An attempt to improve current collocation analysis

Pascual Cantos-Gómez
Universidad de Murcia (Spain)

Abstract

Our goal is to review the main stream of research on collocation analysis and try to overcome some of its intrinsic problems, such as the optimal span and the reason for *undesired* collocates (statistically significant collocates, though lexically and semantically not related with the node word). The idea is not just to calculate significant collocates of a chosen node word or to elucidate which is the best statistical procedure to achieve this goal. However, we shall focus on the way words *socialise* with other words, forming complex network-like structures or units: *lexical constellations*; something that cannot be explained using current collocation analyses.

1. Introduction

Conventional or *prepatterned* expressions are often loosely referred to as *collocations*. However, since the term *collocation* is also used in a stricter sense to denote a special kind of lexical relationship, it is convenient to clarify these concepts.

In its broadest sense, *collocation* is more or less equivalent to: recurrent word combination. In the Firthian tradition, however, it is generally used in a stricter sense: a collocation consists of two or more words which have a strong tendency to be used together. According to Firth (1968: 181), "collocations of a given word are statements of the habitual or customary places of that word." That they are habitually co-occurring lexical items or mutually selective lexical items (Cruse 1986: 40). For example, in English you say *rancid butter*, not *sour butter*, or *burning ambition* instead of *firing ambition*.

Both interpretations imply a syntagmatic relationship between linguistic items, but whereas the broader sense focuses on word sequences in texts, the stricter one goes beyond this notion of textual co-occurrence and emphasises the relationship between lexical items in language (Greenbaum 1974: 80). It follows that the former, but not necessarily the latter, includes idioms, compounds and complex words, and that the latter extends also to discontinuous items.

There are also intermediate uses of the term, where collocations can be thought of as lying on a continuum. At one end of the continuum lie the idioms like *as soft as butter*. At the other end of the continuum lie free combinations. These are word combinations that can be formed by simply applying grammatical rules. For example, *the table* would be the result of the grammatical rule that states that the definite article precedes a countable noun.

The Firthian notion of collocation is thus a more extensive lexical concept than recurrent word combination. Consequently, methodologically it is viewed as essentially a probabilistic phenomenon consisting of identifying statistically significant collocations and excluding fortuitous combinations. This means that collocation analysis can be dealt with quantitatively and to some extent also automatically.

Sinclair, within the Firthian tradition, defines collocation as "the occurrence of two or more words within a short space of each other in a text" (1991: 170). In collocation analysis, interest normally centres on the extent to which the actual pattern of these occurrences differs from the pattern that would have been expected, assuming a random distribution of forms. Any significant difference can be taken as, at least, preliminary evidence that the presence of one word in the text affects the occurrence of the other in some way.

In what follows, we shall briefly discuss: (1) the ways in which actual and expected patterns of co-occurrence can be computed and compared, and (2) the measures that can be applied to the results to assess their significance and explore their implications.

2. Extracting Collocations

A significant collocation can be defined in statistical terms as the probability of one lexical item co-occurring with another word or phrase within a specified linear distance or span being greater than might be expected from pure chance.

It is not our goal here to discuss and compare the various measures, though some points are important to note. Probably, the most commonly used statistical measures to determine collocate significance are *z-score* (Berry-Rogghe 1973), *t-score* and *mutual information* (Church and Hanks 1990). All three significance measures can be used to highlight words that appear to be most strongly

collocated with the node word. Strong collocates are somehow equally highlighted by all three measures; however, *z-score* and the *mutual information* measure artificially inflate the significance of low-frequency co-occurring words because of the nature of their formulae.

There are important differences between the information provided by the three measures: more, perhaps, between *t-score* and the other two than between *z-score* and *mutual information* themselves. It is difficult, if not impossible, to select one measure that provides the best possible assessment of collocates, although there has been ample discussion of their relative merits (see, for example, Church et al. 1991; Clear 1993; or Stubbs 1995). It is probably better to use as much information as possible in exploring collocation and to take advantage of the different perspectives provided by the use of more than one measure.

