7], *elementos únicos* (inspirado por el término internacional; p.ej.: Mellado Blanco 1998) o el término muy original *hápax fraseológico* (González Rey 2005)[8] (destaquemos que los términos que hemos enumerado no son sinónimos del todo: los diferentes autores los enmarcan en contextos diferentes y los definen sirviéndose de criterios diferentes). En checo se ha instalado el término *monokolokabilní slova* (palabras monocolocables, PM; cfr. Čermák 1982 y 2014). František Čermák define la monocolocabilidad como una anomalía colocacional de una palabra, cuyo potencial colocacional es extremadamente restringido (entre 1 y 7 colocaciones documentadas). Este intervalo ha sido fijado de una manera arbitraria, pero el análisis de datos concretos nos hace ver que en las palabras con colocabilidad restringida es más razonable suponer la existencia de unas cuantas colocaciones, aunque efectivamente existan palabras –más bien excepcionales– con una sola colocación (el prefijo *mono-* resulta por ello un tanto exagerado).

Hasta ahora, las palabras monocolocables han sido analizadas ante todo en la fraseología. Como parte inalienable e inseparable de la unidad fraseológica las interpretan Filipec y Čermák (Filipec y Čermák 1985: 173). Por su parte, Dobrovoľskij y Piirainen (Dobrovoľskij y

3 Cfr. Sinclair (1991: 109-115).

4 Véase la introducción de Ignacio Bosque al diccionario REDES (Bosque 2004: LXXXIII-LXXXIV) para una acertada y detallada explicación de la relación entre la *combinatoria restringida* y la *combinatoria libre*. Bosque dice al respecto: "parece que la diferencia entre ambos dominios ha de estar en el tipo de restricciones que se consideran relevantes en cada uno, y no en la presencia o ausencia de restricciones" (Bosque 2004: LXXXIV). El diccionario mismo (Bosque *et alii* 2004) contiene pruebas abundantes de la validez de su afirmación.

5 "[…] destacamos la existencia de palabras únicas, carentes de toda autonomía semántica, reconocidas por el hablante solamente dentro de expresiones fijas (*lirondo, contera, vilo*); las llamamos palabras diacríticas, pues su función es la de constituir y distinguir signos" (Zuluaga Ospina 1980: 102-103);

6 Las palabras diacríticas son "elementos únicos, de uso exclusivo para las unidades que pertenecen a la fraseología" (Ruiz Gurillo 2001: 18).

7 El término *palabra idiomática* "permite dar cuenta de su práctica dependencia del contexto lingüístico en que se circunscriben" (García-Page Sánchez 1990: 279).

8 La autora habla de "el carácter excepcional de esas palabras que se registran sólo una vez en contextos exclusivos, en el sentido de que dichas palabras no son funcionales fuera de ese empleo al que se ven limitadas" (González Rey 2005: 319-320); el hápax fraseológico es "aquella palabra que aparece en una forma única en un entorno lingüístico fijo que la convierte en elemento exclusivo, sin otra posibilidad de existencia fuera de éste dentro de la lengua que utiliza el hablante; […] es aquella forma que funciona como elemento único simultáneamente en el conjunto de la lengua funcional, y en el sistema fraseológico de esa misma lengua" (González Rey 2005: 324).

Piirainen 1994: 449) las conciben como *phraseologisch gebundene Formative.* También en la lingüistica española suelen tratarse solo en la fraseología[9]. Solo trabajos más recientes, basados en el análisis de los datos del corpus, analizan también su presencia en unidades no fraseológicas (cfr. Konecny 2010: 305), ante todo en las denominaciones pluriverbales de carácter terminológico (cfr. Čermák 2014: 13; Čermák, Čermák, Obstová y Vachková 2016: 7, 13).

František Čermák (Čermák 2004) subraya el hecho de que a menudo se trata de formas concretas del paradigma (y no del paradigma completo) las que tienen una colocabilidad restringida. En estos casos, la naturaleza específica de estas palabras, su alto grado de defectividad puede traducirse en el hecho de que –desde el punto de vista sincrónico– realmente exista solo una forma suya: por ejemplo, en checo, *jít k duhu* "traer provecho" (*duhu* es dativo singular; el resto del paradigma ya no existe; la forma *duhu* carece de toda autonomía semántica y sintáctica). Lógicamente, la diferencias en la colocabilidad de los paradigmas completos y de las formas específicas son patentes sobre todo en las lenguas con una flexión nominal rica, como es el caso del checo. En las lenguas que no tienen flexión nominal o la tienen muy reducida no se dan tales diferencias, o bien son solo observables en la diferente colocabilidad de las formas de singular y de plural o en la inexistencia de una de las dos formas ([*under*/*through*/*outside the*] *auspices* [*of sth*]; [*campo di*] *concentramento*; [*en*] *ayunas*).

## 3. LA RECOGIDA DE DATOS Y LOS MÉTODOS DE IDENTIFICACIÓN DE LAS PALABRAS MONOCOLOCABLES

La reflexión teórica sobre la naturaleza de las palabras monocolocables (cualquiera que sea el término que se utilice) suele llevar a la configuración de su tipología (en los estudios dedicados al español son casi siempre tipologías del uso fraseológico de estas formas) o a la creación de listas de estas palabras en diferentes idiomas. Resulta evidente que tanto las listas como las tipologías requieren datos lingüísticos. Hasta ahora, los investigadores se han servido sobre todo del análisis de los textos, de las encuestas realizadas entre hablantes nativos, de la introspección y –en la mayoría de los casos– de los datos extraídos de los diccionarios que –como ya se ha señalado más arriba– suelen reflejar la colocabilidad restringida de la palabra con el comentario "solo en loc." o similar (para más detalles, véanse Dobrovol'skij 1988; Dobrovol'skij y Piirainen 1994; CoDII; Veland 2005 y 2006). En español, casi todos los trabajos que se dedican, al menos parcialmente, a las unidades fraseólogicas con palabras monocolocables contienen listas de ejemplos, acompañadas por un análisis de sus propiedades.

Si tomamos en cuenta el hecho de que la colocabilidad de las palabras cambia con el tiempo (muchas PM han pasado por un proceso durante el cual han ido perdiendo paulatinamente la capacidad de combinarse libremente)[10], cabe suponer que en la lengua aparezcan sin cesar PM "nuevas", no registradas todavía en los diccionarios. Son los corpus lingüísticos los que

9 La colocabilidad restringida constituye solo uno de sus rasgos definitorios. Hay otros: carencia de toda autonomía semántica; uso exclusivo para las unidades que pertenecen a la fraseología (Ruiz Gurillo); ausencia de significado literal o no fraseológico/idiomático (Larreta Zulategui); grado máximo de idiomaticidad o perfecta fosilización de estados arcaicos de la misma lengua histórica o de otras lenguas históricas, etc. (Zuluaga Ospina); dependencia del contexto lingüístico en que se circunscriben; capacidad para determinar el carácter fraseológico del enunciado (García-Page); total restricción sintagmática que impide su uso en el discurso libre (Larreta Zulategui); etc.

10 Este aspecto diacrónico del problema se comenta en Obstová (2018).

nos ofrecen una posibilidad única de cómo identificar las PM en la lengua de hoy, ya que nos permiten buscarlas de una manera sistemática en una cantidad de datos enorme. Precisamente del análisis de datos extraídos del corpus parte el proyecto de la identificación sistémica de las PM en el material de cuatro idiomas (inglés, alemán, italiano, checo) presentado en Čermák, Čermák, Obstová y Vachková (2016). El proyecto constituye el primer intento de utilizar un análisis automático para identificar las PM en diferentes lenguas[11]. El presente artículo constituye otro aporte a esta línea de investigación (cfr. Obstová 2017, 2018 y 2019).

