# Local Document Relevance Clustering in IR Using Collocation Information

**Leo Wanner**[*], **Margarita Alonso Ramos**[⋆]

[*] ICREA and Pompeu Fabra University
Passeig de Circumval·lació, 8; 08003 Barcelona, Spain
leo.wanner@upf.edu
[⋆] Faculty of Philology, University of La Coruña
Campus de Zapateira 15071 La Coruña, Spain
lxalonso@udc.es

## Abstract

A series of different automatic query expansion techniques has been suggested in Information Retrieval. To estimate how suitable a document term is as an expansion term, the most popular of them use a measure of the frequency of the co-occurrence of this term with one or several query terms. The benefit of the use of the linguistic relations that hold between query terms is often questioned. If a linguistic phenomenon is taken into account, it is the phrase structure or lexical compound. We propose a technique that is based on the *restricted lexical cooccurrence* (*collocation*) of query terms. We use the knowledge on collocations formed by query terms for two tasks: (i) document relevance clustering done in the first stage of local query expansion and (ii) choice of suitable expansion terms from the relevant document cluster. In this paper, we describe the first task, providing evidence from first preliminary experiments on Spanish material that local relevance clustering benefits largely from knowledge on collocations.

## 1. Introduction

The performance of an Information Retrieval engine significantly depends on the quality of its query expansion technique. The most popular expansion techniques are based on term co-occurrence: document terms that are found to co-occur significantly often with query terms are considered suitable expansion terms.[1] Depending on the strategy adopted, the query terms are processed in isolation (Sparck Jones, 1971; Schütze and Pedersen, 1994), as a set of terms (Qiu and Frei, 1993; Jing and Croft, 1994; Xu and Croft, 2000), as elements of a phrase (Mitra et al., 1997; De Lima and Pedersen, 1999), or as elements of a compound (Jacquemin and Tzoukermann, 1999; Peñas et al., 2002). Linguistically oriented strategies are largely outnumbered by statistical co-occurrence strategies. However, while the usefulness of linguistic information in IR is still questioned by some scholars, evidence is available that the performance of (especially web-based) IR can be improved by using NLP. As suggested above, so far, mainly two types of linguistic information have been used: phrase structures and lexical compounds. The goal of our work is to explore the use of lexically restricted co-occurrence, i.e., *collocation*, information for *local* query expansion. Local query expansion techniques search for suitable expansion terms the first $n$ top-ranked documents retrieved as response to the original user query.

Following (Xu and Croft, 2000), we hypothesize that top-ranked documents tend to form two clusters—a cluster of documents that are relevant to the query of the user and a cluster of documents that are irrelevant to this query. Our work is thus divided into two stages: (i) relevance clustering of the top-ranked document set; (ii) query expansion using suitable expansion terms from the relevant document cluster. In this paper, we focus on the first stage. We investigate to what extent *collocations* that occur in the original user query can be used for the relevance clustering task. The use of collocation information for query expansion is described elsewhere.

Our working document collection is the Spanish part of the document collection of the CLEF 2002 competition (Peters, 2002). For our experiments, we use top-ranked document sets retrieved within the CLEF 2002 competition by the COLE IR-system (Vilares et al., 2002).

The remainder of the paper is organized as follows. In Section 2., we introduce the phenomenon of collocation underlying our work. Section 3. provides some evidence for the significant co-occurrence of collocations in both queries as used within the CLEF competition and the document collection. Section 4. describes the preliminary experiments we carried out so far and their evaluation. Section 5. presents the conclusions.

## 2. The Phenomenon of Collocation

A collocation is a term combination $t_1 + t_2$ that expresses a concept configuration $c_1 \oplus c_2$ such that $t_1$ (the *base* of the collocation) is a standard "context free" option for the expression of $c_1$, while the choice of $t_2$ (the *collocate* of the collocation) for the expression of $c_2$ depends on the availability of $t_1$. Cf. the concept combinations 'sanctions'$\oplus$'installation', 'state of emergency'$\oplus$'installation', and 'customs duty'$\oplus$'installation'. In all three of them, $c_2$ is 'installation'. However, in connection with *sanctions*, it is expressed by the term *imposition*, in the case of *state of*

[1] Some other techniques involve, e.g., the exploitation of terms in the syntactic context of the query terms found in the document collection (Grefenstette, 1992; Ruge, 1992), or the use of most common terms in the $n$-top ranked documents obtained for the original query. The use of thesauri and lexica as source of expansion terms (e.g., hyperonyms and synonyms of query terms in the case of a thesaurus, and morphological derivates in the case of lexica) has also been suggested; cf. among others, (Hersh et al., 2000; Woods et al., 2001).