Geffroy et al. (1973) produced a formula for the strength of collocation (called *C*) which took into account both the frequency of co-occurrence and the proximity of the collocates to each other. They employed cut-off points for minimum values of *C* and *f* (frequency) such as 5 and 20, respectively.

Similarly, Smadja et al. (Smadja and McKeown 1990, Smadja 1991) take into account word distance as well as word strength (spread) for a measure of word association (height) within a limited span: 5 words on either side (see Martin et al. 1983). By means of *XTRACT,* Smadja proposes an approach for automatically acquiring co-occurrence knowledge from statistical analysis of large corpora. Smadja assumes that two words are associated if they appear in the same sentence and are separated by less than five words. For each pair of collocating words, a vector of 10 values is created. In a flexible collocation, the words may be inflected, the word order may vary and the words can be separated by any number of intervening words (i.e. *to take steps, took immediate steps, steps were taken*). If the word order and inflections are fixed and the words are in sequence, the collocation is said to be rigid (e.g. *International Human Rights*).

Kita et al. (1994) also propose a mathematical approach to automatically compiling collocations: the *cost criteria*. However, their approach differs slightly from the others mentioned so far in that they consider collocations to be cohesive word clusters, including idioms, frozen expressions and compound words. This approach focuses just on linearly fixed multiple word units, being unable to determine discontinuous co-occurrences and/or word associations. Kita et al.'s measure relies on the length of the collocate or word sequence and on its frequency. This approach is much more in the line of research that understands collocation within the broader definition (see above) and consequently fails to account for word associations, lexical relations or discontinuous collocations.

There are also other approaches such as Lafon's use of combinatorics to determine collocational significance (Lafon 1984), Daille's approach to monolingual terminology extraction (Daille 1995) or Dunning's *log-likelihood measure* also known as *g-square* or *g-score* (Dunning 1993), among others.

3. Collocations Revisited

In what follows, we shall re-examine some basic notions regarding collocations with reference to real data. This, we are confident, will give us a more comprehensive view of the complex interrelationships (semantic, lexical, syntactic, etc.) between co-occurring linguistic items.

Three starting assumptions were made: (1) we understand collocates in the stricter sense, as already discussed: that is, words that have a strong tendency to co-occur around a node word in a concordance; (2) in order to avoid being biased in favour of any data, we concentrated on the occurrences of a lemma within a single semantic domain: the Spanish noun *mano* (extracted from the *CUMBRE Corpus* –16.5 million tokens–, by SGEL); and (3) no optimal span was taken for granted. This explains why we took the full sentence as the concordance (though important collocates can be missed due to anaphoric reference) .

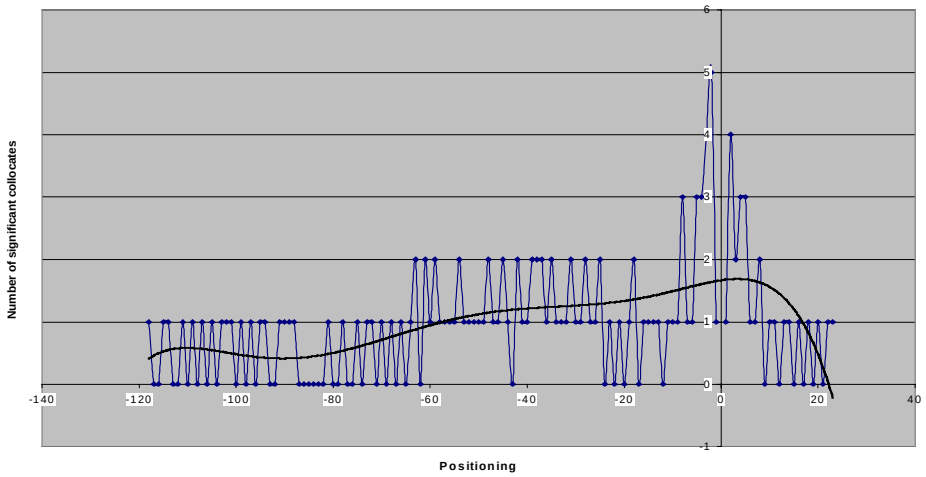
To make sure that we were actually focusing on the same semantic contexts, we first extracted all instances (full sentences) containing *mano*, both in singular and plural form and next, classified all occurrences semantically by means of the various definitions for *mano*.