### 3.1. Descripción del método utilizado

Se desprende de todo lo dicho hasta ahora que lo que buscamos es un instrumento que nos permita identificar palabras que cumplan con dos requisitos:

- su colocabilidad es restringida, es decir, tienen un número extremadamente reducido de colocaciones[12];
- su frecuencia es relativamente alta (las palabras que aparecen solo pocas veces en el corpus no nos sirven porque no permiten valorar su colocabilidad; el hecho de que un fenómeno aparezca poco en un corpus –por ejemplo, solo una vez– no significa automáticamente que se trate de un fenómeno marginal, un "hápax").

El método que utilizamos en este estudio se sirve del índice Herfindahl-Hirschman (HHI), que se utiliza para cuantificar la diversidad del contexto (en nuestro caso, para identificar unidades con una colocabilidad anómala; cfr.: Cvrček 2013: 119)[13]. Simplificando un poco, podemos decir que el índice es capaz de encontrar –en una cantidad de datos enorme– palabras que tienen una frecuencia mínima establecida por el investigador y que al mismo tiempo aparecen combinadas con un número muy reducido de palabras (colocados)[14]. El índice HHI puede adquirir valores entre 0 y 1: cuanto más alto es el valor del índice, tanto más restringida es la colocabilidad de la palabra. Se desprende de eso que el índice de las palabras monocolocables se acerca al valor numérico 1. Una gran parte de las palabras adquiere valores muy bajos (por ejemplo el valor medio de nuestro corpus AHMa es 0,1789): son palabras que tienen una colocabilidad libre o muy poco restringida.

---

11 Los autores identifican las PM en cuatro corpus de cca 100 millones de palabras (BNC, CORIS, SYN 2010 y un corpus alemán, creado *ad hoc* con los textos de Wikipedia).

12 No queremos entrar aquí en la discusión sobre la definición del controvertido concepto de *colocación*. Dado el método que estamos utilizando, nuestra interpretación de la colocación es estadística y se basa más en el aspecto de la frecuencia (*frequency-based*, Firth 1957; Halliday 1966; Sinclair 1991 y otros) que en el fraseológico (Cowie 1998; Hausmann 1984; Koike 2001; Corpas Pastor 2001, y otros). En el presente artículo entendemos por colocados las palabras con las que se unen las palabras monocolocables detectadas.

13 El índice HHI es un instrumento estadístico, un parámetro empleado en economía para medir la concentración económica de un mercado (ante todo para identificar monopolios en el mercado). Viene dado por la suma de los cuadrados de las cuotas de mercado de todas las empresas presentes en el sector. El índice puede utilizarse también en la lingüística (cfr. Cvrček 2013) para cuantificar la diversidad del contexto de una palabra, esto es, para ver si una determinada posición puede estar ocupada por varias palabras diversas o si más bien una (o unas pocas palabras) tienen el "monopolio" de esta posición.

14 Una frecuencia relativamente alta de la combinación es necesaria para poder identificar la PM: en palabras poco frecuentes en el corpus, la monocolocabilidad real puede confundirse con una monocolocabilidad aparente (una palabra poco frecuente puede aparecer en el corpus solo en una combinación, por ejemplo, aunque en realidad su combinatoria es libre).

Nos hemos servido de un programa informático que ha atribuido el índice HHI a todas las palabras semánticas de varios corpus lingüísticos españoles e italianos (su descripción aparece en los apartados **4**. y **5.**), aplicando los criterios que siguen:

1) Hemos trabajado con formas de palabras (*word form*), no con lemas (*lemma*), que incluyen el paradigma entero de una palabra (véase el comentario al respecto en el apartado **2.**). Los estudios que ya se han dedicado al tema hacen ver que el análisis de los lemas corre el riesgo de dejar de un lado muchos casos evidentes de monocolocabilidad (este peligro es más patente en las lenguas flexivas, pero en español y en italiano tiene también relevancia, sobre todo en la cuestión del número de los sustantivos y adjetivos).

b) Siguiendo el estudio de Obstová, hemos analizado únicamente palabras cuyo índice es superior a 0,7 (cfr. Obstová 2017: 229 para la detallada argumentación de esta decisión). Podemos decir que las palabras con un índice inferior a 0,7 son palabras con varios colocados dominantes u otras palabras cuya monocolocabilidad es discutible; dicho de otra manera, cuanto más bajo es el valor del índice, más discutible es la monocolocabilidad de la palabra.

c) El programa ha buscado solo palabras cuyo número de ocurrencias ha sido superior a 150 (véase la nota 14). Lógicamente, este criterio elimina varias PM menos utilizadas (véase el apartado **6.1.**). Se desprende de eso que en este estudio no analizamos el problema de la relación entre la frecuencia y la idiomaticidad que suele comentarse mucho en los estudios fraseológicos y que pone en duda Ignacio Bosque (Bosque 2004)[15].

### 3.2. Criba manual de los datos

Los datos extraídos del corpus se han sometido a una criba manual en dos fases.

En una primera fase han sido eliminadas las palabras, cuya monocolocabilidad viene dada por razones gramaticales o léxico-gramaticales. En concreto, han sido eliminadas:

- formas gramaticales compuestas;
- construcciones con valencia/régimen de las palabras (*propenso a*)
- nombres propios compuestos (*Gran Bretaña*);
- palabras como *besete* o *besazo / bacione* (que aparecen casi solo con *un*, por ejemplo, al final de una carta);
- palabras que expresan cantidad, como *poquitín* o *pochetto*, que aparecen de manera casi exclusiva acompañados por el artículo indefinido y la preposición *de / di*.

Todos estos "ruidos sistémicos" son fáciles de eliminar, y además, es imaginable la creación de un instrumento automatizado que los elimine sin intervención del investigador.

No obstante, en nuestros datos ha aparecido otro tipo de "ruido", más difícil de eliminar. En una segunda fase de criba hubo que eliminar palabras cuyo índice HHI apuntaba hacia la monocolocabilidad, aunque la apreciación personal del investigador decía que tenían una colocabilidad mucho más libre. Por la configuración del corpus, el programa ha atribuido un índice muy alto a formas como *alcohólicas* (prevalece *bebidas*), *espirituosas* (domina *bebidas*), *intrafamiliar* (predomina *violencia*), *olímpicos* (prevalece *juegos*) y otras. En

15 Bosque llega a la conclusión de que la frecuencia de coaparición no es consecuencia de la idiomaticidad, sino más bien de la sistematicidad (Bosque 2004: LXXXIV). Como veremos, nuestros datos parecen corroborar esta afirmación.

todos estos casos, prevalece claramente una combinación: estadísticamente, una palabra tiene "monopolio" en nuestro corpus, pero ello no se corresponde con la situación real en el sistema de la lengua (véase el apartado **6.** para la discusión del problema). Nuestra decisión de eliminar estas palabras puede apoyarse también en el comentario teórico de este tipo de combinaciones que aporta Ignacio Bosque (Bosque 2004: CLIII-CLV)[16].