*emergency* by *declaration* and in the case of (*customs*) *duty* by *putting on*. In contrast, in German, in all three combinations, 'installation' is expressed by the term: *Verhängen* lit. 'hanging over'.

Collocations reveal the following four main features that are immediately relevant to IR (with respect to both indexing and query expansion):

- They can be classified according to a semantic typology; for instance, *pest extermination*, *overthrowing the government*, *lift of the embargo*, etc. can be viewed as being of the same type $\tau$, namely 'putting an end to X'.
- Over the semantic collocation typology, a similarity metric can be defined. Two collocation types $\tau_1$ and $\tau_2$ can be judged as similar, as, e.g., 'installation of X' and 'continuation of X' (cf. *imposition of the embargo* vs. *maintain the embargo*) or as opposite, as, e.g., 'installation of X' and 'putting an end to X' (cf. *imposition of the embargo* vs. *lift of the embargo*.
- The collocate of a collocation cannot be considered as an isolated term since its meaning as an isolated item often deviates from its collocate meaning; cf. *disposal* in *waste disposal* vs. *disposal* in *house disposal*, *raising* in *fund raising* vs. *raising* in *child raising*, *fall* in *fall of the regime* vs. *fall* in *fall of the stock*. Rather, in order to ensure its correct disambiguation, it must be considered as an element of a complex unit.
- The base of a collocation may form a collocation of the same meaning with several collocates; e.g., *deposition* in *deposition of the king* can be replaced by *dethronement* or *dethroning*; *overthrowing* in *overthrowing the regime* (*in Iraque*) can be substituted by *bringing down* or *ousting*.

The most fine-grained collocation typology available to date is the typology based on *lexical functions* (LFs) from the *Explanatory Combinatorial Lexicology* (Mel'čuk, 1996). A lexical function is a (directed) relation with a STANDARD ABSTRACT meaning that associates a lexeme $L_1$ with another lexeme $L_2$.[2] 'Standard' means that this relation applies to a large number of collocations. For instance, the relation that holds between *Shah* and *deposition* is the same as the one that holds between *government* and *overthrowing*, *minister* and *removal*, *director* and *dismissal*, and so on. It is the same in the sense that the $L_2$ lexemes provide the same linguistic features (i.e., modifications of semantic content and/or syntactic structure) to their respective $L_1$ lexemes. 'Abstract' means that the meaning of this relation is sufficiently vague and can therefore be exploited for purposes of classification.

In total, about 60 simple standard LFs are distinguished. Most common, and for IR most significant, collocations are (nominalized) verb-noun and noun-adjective collocations. In this paper, we focus on verb-noun collocations. Some of the most common standard verb-noun LFs are (in the examples below, the arguments of the LFs, i.e. the bases, are written in small capitals, their values, i.e. the collocates, in a slanted font):[3]

1. 'perform', 'do'; cf. EMBARGO:*imposition*, MOTION:*proposition* [*of*], *support*:*demonstration* [*of*]
2. 'cause existence'; cf. CONGRESS:*convention*, ELECTION:*scheduling*, OPPOSITION:*stir up*, FORTUNE:*accumulation*
3. 'put an end to'; cf. SUPPORT:*withdrawal*, FIRE:*extinguish*
4. 'act accordingly to the situation'; cf. THEOREM:*proof* [*of*], LAW:*enforcement*, RULE:*application*; RESPONSIBILITY:*acceptance*
5. 'react accordingly to the situation'; cf. ORDER:*execution*, EXAM:*pass*, ACCUSATION:*accept*

Verb-noun LFs can appear in a document as noun-verb, as noun-noun (cf. the illustrations above), and as noun-participle combinations.

In (Wanner, 2004; Wanner et al., in print), we discuss a series of machine learning techniques for the automatic recognition and classification of word bigrams in terms of the LF-typology. These techniques can be applied to the identification and classification of collocations in queries and document collections.