This preliminary distributional analysis allowed us to isolate the various meanings and to concentrate on individual meanings or identical semantic contexts. In the analysis that follows, we shall concentrate on one of the meanings of *mano* ("layer of paint, varnish or other substance put on a surface at one time").

As already noted, no preliminary assumptions were made regarding relevant or optimal window size or span. We simply started by taking the whole sentence in the belief that full sentences are more likely to contain complete ideas, meanings, etc. Fixing a span beforehand might often distort the *unity of meaning* or *complete meaning* in a sentence. We tend to think that words outside the range of a sentence are not likely to affect the *meaning unity* of other sentences, nor are they strongly associated to node words within other sentences.

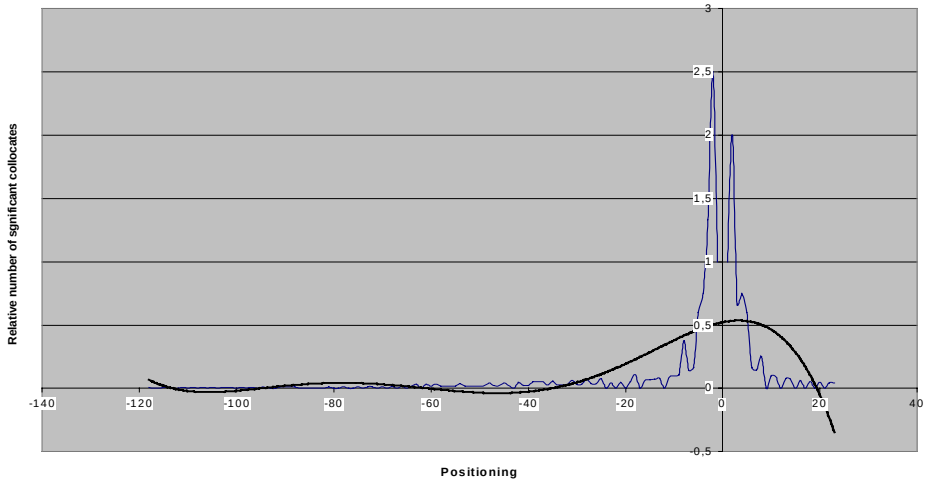
The distribution of significant collocates (*z-scores*) results in the following graph (*Figure 1*), where the *x-axis* displays the position relative to the node word and the *y-axis* the number of statistically significant collocates.

Fig.1. **Distribution of significant collocates vs trendline (z-score: *mano*)**



In order to get a more accurate and neater visual representation of the distribution of the significant collocates than the zigzag curves above, we normalised the data of the *y-axis* (number of significant collocates), by dividing the number of significant collocates found at each position by its positioning with respect to the node word (*Figure 2*). The *y-axis* represents the relative number of significant collocates. *Figure 2* visually represent the extent of influence of the node on its immediate environment: the *lexical gravity*. Note that our notion of gravity differs radically from the one proposed by Sinclair et al. (1998; see also Mason 1997): (1) we do not calculate the extent of influence of the node on its immediate verbal environment, but take the node itself; (2) Mason's one is a more heavily calculated statistic one and (3) instead of displaying gravity as crater shapes, we do it in peak shapes. The graph stresses the attraction of *MANO(layer of paint, etc.)*, which is greatest in the immediate environment and wears off with distance, which is indeed an already established fact.

Fig. 2. **Gravity vs trendline (*mano*)**



If we observe and compare the actual gravity with its trendline, we realize that there is a kind of correlation between the relative number of significant collocates and the node distance, particularly close around the node word. Intuitively, we can with some confidence state that the most relevant collocational context for *MANO(layer of paint, etc)* is between –8 and +6. It is also interesting that data around the node tends to have a more normal distribution than overall. However, evidence shows that fixing an optimal span could be misleading as the appearance of significant collocates is likely to exceed these pre-established limits.