## 4. PALABRAS MONOCOLOCABLES EN ESPAÑOL Y EN ITALIANO

Tanto los datos españoles como los italianos han sido obtenidos de tres corpus lingüísticos, tal y como refleja la tabla 1:

| Corpus | Abreviatura | Número de tokens | Autores |
|---|---|---|---|
| Español | | | |
| *Araneum Hispanicum Maius* | AHMa | 1.200.000.617 | Benko (2015) |
| *Araneum Hispanicum Minus* | AHMi | 121.570.580 | Benko (2015) |
| *InterCorp*, versión 11 | IC | 136.240.549 | Čermák y Vavřín (2018) |
| Italiano | | | |
| Araneum Italicum Maius | AIMa | 1.200.000.174 | Benko (2015) |
| Araneum Italicum Minus | AIMi | 119.284.520 | Benko (2015) |
| CORIS | CORIS | 130.294.357 | Rossini Favretti |

Tabla 1. Corpus lingüísticos utilizados

Los cuatro corpus *Araneum* (AHMa, AHMi, AIMa, AIMi) constan de textos descargados de páginas web[17]. Los corpus AHMi y AIMi constituyen una versión reducida de los corpus AHMa y AIMa, respectivamente, y abarcan el 10 % de los textos de AHMa/AIMa. Los hemos incluido en el estudio para poder valorar la influencia del tamaño del corpus sobre los resultados adquiridos. El corpus *InterCorp* se compone de textos literarios, periodísticos, administrativos y jurídicos[18]. Finalmente el corpus italiano CORIS es un corpus equilibrado, representativo y de referencia que contiene los siguientes materiales: prensa, textos literarios, textos académicos, textos jurídicos y administrativos, ephemera[19].

Los datos de los corpus han sido analizados mediante un programa automatizado específico que atribuye un valor de HHI a cada forma de palabra (como ya se ha señalado más arriba, hemos trabajado con formas, no con lemas, y solo han entrado en el análisis aquellas formas con un número de ocurrencias superior a 150). Después, todas las formas con el HHI superior a 0,7 han sido analizadas y cribadas por el investigador (véanse los criterios de la criba en el apartado **3.2.**; además, hemos consultado los diccionarios REDES (Bosque et alii, 2004) y GRADIT (De Mauro 1999) para ver si estas importantes fuentes recogen la construcción en cuestión).

16 Entre otras cosas, Bosque comenta que "la frecuencia –aunque elevada– de algunas expresiones no nos dice nada acerca del idioma, sino que nos informa a lo sumo de algunos de nuestros hábitos" (Bosque 2004: CLIV).

17 http://ucts.uniba.sk/aranea_about/index.html

18 En concreto son textos literarios, periodísticos (provenientes de las páginas Project Syndicate y VoxEurop) y jurídicos (de la Unión Europea, Acquis Communautaire), actas del Parlamento Europeo de los años 2007–2011 (Europarl) y subtítulos de películas de la base de datos OpenSubtitles.

19 CORIS/CODIS, Università di Bologna, accesible en http://corpora.ficlit.unibo.it/coris_ita.html

### 4.1. Caracterización básica de los datos

La Tabla 2 recoge los datos estadísticos básicos de los resultados:

| Español | | Italiano | |
|---|---|---|---|
| **Corpus** | **Formas monocolocables** | **Corpus** | **Formas monocolocables** |
| **AHMa** | **193** | **AIMa** | **236** |
| **AHMi** | **63** | **AIMi** | **79** |
| **IC** | **80** | **CORIS** | **109** |
| **AHMa + AHMi[20]** | **63** | **AIMa + AIMi** | **72** |
| **AHMa + IC** | **51** | **AIMa + CORIS** | **68** |
| **AHMi + IC** | **41** | **AIMi + CORIS** | **50** |
| **AHMa + AHMi + IC** | **41** | **AIMa + AIMi + CORIS** | **48** |

Tabla 2. Formas monocolocables en los corpus españoles e italianos

Los datos de la Tabla 2 nos permiten realizar las observaciones siguientes:

1) El tamaño del corpus es relevante

- en español, en la versión menor (10 %, AHMi, AIMi) de otro corpus (AHMa y AIMa, respectivamente), el programa identifica todas las formas monocolocables identificadas en la versión mayor; para el italiano vale lo mismo, pero, además, el programa reconoce también 5 formas (*lizza, vertebrale, spasso, occorrenza, vizioso*) no identificadas en el corpus grande (AIMa) porque su índice es levemente inferior a 0,7 (entre 0,69-0,61);

- en la versión mayor del corpus (AHMa, AIMa) –que en total tiene diez veces más palabras que la menor (AHMi, AIMi)– el programa identifica 3,1 veces más (en español) y 3 veces más (en italiano) formas monocolocables (resulta que el aumento de la dimensión del corpus no es idéntico al aumento del número de formas monocolocables).

2) La configuración del corpus es relevante (como ya se ha dicho, los corpus AHMa, AHMi, AIMa y AIMi contienen textos descargados de las páginas web, mientras que el corpus IC textos literarios, administrativos y jurídicos, y el corpus CORIS, mayoritariamente, textos literarios y de prensa):

- el corpus IC es un 12 % más grande que el corpus AHMi y contiene un 29 % formas monocolocables más;

- solo un 64 % de formas monocolocables del corpus IC aparece también en el corpus AHMa; el 36 % restantes son diferentes (dicho de otra manera, el corpus IC, cuyo tamaño equivale solo a una décima parte del tamaño de AHMa, contiene un 36 % de formas monocolocables que no aparecen en AHMa);

- el corpus CORIS es en un 8 % más grande que el corpus AIMi y contiene unos 28 % formas monocolocables más;

20 Formas identificadas como monocolocables en los dos corpus.

- solo un 62 % de formas monocolocables del corpus CORIS aparece también en el corpus AIMa, unos 38 % son diferentes (dicho de otra manera, el corpus CORIS, cuyo tamaño equivale solo a una décima parte del tamaño de AIMa, contiene un 38 % de formas monocolocables que no aparecen el AIMa).

Nuestros datos pueden dar cuenta también de las formas "más monocolocables". Dos formas españolas (*cabelludo* en AHMi y *valorem* en IC) y una italiana (*malapena* en CORIS) adquieren el valor 1 del HHI, o sea, aparecen solo en una combinación en el corpus correspondiente. Valores extremos del índice se dan por ejemplo también en las palabras españolas *bruces*, *través*, *obstante* o *repente*, y en las italianas *frattempo*, *vanvera*, *vitro* o *disparte*.

## 5. TIPOLOGÍA DE LAS CONSTRUCCIONES CON FORMAS IDENTIFICADAS COMO MONOCOLOCABLES

En este apartado vamos a esbozar una tipología de las construcciones que incluyen formas identificadas como monocolocables mediante nuestro método. Pretendemos ofrecer una visión de conjunto del tema, por lo que dejamos a un lado las diferencias entre los corpus comentadas en el apartado anterior y analizamos conjuntamente las 219 formas monocolocables españolas y las 277 italianas que han sido identificadas como tales al menos en uno de los tres corpus.

Nuestros datos hacen ver que en las dos lenguas las PM aparecen en diferentes tipos de construcciones:

| **Español** | | | **Italiano** | |
|---|---|---|---|---|
| Tipo | Número de PM | % | Número de PM | % |
| Locuciones nominales / términos | 71 | 32,42 % | 126 | 45,49 % |
| Latinismos | 27 | 12,33 % | 29 | 10,47 % |
| Anglicismos, galicismos | 0 | 0 % | 13 | 4,69 % |
| Locuciones conjuntivas y preposicionales | 23 | 10,5 % | 13 | 4,69 % |
| Locuciones adverbiales y adjetivales | 65 | 29,68 % | 71 | 25,63 % |
| Locuciones verbales | 23 | 10,5 % | 23 | 8,3 % |
| Casos especiales | 10 | 4,57 % | 2 | 0,73 % |
| **En total** | **219** | **100 %** | **277** | **100 %** |

Tabla 3. Tipología de las construcciones con formas identificadas como monocolocables

En general, son las locuciones nominales/los términos los que prevalecen claramente (su predominio es aún más claro en italiano).