As mentioned above, a similarity metric can be defined over the LF-typology. For our purposes, two simple similarity clusters turned out to be already helpful:

$CL_1$ := $\{LF$, 'cause $LF$', 'begin $LF$', 'continue $LF$'$\}$

$CL_2$ := $\{$'cause end $LF$', 'end $LF$'$\}$

with $LF \in \{$'perform', 'undergo', $\ldots\}$. Elements of $CL_1$ and $CL_2$ are considered to be opposed to each other.

## 3. Collocation Distribution

Collocations are a regular phenomenon. In an average text document, a predicative key term is assumed to make part of a collocation in about 50% of its occurrences. Our evaluation shows that in the Spanish part of the CLEF 2002 material, collocations occur significantly often in both the topic descriptions of the competition (which have commonly been used as (parts of) queries) and the document collection itself. Thus, from 100 randomly chosen topic descriptions, 52 contained at least one collocation. Consider also a fragment of a document from the document collection annotated with Mel'čuk's labels of LFs.

> Los principales partidos políticos de la oposición insistieron hoy en que Felipe González <1 FinReal1 abandone> la <2Cap<1 presidencia>> del <2 Gobierno> tras conocer la <1 FinReal1 dimisión> <1 del ministro> de Agricultura, Vicente Albero, por <1 S0nonReal1 incumplimiento> de sus <1 deberes> fiscales, mientras el PSOE consideraba "adecuada" esta decisión. El secretario general del PP, Francisco Alvárez Cascos, en <1 rueda de prensa> <1 Oper1 celebrada> en un descanso de la <1 reunión> de

[2] We consider only the regular "standard" lexical functions. See (Mel'čuk, 1996) for a comprehensive overview of all types of LFs.

[3] In order not to confuse the reader, we use semantic glosses instead of Latin name abbreviations suggested by Mel'čuk as LF-labels.

> la comisión ejecutiva del PP <1 CausFunc0 convocada> con carácter de "urgencia", señaló que su partido desea trasladar a la sociedad española "un mensaje de tranquilidad, serenidad y responsabilidad ante la <1 Magn grave> <2<1 crisis>> en que González <2 CausOper1 ha sumido> al gobierno de la nación". Consideró que "a cualquier <1 Cap presidente> del <1 Gobierno> le puede salir un Roldán, pero tantos "roldanes" no le pueden salir sino a quien ha contribuido a crear un caldo de cultivo que permite que se generalice esta situación desde el ejercicio autoritario del Gobierno". El secretario de organización de IU, Mariano Santiso, calificó de "emergencia" la <situación> <CausFunc0 creada> tras la dimisión de Albero y reiteró la petición de que Felipe González <1 FinReal1 cese> <1 como presidente> del Gobierno.

Collocation elements are enclosed in angle brackets. Elements that form a collocation carry the same number, and the collocate further carries the name of the LF of which it is the value when the LF is applied to the corresponding keyword.

Individual terms often occur in different collocations. Table 1 lists the collocates with which the noun CRISIS co-occurs in the document collection and to which type of collocation the collocations formed belong. *provocar*, *plantear*, *desatar* and *abrir* are values of the same LF. They contribute the same meaning to the collocation in which they participate.

| | |
|---|---|
| 'intense' | grave, profunda |
| 'cause existence' | provocar, plantear, desatar, crear, abrir |
| 'put an end to' | atajar, combatir, superar, resolver, solucionar, poner fin, dar salida, zanjar, salir |
| 'perform' | sufrir, vivir, pasar, atravesar |
| 'cause performance' | sumir, llevar |
| 'end performance' | sacar |
| 'act appropriately' | hacer frente |
| 'continue existence' | prolongar |
| 'intensify' | agravar |
| 'begin exist' | estallar |

Table 1: Collocations with CRISIS

To estimate the frequency of the occurrence of collocations in the document collection, we selected some key terms from a number of topic descriptions and examined their co-occurrence. Cf., as illustration, the following five somewhat shortened topic descriptions:

1. *Francia extraditó a Teherán los dos iraníes sospechosos de haber asesinado a Kazem Radjavi en Suiza.*
2. *La situación política en Afganistán que llevó a la guerra civil tras la deposición del líder comunista Najibullah.*
3. *Qué efectos ha tenido el embargo de la ONU en la vida del pueblo iraquí?*
4. *Qué medidas ha tomado Irak para lograr el levantamiento del embargo económico de la ONU, así como de las sanciones políticas impuestas después de su invasión de Kuwait en 1990?*
5. *Peticiones públicas de dimisión para Felipe González por parte de personalidades políticas en España.*

For each underlined key term, we examined its occurrence in collocations in the first 100 top-ranked documents retrieved by the COLE-system (Vilares et al., 2002) for the query that contained the topic description in question; cf. Table 2 for the distribution figures.

| key term | in colloc. | isolated |
|---|---|---|
| extradición | 31 | 62 |
| sospechoso(s) | 0 | 30 |
| asesinato | 11 | 42 |
| situación | 17 | 13 |
| guerra civil | 32 | 46 |
| guerra | 36 | 135 |
| deposición | 0 | 0 |
| efecto(s) | 9 | 2 |
| embargo | 269 | 84 |
| vida | 1 | 9 |
| medidas | 25 | 22 |
| levantamiento | 214 | 2 |
| embargo | 165 | 35 |
| sanciones | 246 | 73 |
| invasión | 15 | 106 |
| petición | 18 | 20 |
| dimisión | 133 | 140 |

Table 2: Frequency of the occurrence of certain key terms in collocations

## 4. Experiments

The distribution of collocations in the queries and in the document collection led us to the following two assumptions: (i) if the user uses a collocation $Co$ in a query at least some of the documents that contain $Co$, will be relevant to this user; (ii) at least some of the documents that do not contain $Co$, will be irrelevant to the user.

We can thus reformulate the hypothesis underlying *Local Context Analysis* in (Xu and Croft, 2000) as follows:

> If a user query $Q$ contains an instance $I$ of an LF $L$, documents that are relevant to $Q$ are likely to contain $I$ or instances of typologically similar LFs applied to the same base as $L$. But they are unlikely to contain instances of LFs (applied to the same base as $L$) that are typologically distant to $L$.

In other words, the documents of the top-ranked set can be clustered into relevant and irrelevant subsets, depending on the extent to which they contain the same or similar instances of LFs that occur in $Q$

To verify this hypothesis, we use the metric in (1) that calculates, for a document $D$, a collocational weight $W_D$. $W_D$ reflects the degree to which the collocations in $D$ match the collocations in $Q$. If $W_D \geq \Delta$ ($\Delta$ being empirically determined), $D$ is considered relevant; otherwise it is considered irrelevant. The experiments below have been carried out with $\Delta = 0.5$.

$$W_D := \frac{N_{Co_D}}{N_{Co_Q}} \times \sum_{Co \in Q} \frac{f(Co)}{f(B)(1 + N_{Co^-})} \qquad (1)$$

$N_{Co_Q}$ being the number of different LFs in $Q$, $N_{Co_D}$ the number of different LFs from $Q$ that occur $D$, $f(co)$ the frequency of the collocation $co$ or a similar instance of the same LF in the document $D$, $f(B)$ the frequency of the base $B$ of $co$ in $D$, and $N_{Co^-}$ the total number of the LF-instances with $B$ in $D$ that are opposed to the LF of which $co$ is an instance.

The advantage of an LF-metric over a term co-occurrence metric is its generalization potential: it covers all semantically similar term sequencies rather than only one single term sequence.

In a series of preliminary experiments, we clustered a number of 100 top-ranked documents sets returned by the IR-system of the COLE-group of the University of La Coruña. Table 3 shows the *fallout* $f$, *precision* $p$ and *recall* $r$ of five clustering runs on five different 100 top-ranked sets.

| $f$(allout) % | $p$(recision) % | $r$(ecall) % |
|---|---|---|
| 97.77 | 50.00 | 28.57 |
| 61.40 | 56.86 | 67.44 |
| 51.35 | 67.92 | 65.45 |
| 80.64 | 10.00 | 50.00 |
| 97.80 | 66.67 | 50.00 |

$f = \frac{NR_Q}{NR_d}$, with $NR_Q$ as the number of non-relevant documents recognized by the metric, $NR_d$ as the number of non-relevant documents in the corresponding top-100 set; $p = \frac{R_Q}{R_m}$, with $R_Q$ as the number of relevant documents recognized by the metric and $R_m$ as the number of documents classified by the metric as relevant; $r = \frac{R_Q}{R_d}$, with $R_d$ as the number of relevant documents in the corresponding top-100 set.