### 5.1. Locuciones nominales (términos)

*5.1.1. Español*

*meteduras, quebradero, cabelludo, tomadura, moscada, meteduras, leporino, quebraderos, amiotrófica, refundido, ferina, lesa, dictaminadoras, sulfurizado, antipersona, espaciadora, aerostático, aerostáticos, úrico, expiatorios, cocleares, velatoria, fólico, inoxidable, pantoténico, conejillos, potable, bariátrica, ferropénica, adquisitivo, pistoletazo, filosofal, harineros, falciformes, hialorúnico, tomadores, pernada, sudoríparas, lacrimógenos, dactilares, antipersonal, bórico, inmemoriales, subatómicas, calloso, acondicionado, conectivo, cautelares, extrasolares, supermasivo, conjuntivo, inflexión, huracanados, dictaminadora, carpiano, vertebral, maché, levadizo, sebáceas, tabáquico, marciales, terráqueo, coclear, umbilical, suspensivos, alfabético, laborables, domicialario, integrante, angular, ponedoras*

Prevalecen claramente la combinaciones *sustantivo + complemento adnominal* (adjetivo, preposición + sustantivo). Tanto el sustantivo como el adjetivo pueden ser monocolocables.

El grupo[21] abarca dos subtipos:

1) palabras que forman parte de la terminología científica especializada, con una predominancia de términos de química y medicina; según muestran nuestros datos, estos términos pueden aparecer también fuera del ámbito estrictamente especializado.

- *amiotrófica* (*esclerosis*), *úrico* (*ácido*), *fólico* (*ácido*), etc;

- puede que estas formas se combinen –en la disciplina científica en cuestión– con más palabras, pero en nuestros datos aparecen solo en una o dos combinaciones;

2) términos utilizados en la vida cotidiana, locuciones nominales no especializadas, etc. (somos conscientes de que la frontera entre los dos subgrupos es a veces borrosa: el criterio lo constituye el grado de especialización del término y su uso en el lenguaje cotidiano).

- *cocleares* (*implantes*), *lacrimógenos* (*gases*), *leporino* (*labio*), *dactilares* (*huellas*), *carpiano* (*túnel*), *metedura*/s (*de pata*), *quebradero*/*s* (*de cabeza*), *tomadura* (*de pelo*), *sulfurizado* (*papel*), *expiatorios* (*chivos*), *velatoria* (*sala*), *potable* (*agua*), *pistoletazo* (*de salida*), *tomadores* (*de decisiones*), *inmemoriales* (*tiempos*), *maché* (*papel*), *levadizo* (*puente*), *umbilical* (*cordón*), *ponedoras* (*gallinas*) etc.; algunas de estas palabras tienen un HHI muy alto e incluso en nuestros datos tienden a una monocolocabilidad total (*cabelludo*, *lesa*);

- tiene validez lo señalado más arriba: nuestros datos son el resultado de la aplicación del método a un corpus concreto; por ejemplo, la forma *inoxidable*, con un índice de 0,86382 y con el número de apariciones de más de 2500, se combina de manera casi exclusiva con el sustantivo *acero* en nuestro corpus (además, aparecen unas cuantas combinaciones con *material*, *metal* y *lata*).

Es precisamente en el segundo subtipo donde hubo de aplicarse la segunda fase de la criba descrita en el apartado **3.2.** El programa atribuyó un índice superior a 0,7 a palabras que, por un lado, forman parte de una combinación muy frecuente y dominante, pero, por otro, pueden imaginarse fácilmente en otras combinaciones (cfr. *alcohólicas*, *espirituosas*, *intrafamiliar*, *olímpicos*, *porcentuales*, etc).

21 Bosque (2004: CLIV-CLVI) ofrece una reflexión lexicográfica muy interesante sobre este grupo.

Es interesante observar el tratamiento de estas palabras en el diccionario combinatorio REDES. Partiendo de las observaciones teóricas de Ignacio Bosque, sus autores prefieren no introducir lemas relacionados con las formas que estamos comentando en este apartado. Son muy pocas las excepciones: en el diccionario aparecen *levadizo* (como colocados sistémicos se mencionan *barrera*, *puente*, *puerta*), o *leso*. Y precisamente la palabra *leso* puede dar cuenta del hecho de que no es lo mismo ser frecuente que ser sistémico y, por ende, de ciertas limitaciones que tiene el método que ponemos a prueba: el índice documenta una presencia muy frecuente de una combinación que predomina claramente (en comparación con otras combinaciones posibles) en un conjunto enorme de datos lingüísticos (*lesa humanidad*). No obstante, por más grande que sea el conjunto de datos, sigue siendo solo una muestra específica de la lengua y no la lengua en su totalidad, por lo que no refleja todas las potencialidades sistémicas. Los autores del lema *leso*, partiendo de sus fuentes, mencionan como los más frecuentes los colocados *humanidad*, *majestad* y *patria*, pero además, mencionan otros 22 colocados posibles aportando citas concretas (Bosque *et alii* 2004: 1258-1259).

*5.1.2. Italiano*

Predominan claramente la combinaciones *sustantivo + adjetivo.* El grupo abarca dos subtipos:

1) palabras que forman parte de la terminología científica especializada (84 casos), con una predominancia de términos de química y medicina:

*capelluto, folico, nucleici, fumarie, giudicatrice, ialuronico, botulinica, alcolometrico, urico, elettrogeni, pectoris, zooprofilattico, volumico, amiotrofica, giudicatrici, binarie, acquifere, zooprofilattici, calorifico, focaia, clorogenico, fumogena, carbonica, fotostatica, spaziatrice, cloridrico, esaminatrice, elettrogeno, antiparticolato, senzienti, autogeno, reumatoide, cauzionale, lattico, circondariale, antiuomo, accomandita, cordonale, cessante, cistica, fabbricabili, anidride, metallifere, aumentativa, cranico, salicilico, termoelettriche, ausiliatrice, miocardico, palliative, centigradi, fumaria, confirmatoria, glicolico, anatocistici, ondoso, pulsata, uninominali, imbrifero, varicose, alcolemico, lattiero-caseari, fluorurati, cinerarie, anafilattico, sudoripare, ombelicale, olografo, androgenetica, ricino, solforico, scamosciata, capitaneria, coloniche, arachidonico, carpale, termovettore, preadottivo, emendative, esattoriali, antidiscriminazioni, sciatico, vertebrale, aggiudicatrici.*

Teóricamente, en la disciplina en cuestión, muchas de estas formas son combinables también con otras palabras, pero en los textos no especializados que forman parte de los corpus solo se unen con una o dos palabras (dicho de otra manera, solo uno de los términos posibles es de interés general).

El diccionario GRADIT opta por la misma solución que el REDES en el caso del español y no incluye –por su carácter especializado– la mayoría de estos términos.