Table 3: Quality figures of five clustering runs.

The table reveals that $f$ is consistently high, which means that the metric functions well for filtering out irrelevant documents from the top-ranked sets. $p$ and $r$ may vary significantly—depending on the LFs involved in the query: instances of certain LFs are better discriminators than instances of other LFs.

## 5. Conclusions

We argued that collocation information is an important type of linguistic information that must be taken into account for local document relevance clustering and for query expansion. We presented some initial evidence for the importance of collocations for local document relevance clustering. More work is needed to obtain reliable figures that reveal to what extent query expansion improves when collocations in queries are being taken into account. A topic we still did not explore so far at all is the role of collocations in indexing.

## 6. References

E. De Lima and J. Pedersen. 1999. Phrase Recognition and Expansion for Short, Precision-Biased Queries Based on a Query Log. In *Proceedings of the 22nd SIGIR Conference*, pages 145–152. ACM Press, NY.

G. Grefenstette. 1992. Use of Syntactic Context to Produce Term Association Lists for Retrieval. In *Proceedings of the 15h SIGIR Conference*, pages 89–97. ACM Press, NY.

W. Hersh, S. Price, and L. Donohoe. 2000. Assessing Thesaurus-Based Query Expansion Using the UMLS Metathesaurus. *Journal of Americian Medical Informatics Association*, Symposium Supplement.

C. Jacquemin and E. Tzoukermann. 1999. Nlp for term variant extraction: Synergy between morphology, lexicon, and syntax. In T. Strzalkowski, editor, *Natural Language Processing Information Retrieval*, pages 25–74. Kluwer, Boston.

Y. Jing and W.B. Croft. 1994. An association thesaurus for information retrieval. In *Proceedings of the RIAO Conference*, pages 146–160.

I.A. Mel'čuk. 1996. Lexical Functions: A Tool for the Description of Lexical Relations in a Lexicon. In L. Wanner, editor, *Lexical Functions in Lexicography and Natural Language Processing*, pages 37–102. Benjamins Academic Publishers, Amsterdam.

M Mitra, C. Buckley, A. Singhal, and C Cardie. 1997. An analysis of statistical and syntactic phrases. In *Proceedings of the Fifth RIAO Conference*.

A. Peñas, J. Gonzalo, and F. Verdejo. 2002. Distinción semántica de compuestos léxicos en recuperación de información. *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*, 26.

C. Peters. 2002. *Results of the CLEF 2002 Cross-Language System Evaluation Campaign, Working Notes for the CLEF 2002 Workshop*. Rome, Italia.

Y. Qiu and H.P. Frei. 1993. Concept based query expansion. In *Proceedings of the 16th SIGIR Conference*, pages 160–169. ACM Press, NY.

G. Ruge. 1992. Experiments on linguistically-based term associations. *Information Processing and Management*, 28(3):317–332.

H. Schütze and J. Pedersen. 1994. A cooccurrence-based thesaurus and two applications to Information Retrieval. In *Proceedings of the RIAO Conference*, pages 266–274.

K. Sparck Jones. 1971. *Automatic KeyWord Classification for Information Retrieval*. Butterworths, London.

J. Vilares, M.A. Alonso, F.J. Ribadas, and M.Vilares. 2002. COLE Experiments at CLEF 2002 Spanish Monolingual Track. In C. Peters, editor, *Results of the CLEF 2002 Evaluation Campaign*, pages 153–160.

L. Wanner, B. Bohnet, and M. Giereth. in print. Making Sense of Collocations. *Computer Speech and Language*.

L. Wanner. 2004. Towards Automatic Fine-Grained Semantic Classification of Verb-Noun Collocations. *Natural Language Engineering Journal*, 10(2):95–143.

W. Woods, S. Green, P. Martin, and A. Houston. 2001. Aggressive morphology and lexical relations for query expansion. In *The Tenth Text REtrieval Conference (TREC 2001)*.

J. Xu and W.B. Croft. 2000. ACM Transactions on Information Systems. *Improving the Effectiveness of Information Retrieval with Local Context Analysis*, 18(1):79–112.