2) palabras que forman parte de combinaciones no especializadas (41); son términos y combinaciones de palabras utilizados en la vida cotidiana (*sistema immunitario*, *acqua piovana, vitello tonnato, filo spinato*), otras unidades pluriverbales de uso cotidiano (*datore di lavoro, casa editrice, stragrande maggioranza*) o unidades fraseologícas nominales (*capro espiatorio, fiore all'occhiello*):

*moscata, datori, stragrande, piovana, datore, fanalino, semolato, miliare, espiatori, filosofale, editrici, immunitario, vandalici, figliol, vedenti, morsicata, datrice, pindarici,*

*integrante, occhiello, capro, spiritica, cipollina, paliative, serramanico, quartier, marziali, extravergine, vedente, concentramento, spinato, tonnato, vizioso, rettorale, lasso, espiatorio, terrieri, levatoio, rettor, sbarrati.*

La mayoría de las palabras de este tipo viene registrada por el diccionario GRADIT como parte de unidades pluriverbales (en algunas de ellas el diccionario registra colocados que no aparecen en nuestros datos; p.ej.: *noce/achillea/uva moscata*), no obstante, su carácter monocolocable se menciona solo excepcionalmente.

### 5.2. Latinismos

*5.2.1. Español*

*cápita*, *témpore*, *posteriori*, *fraganti*, *aequo*, *statu*, *máter*, *priori*, *vitro*, *situ*, *hominem*, *honorem*, *operandi*, *honoris*, *facies*, *memoriam*, *nihilo*, *extremis*, *crucis*, *hoc*, *facto*, *vadis*, *mortem*, *valorem*, *mutatis*, *mutandis*, *minimis.*

Su presencia no es sorprendente, ya que son construcciones muy lexicalizadas, a veces incluso petrificadas. Aparecen más en el corpus IC que en los otros dos corpus, lo que tampoco sorprende, dado su componente importante de textos literarios y administrativos.

*5.2.2. Italiano*

*primis, extremis, pectore, hoc, generis, partum, pontificum, mutatis, scriptum, volenti, honoris, nuce, minimis, honorem, volente, dulcis, nolente, litteram, tantum, conditio, operandi, priori, alter, tempore, stabat, crucis, personam, itinere, vitro.*

Vale lo dicho sobre el español (**5.2.1**.). El porcentaje alto de latinismos en nuestros datos muestra además su importante papel en el texto escrito. A diferencia de los datos españoles, en italiano es observable una presencia frecuente de estas formas en el corpus AIMa.

### 5.3. Anglicismos y galicismos

*5.3.1. Italiano*

*stager, operator, mailing, task, made, mountain, part, talk* (anglicismos)
*époque, roulant, prodige, tout, tapis* (galicismos).

Este tipo de palabras solo es observable en italiano.

### 5.4. Locuciones conjuntivas o preposicionales

*5.4.1. Español*

*través, obstante, embargo, doquier, torno, anterioridad, posterioridad, partir, ende, pesar, aras, postrimerías, transcurso/trascurso, consonancia, pese, respecta, detrimento, albores, desmedro, tocante, consiguiente, pos, amparo.*

El grupo consta de palabras monocolocables que forman parte de construcciones con una función gramatical:

1) locuciones preposicionales; se trata de:

- palabras que casi no aparecen en otro contexto, cfr. *aras* (*en a. de*), *pos* (*en p. de*);
- palabras que –además de formar parte de una locución preposicional, función claramente predominante en ellas– pueden ser también sustantivos, aunque no muy frecuentes (cfr. *detrimento*, *amparo*[22]*, postrimerías*); el índice muy alto nos informa, no obstante, de que, al menos en nuestros corpus, el uso no preposicional es puramente marginal;
- palabras que parecen no ser monocolocables, ya que tienen, además, un uso nominal: *torno* (*en t. a.*), *trascurso* (*en el t. de*), *consonancia* (*en c. con*), *anterioridad* (*con a.*), *partir* (*a p.*); no obstante, el programa les ha atribuido un índice muy alto a estas palabras, de lo que se desprende que en el uso (en nuestros corpus) el predominio de la construcción preposicional es abrumador;

2) locuciones conjuntivas; se trata de:

- palabras que no suelen aparecer en otro contexto, cfr. *ende* (*por e.*), *obstante* (*no o.*);
- palabras que, aunque pueden ser también nombres, adjetivos o verbos, por su índice han sido identificadas como monocolocables (lo que demuestra otra vez el predominio total del uso conjuntivo), cfr. *embargo* (*sin e.*), *pese* (*p. a*), *tocante* (*t. a, en lo t. a*), *respecta* (*por lo que r.*, *en lo que r.*), etc.

Las palabras mencionadas forman parte de unas construcciones fijas –a veces muy cultas o literarias– compuestas de dos o tres miembros con un significado mayoritariamente gramatical (las palabras *pos* y *amparo* solo se han identificado como monocolocables en el corpus IC, lo que viene dado por su configuración, por la presencia de textos literarios).

El papel de la configuración del corpus es decisivo en las formas que pueden tener también función nominal: por ejemplo, la palabra *pesar* tiene el índice 0,907 en el corpus AHMa y se comporta como monocolocable, mientras que en el corpus IC, que abarca textos literarios, el HHI adquiere solo el valor de 0,41 (la presencia frecuente del uso nominal cambia por completo la situación).

### *5.4.2. Italiano*

*procinto, pressi, scapito, decorrere, istato, suon, discapito, disopra, egida, dispetto, confronti, stregua, prescindere.*

Son sustantivos o infinitivos que aunque puedan aparecer como tales, casi siempre forman parte de la locución preposicional (muchas veces son palabras anticuadas o formalmente anómalas, cfr. *istato* o la forma apocopada *suon*).

22 Volvemos a subrayar lo sustancial: lo que estamos analizando son los resultados de la aplicación de nuestro método a corpus concretos. La intuición del hablante nativo podría ponerlos en duda; por ejemplo, un hablante nativo podría opinar que *amparo* tiene una colocabilidad mucho mayor que *detrimento* o *postrimerías*: *buscar amparo, encontrar amparo, sentirse sin amparo,* etc. Además, se encuentra en la expresión jurídica *recurso de amparo* que aparece con gran frecuencia en la prensa.

## 5.5. Locuciones adverbiales y adjetivales

### *5.5.1. Español*

*bruces, repente, antemano, ultranza, bocajarro, rechupete, antonomasia, reojo, alimón, dedillo, marras, ayunas, agigantados, tutiplén, cabalidad, refilón, ristre, contrapelo, tapadillo, bordo, juntillas, vilo, sopetón, plumazo, regañadientes, jamases, volandas, creces, destiempo, pacotilla, albedrío, horcajadas, puridad, demasía, rajatabla, destajo, quemarropa, improviso, rechistar, tapujos, raudales, rededor, unísono, soslayo, ambages, menudo, deshora, desuso, postín, continuación, intemperie, borbotones, trompicones, añadidura, vano, chistar, hurtadillas, sucesivo, granel, rastras, rabillo, veras, cuclillas, supuesto, puntillas.*

Prevalecen las construcciones con significado adverbial (mayoritariamente temporal o modal), aunque a veces las dos interpretaciones –adverbial y adjetival– son posibles, por lo que los diccionarios suelen utilizar el comentario *loc. adv. / loc. adj.* (en los diccionarios la locución casi siempre es tratada como un todo; el diccionario REDES a veces comenta la combinatoria de la locución entera, es decir, no analiza su estructura interna; cfr. la locución *a ultranza*).

Algunas de las locuciones mencionadas pueden formar parte de una locución verbal (por ejemplo, el DRAE considera la locución *de pacotilla* como *loc. verb.*, o sea, *ser de pacotilla*; no obstante, en nuestro material prevalece el uso adjetival, por eso la incluimos en este subgrupo). A veces, el límite entre los tipos de locución queda poco claro, cfr. *el rabillo de ojo*, considerado por DRAE como locución verbal (*mirar con el rabillo del ojo*), mientras que nuestro material muestra también una presencia marginal de otros verbos (*vigilar*, *atisbar*…). De todos modos, como ya se ha señalado más arriba, no es nuestro objetivo comentar la naturaleza de estas locuciones: solo queremos hacer ver que nuestro programa ha identificado un componente suyo como palabra monocolocable.

### *5.5.2. Italiano*

Estas locuciones son muy frecuentes en nuestros datos. Igual que en español la diferenciación entre los diferentes tipos de locuciones puede ser difícil. Eso vale para la diferenciación entre las locuciones adverbiales y verbales (*a squarciagola x gridare/cantare a squarciagola*; en tales casos el diccionario GRADIT menciona muchas veces dos posibilidades), o para las adverbiales y adjetivales (*essere in disuso, fabbriche cadute in disuso, fabbriche in disuso*; también en este caso GRADIT ofrece más interpretaciones).

*sbando, spicco, zeppo, lattiero, strapazzo, capogiro, zeppi, vegeto, auge, ingrosso, diametralmente, diporto, tracolla, stiro, vigore, voga, disuso;*
*frattempo, vanvera, disparte, soppiatto, ridosso, malapena, capofitto, particolar, altronde, malincuore, ritroso, tantino, infuori, antonomasia, sbaraglio, casaccio, incirca, bruciapelo, rinfusa, insù, bizzeffe, impazzata, dismisura, sgoccioli, bilico, rado, battibaleno, sommato, braccetto, subordine, sbafo, unisono, ingiù, bagnomaria, catinelle, sordina, tentoni, precedenza, spasso, occorrenza, squarciagola, rallentatore, sottecchi, baleno, cavalcioni, stento, pian, sbieco, erta, perdifiato, galla, soqquadro, tilt, palio.*

En la mayoría de los casos, un sustantivo monocolocable (raras veces otra clase de palabras)[23] es el término de una preposición (*nel frattempo, a casaccio*). También en este grupo, el diccionario GRADIT vacila a la hora de identificar la clase de locución: a veces la interpreta como adverbial, otras veces como verbal, y a menudo admite las dos soluciones.

### 5.6. Locuciones verbales

#### *5.6.1. Español*

*garete*, *omiso*, *bledo*, *pairo*, *entredicho*, *riendas*, *escarpias*, *garbeo*, *tabarra*, *vistazo*, *palestra*, *vueltecita*, *hincapié, andadas*, *rienda*, *tintero*, *calzador*, *vueltita, respingo*, *solfa*, *vigor*, *paces*, *trizas.*

Las palabras identificadas como monocolocables forman parte de una locución verbal, teniendo mayoritariamente la función sintáctica de objeto o complemento circunstancial del verbo. En los diccionarios (por ejemplo en el DRAE) figuran como un todo con el comentario *loc. verb.*, a veces con una referencia al registro (p.ej.: *coloquial*).

Se trata de casos de lo que se conoce como solidaridad léxica. Formalmente prevalece la combinación verbo + (artículo) + forma monocolocable; a veces, la estructura interna de la locución es más compleja (cfr. *hacer caso omiso*, *dar rienda suelta*)[24].

El programa ha considerado como monocolocables algunos derivados diminutivos de palabras que no son monocolocables, cfr. *dar una vueltecita*, *dar una vueltita* (la forma *vuelta* no ha sido identificada como monocolocable).

#### *5.6.2. Italiano*

*repentaglio, sopravvento, stremo, rendersene, convolare, rimbocca, rincarare, rendercene, storcere, spallucce, capolino, mozzare, lunario, affermativamente, spola, briga, furie, scrollarsi, cagnesco, zonzo, crepapelle, visibilio, lizza.*

En la mayoría de los casos, la forma monocolocable es un sustantivo, pero aparecen también verbos monocolocables (*mozzare*), ejemplos de la solidaridad léxica. Predominan las construcciones que suelen considerarse como unidades fraseológicas.

### 5.7. Casos especiales

#### *5.7.1. Español*

Algunas palabras identificadas no cumplen con los criterios definitorios de los grupos que hemos comentado hasta ahora.

Se trata de los siguientes casos:

---

23 Es bastante frecuente que las formas monocolocables sean difíciles de asignar de una manera inequívoca a una clase de palabras (por ejemplo, las palabras *malapena* o *malincuore* se consideran como sustantivos en unos diccionarios y como adverbios en otros).

24 Acerca del problema de composicionalidad puede consultarse Bosque, 2004, CXXVI.

- un miembro de un paradigma verbal concreto se une con una forma de un sustantivo concreto (los demás miembros de los paradigmas no han sido identificados como monocolocables), cfr. *hincarle* (*el diente*), *chuparse* (*los dedos*), *habida* (*cuenta*), *encogiéndose* (*de hombros*), *surtirá* (*efecto*, *efectos*);

- el programa ha asignado un índice alto a las dos partes de una combinación, cfr. *frunció el ceño*/*entrecejo*, *fruncí el ceño*/*entrecejo*, o sea, se consideran como monocolocables las formas *ceño* (*frunció, fruncí el c*.), *frunció* (*el ceño*, *entrecejo*), *fruncí* (*el ceño*, *entrecejo*), *entrecejo* (*frunció, fruncí*); resulta que se trata de unas locuciones verbales compuestas de dos palabras monocolocables;

- la palabra *maestrillo*, que forma parte de la locución *cada maestrillo tiene su librillo*.

*5.7.2. Italiano*

*intenderci, diciamocela*

Dos palabras monocolocables forman parte de una unidad fraseológica compleja, diferente de las construcciones que hemos comentado en los apartados anteriores (*per intenderci*, *diciamocela tutta*).

## 6. DISCUSIÓN

### 6.1. Eficacia del método

Si consideramos los resultados obtenidos en los dos idiomas, podemos hacer constar las siguientes ventajas y desventajas del método automatizado basado en el uso del índice HHI:

1) El método identifica muy bien latinismos, anglicismos y galicismos, esto es, formas monocolocables que por definición tienden a formar parte de construcciones casi petrificadas en español y en italiano.

2) En las dos lenguas, es muy eficaz en la identificación de las locuciones adverbiales, adjetivales y verbales.

3) Identifica sin problemas las locuciones de carácter gramatical, es decir, las conjuntivas y las preposicionales.

4) Es capaz de identificar muy bien los términos científicos muy especializados.

5) En cuanto a los términos científicos más comunes, la eficacia del método es menor. Por una parte, identifica formas monocolocables cuya monocolocablidad restringida a veces pasa desapercebida (*potable*). Por otra, asigna un índice HHI muy alto a formas que cuentan con un colocado muy frecuente en nuestros datos, pero para las que la intuición del hablante puede imaginarse una colocabilidad potencial más extensa (tipo *olímpicos*, véase el apartado **3.2.**). Si no nos interesa la monocolocabilidad de la palabra solo en un corpus concreto, si pretendemos buscar la monocolocabilidad sistémica, hay que dejar de lado las formas de este tipo. El problema es que el proceso de eliminación debe ser manual, por lo que es laborioso y al mismo tiempo subjetivo.

6) Hay un punto débil del método que hace problemático su uso en los estudios puramente fraseológicos: pasan desapercibidas muchas palabras monocolocables que forman parte de las construcciones que suelen ser objeto de estudio de la fraseología, pero que tienen un

número de ocurrencias muy bajo en los textos (a veces por pertenecer a un registro marcado): en nuestras listas no aparece por ejemplo la palabra *lirondo* (*mondo y lirondo*), comentada en el apartado **2**. No hay manera de evitar este inconveniente porque si bajáramos el límite de frecuencia, el método perdería sentido. Además, nuestros datos nos hacen ver que el tipo de corpus tiene relevancia en el número de ocurrencias que debe tener una forma para ser identificada como monocolocable: las palabras *granel*, *rabillo*, *cuclillas*, *puntillas* solo se identifican como monocolocables en el corpus IC.

7) Por definición, el método identifica solo formas con un colocado dominante: si hay más colocados (aunque no sobrepasen el límite de 7, fijado en nuestro estudio), el índice baja rápidamente. Resulta que por ejemplo una palabra con cuatro colocados (pero no más) puede tener un índice tan bajo que no entra en el intervalo predeterminado.

8) De todo lo anterior se desprende que el método a veces no puede funcionar sin una intervención manual/intuitiva del investigador, lo que pone en duda su ventaja más importante, el carácter automático.

### 6.2. Criba de datos

La utilización de un método automatizado, basado en un análisis estadístico y en su aplicación a grandes cantidades de datos, lleva consigo un ahorro de tiempo enorme. Por desgracia, este ahorro se ve parcialmente reducido por la necesidad de una criba manual (descrita en el apartado **3.2.**). En un futuro perfeccionamiento del método es imaginable una reducción de esta criba (por ejemplo, sería posible definir una parte de las construcciones no deseadas de manera que el programa las elimine sin intervención manual del investigador), pero parece probable que una parte de la criba (por ejemplo la que elimina las formas como *olímpicos*) seguirá siendo necesaria, lo que hace el método más laborioso y más subjetivo.

Otro problema de la criba lo constituyen las formas como *embargo*, *torno* o *menudo*, que tienen –además del significado de la forma monocolocable que forma parte de construcciones *sin embargo, en torno a* y *a menudo*– otro uso sustantivo o adjetivo. Como podemos observar, el método las ha identificado como monocolocables, ya que efectivamente lo son según los principios estadísticos en los que se basa el programa (en el corpus utilizado, el predominio de la forma monocolocable es suficiente para que se le asigne un índice alto). Desde el punto de vista diacrónico el problema es probablemente mucho más complejo: hay pruebas de que a veces las palabras se "monocolocabilizan" en la lengua cotidiana (por ejemplo, la palabra italiana *galla*, que forma parte de la construcción muy frecuente *a galla* „en la superficie" –véase **5.5.2.**– tiene el significado original „agalla, cecidia"; se utilizaba para producir tinta y las agallas –por pesar poco– quedaban a la superficie; hoy, la mayoría de los hablantes –ajena al lenguaje especializado de los biólogos– solo conoce el uso monocolocable y no tiene conciencia de su origen; cfr. Obstová 2018: 75-76). De todos modos, el aspecto diacrónico queda fuera del presente estudio.

### 6.3. Influencia del corpus

Como el número de ocurrencias es uno de los parámetros constitutivos del método es trivial la conclusión de que el tamaño del corpus influye sobre los resultados: el corpus AHMa es mucho más grande que los AHMi y IC, por eso contiene más palabras monocolocables (la

situación en los corpus italianos es análoga). No obstante, el aumento del número de PM no se corresponde exactamente con la diferencia de tamaño de los corpus, como se ha mostrado en **4.1.** (un análisis de estas correspondencias constituye un desiderátum de la futura investigación en este campo).

Se ha probado la importancia de la configuración del corpus: en el IC (que consta de textos de otra índole que los demás corpus) se han encontrado formas monocolocables que no han sido identificadas en los corpus AHMa y AHMi. La misma situación se ha observado en italiano de una forma muy marcada: por ejemplo, los corpus AIMa y AIMi, formados por textos descargados de la red, contienen muchos términos (locuciones nominales).

### 6.4. *Forma* vs. *lema*

Aunque ni el español ni el italiano no tienen declinación nominal, parece razonable trabajar con formas y no con lemas.

Los datos nos hacen ver que a veces se consideran como monocolocables tanto la forma del singular como la del plural (*meteduras* y *metedura de pata*, *quebradero* y *quebraderos de cabeza*), otras veces solo una de las formas (*ayunas*). Lo mismo vale para algunas formas verbales (*surtirá efecto*).

### 6.5. El español y el italiano

Como se desprende de los apartados anteriores, el español y el italiano se comportan de una manera muy semejante. Las tipologías de las formas monocolocables son muy parecidas (a diferencia de lo que ocurre en español, en italiano se han identificado también unos cuantos anglicismos y galicismos monocolocables, cosa que no se ha producido en español, pero habría que tener corpus más grandes y de diferentes configuraciones para extraer conclusiones definitivas al respecto). Es precisamente la configuración diferente del corpus lo que nos hace ver la necesidad de poner a prueba nuestros datos en otros corpus: todo parece indicar que en los corpus utilizados es CORIS el corpus más equilibrado, lo que se refleja en unos resultados un poco diferentes de los de los otros corpus.

## 7. CONCLUSIONES

En este trabajo hemos presentado de una manera sucinta el estado actual del estudio del fenómeno de la colocabilidad extremadamente restringida y lo hemos ejemplificado en el español y el italiano. Hemos mostrado que aunque el fenómeno suele tratarse sobre todo en fraseología o (en su dimensión práctica) en lexicografía, existen muchas palabras con una colocabilidad muy restringida que tienen que ver más bien con la terminología. Nuestro objetivo principal ha sido poner a prueba un método automatizado que trabaja con el índice HHI y pretende –sirviéndose de grandes cantidades de datos adquiridos de los corpus lingüísticos– identificar palabras con una colocabilidad restringida sin que sea necesaria una intervención decisiva del investigador y de su intuición. Los resultados nos hacen ver que el método puede utilizarse como un instrumento auxiliar, complementario, que le permite al investigador adquirir una colección de datos potencialmente interesante para su posterior análisis, una colección que refleja el uso lingüístico en una cantidad enorme de

datos lingüísticos (son dos ventajas evidentes del método: no solo ahorra tiempo, sino que también permite trabajar con cantidades de datos inimaginables hace unos decenios). Este análisis posterior es necesario porque, aunque el método logra identificar adecuadamente las palabras monocolocables que suelen tratarse en la fraseología (como parte de las locuciones adverbiales, adjetivales o verbales) y también los latinismos, en los términos (las locuciones nominales) afronta serios problemas. Puede decirse que el proceso de identificación logra encontrar palabras que pueden considerarse como monocolocables, pero –por desgracia– al mismo tiempo encuentra bastantes palabras que no resisten la confrontación con la intuición lingüística del investigador. Dicho de otra manera, el método requiere un análisis posterior basado en la introspección, lo que contamina el proceso automático con un trabajo manual y con una carga de subjetividad. Otro problema consiste en que a veces hay diferencia entre la colocabilidad, tal y como la presenta un corpus, por grande que sea, y la colocabilidad sistémica, tal y como la concibe un hablante nativo. En el apartado **5.1.1.** hemos visto que en la forma *lesa*, que en nuestro corpus tiende a una monocolocabilidad casi absoluta, los investigadores que se servían de métodos filológicos tradicionales y de su intuición lograron identificar muchas más colocaciones, documentadas con citas de textos reales. Es una prueba más del hecho de que la imagen de la lengua que nos ofrecen los corpus es y al mismo tiempo no es reflejo fiel de la lengua real.

## REFERENCIAS BIBLIOGRÁFICAS


Aronoff, M. (1976). *Word Formation in Generative Grammar*. Massachussetts: The MIT Press.

Benko, V. (2015). *Srovnatelné webové korpusy Aranea* [Aranea, corpus comparables con datos sacados de la web]. Praga: Ústav Českého národního korpusu FF UK. http://www.korpus.cz (05-05-2020).

Bloomfield, L. (1933). *Language*. London: Allen Unwin.

Bosque, I. et alii (2004). *REDES. Diccionario combinatorio del español contemporáneo,* Madrid: Ediciones SM.

Bosque, I. (2004). "Combinatoria y significación. Algunas reflexiones". En Bosque I. et alii, pp. LXXVIII-CLXXIV.

*CoDII, Collection of Distributionally Idiosyncratic Items*. https://www.english-linguistics.de/codii/ (05-05-2020).

CORIS*, Corpus di Italiano Scritto* (coord. R. Rossini Favretti). http://corpora.dslo.unibo.it/coris_ita.html (05-05-2020).

Corpas Pastor, G. (2001). "Apuntes para el estudio de la colocación", *Lingüística española actual*, 23, 1, pp. 41-56.

Cowie, A. P. (ed.) (1998). *Phraseology. Theory, Analysis and Application.* Oxford: Oxford University Press.

Cvrček, V. (2013). *Kvantitativní analýza kontextu* [Análisis cuantitativo del contexto]. Praga: Nakladatelství Lidové noviny.

Čermák, F. (2014). *Periferie jazyka. Slovník monokolokabilních slov* [Periferia de la lengua. Diccionario de palabras monocolocables]. Praga: Nakladatelství Lidové noviny.

Čermák, F., Čermák, J., Obstová, Z. y M. Vachková (2016). *Language Periphery. Monocollocable Words in English, German, Italian and Czech*. Amsterdam: John Benjamins.

Čermák, P. y M. Vavřín (2018). *Korpus InterCorp – španělština, verze 11 z 19. 10. 2018* [Corpus InterCorp – español, versión 11 del día 19-10-2018]. Praga: Ústav Českého národního korpusu FF UK. http://www.korpus.cz (05-05-2020).

De Mauro, T. (1999). *GRADIT, Grande dizionario italiano dell'uso*. Torino: UTET.

*Dizionario Garzanti di italiano* (1998). Milano: Garzanti Editore.

Dobrovoľskij, D. (1988). *Phraseologie als Objekt der Universalienlinguistik.* Leipzig: VEB.
Dobrovoľskij, D. y E. Piirainen (1994). "Sprachliche Unikalia im Deutschen. Zum Phänomen phraseologisch gebundener Formative"*, Folia Linguistica*, 28, 3-4, pp. 449-473.
Filipec, J. y F. Čermák (1985). *Česká lexikologie* [Lexicología checa]. Praga: Academia.
Firth, J. R. (1957). "A synopsis of linguistic theory, 1930–1955". En J. R. Firth et al. *Studies in Linguistic Analysis. Special volume of the Philological Society*. Oxford: Blackwell.
García-Page Sánchez, M. (1990). "Léxico y sintaxis locucionales: algunas consideraciones sobre las palabras idiomáticas", *Estudios humanísticos. Filología,* 12, pp. 279-290.
González Rey, M. I. (2005). "La noción de «hápax» en el sistema fraseológico francés y español". En Almela Pérez, R., Ramón Trives, E. y G. Wotjak. *Fraseología contrastiva: con ejemplos tomados del alemán, español, francés e italiano*. Murcia: Servicio de publicación de la Universidad de Murcia.
Halliday, M. A. K. (1966). "Lexis as a Linguistic Level", *Journal of Linguistics,* 2, 1, pp. 57–67.
Hausmann, F. J. (1984). "Wortschatzlernen ist Kollokationslernen. Zum Lehren und Lernen französischer Wortverbindungen", *Praxis des neusprachlichen Unterrichts*, 31, pp. 395–406.
Koike, K. (2001). *Colocaciones léxicas en el español actual: estudio formal y léxico-semántico*. Alcalá de Henares: Universidad de Alcalá de Henares.
Konecny, Ch. (2010). *Kollokationen. Versuch einer semantisch-begrifflichen Annäherung und Klassifizierung anhand italienischer Beispiele*. München: Meidenbauer [Forum Sprachwissenschaften 8].
Konecny, Ch. (2018). "La monocollocabilità: un fenomeno di interfaccia tra sincronia e diacronia". *PHRASIS. Rivista di studi fraseologici e paremiologici*, 2, pp. 60-76.
Mellado Blanco, C. (1998). "Aproximación teórico-práctica a los "elementos únicos" del alemán actual en su calidad de fósiles léxicos". En *Actas del 1er Congreso Hispalense de Germanistas,* Sevilla, pp. 493-501.
Obstová, Z. (2017). "Monokolokabilita ve dvou typologicky odlišných jazycích: srovnání češtiny a italštiny" [Monocolocabilidad en dos lenguas tipológicamente diferentes: comparación del checo y el italiano], *Časopis pro moderní filologii*, 99, 2, pp. 225-244.
Obstová, Z. (2018). "Esiti di un processo unicizzante o parole storicamente sprovviste di autonomia collocazionale? Uno sguardo alla diacronia delle cranberry words in italiano", *Linguistica e Filologia*, 38, pp. 57–84.
Obstová, Z. (2019). "Cranberry words tra tipologia e diacronia: l'italiano e il ceco a confronto". En Balaş, O.-D., Gebăilă, A. y R. Voicu (eds.). *Fraseologia e paremiologia: prospettive evolutive, pragmatica e concettualizzazione*, Edizioni Accademiche Italiane (Omniscriptum Group). Riga: Lettonia.
Ruiz Gurillo, L. (1997). *Aspectos de fraseología teórica española*. Anejo nº XXIV de la Revista Cuadernos de Filología. Valencia: Universitat de València.
Ruiz Gurillo, L. (1998). *La fraseología del español coloquial*. Barcelona: Ariel.
Ruiz Gurillo, L. (2001). *Las locuciones en español actual*. Madrid: Arco Libros.
Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
Veland, R. (2005). "I vocaboli a collocazione unica nell'italiano di oggi", *Mémoires de la societé néophilologique de Helsinki*, 68, pp. 331–339.
Veland, R. (2006). "Il concetto di collocazione unica e il valore di predizione della dicitura 'solo nella loc.' in uso nella pratica lessicografica", *Zeitschrift für Romanische Philologie*, 122, pp. 260–280.
Wotjak, G. (2012). "Valencia y colocabilidad: aspectos cognitivo-semánticos, morfosintácticos y pragmático-situativos". En Jiménez Juliá, T., López Meirama, B., Vázquez Rozas, V. y A. Veiga (eds.). *Cum corde et in nova grammatica. Estudios ofrecidos a Guillermo Rojo*. Santiago de Compostela: Universidade Santiago de Compostela, pp. 897–927.
Zuluaga Ospina, A. (1980). *Introducción al estudio de las expresiones fijas.* Frankfurt am Main: Verlag Peter D. Lang.