Against the background of this data, it becomes clear that the distribution of significant collocates is not necessarily normal. This has important consequences as it speaks against generic assumptions on optimal spans (see i.e. Jones and Sinclair 1973; Martin et al. 1983; Smadja 1989; or Berber 1997, among others). Phillips is right to claim that "it is probably sensible to err in favour of over-inclusiveness" (Phillips 1989: 23).

This led us to state two hypotheses on collocates: H_1: each word has a unique idiosyncratic linguistic behaviour (different attraction strength or power), at least lexically speaking; and H_2: not all significant collocates of a node word are actually attracted by it, only some. Others are likely to be attracted by other significant collocates within the same context (sentence, phrase, concordance line, etc.).

Regarding H_1, it is clear that our empirical data is not sufficient to accept or reject this hypothesis. However, other related research such as Mason's (1997) gravity study and Sinclair et al. (1998) clearly support it. They talk about the importance of establishing a reasonable span for the collocational analysis of a word. They go even further by saying that there is a need to have some means of justifying the span to be used for each word. The key is to discover the extent of node effect on its environment (whether the optimum span varies according to the node's grammatical class, semantic range, between word-forms of a lemma, *richness* of lexis and so on, remains to be seen). These findings clearly indicate a change compared with those of Martin et al. (1983) and Jones and Sinclair (1973), who fixed the span of influence of the node on its environment, irrespective of the grammatical class of the node, its semantic range, etc, and concluded that the optimal span is –5 +5 and –4 +4, respectively. Summing up, this means that each word has a distinctive range of influence on other words and we, necessarily, need to determine the gravity of each word in order to arrive at or calculate its significant span.

H_2 states that not all significant collocates are actually attracted by the same word or node; only some are. Others are likely to be attracted by other significant collocates. This hypothesis does not focus on the span; it is even *span-independent*. What matters here is not just the span or extent of the influence of the node on its immediate environment but the fact that collocates (statistically significant ones) within the same environment are not necessarily attracted by the same node. This introduces a striking difference with regard to H_1. By H_2, we understand that collocates within the same environment build ordered and structured frames: not flat lexical frames (traditional view of collocates, see H_1), but complex interrelated hierarchies similar to *constellations* consisting of a nucleus (e.g. the Sun) which attracts various other stars with each star attracting various other moons. Collocates form lexical conceptual multi-dimensional frames: lexical constellations. Constellations themselves can be substructures (subsets) of others (e.g. the Solar System as part of the Milky Way) or superstructures (supersets) subsuming other structures (e.g. the Solar System containing Jupiter and its Moons).

Both hypotheses seem plausible, particularly the first one, if we consider the main stream and recent advances made in corpus linguistic research on collocational analysis (see e.g. Daille 1995; Mason 1997; Smadja 1992; Samdja et al. 1996; Sinclair et al. 1998). However, H_1 somehow fails to comfort to one major psychological feature of linguistic data: efficiency of storage of lexical items (Worff 1991). This speaks against the economising role or purpose of language production (Coulmas 1981, Cowie 1981, Nattinger 1980, 1988, etc.). There are features that indicate that the effect or influence of a word on others varies significantly among words (see Berry-Rogghe 1974). However, the range of variability cannot be unlimited or very large. Consequently, H_1 is, at least to its full extent, not fully plausible psychologically, as humans have a limited memory and storage capacity.

H_2 seems in this respect more realistic, both psychologically and linguistically. Evidence shows that words exert some influence on others, forming units (semantic units, tone units, syntactic structures, such as phrases, etc.). Furthermore, these units are to some extent cognitively, spatially and temporarily limited, though not necessarily universally fixed or equally limited. Several principles speak in favour of this: minimal attachment, right association and lexical preferences (Allen 1995: 160-2). This means that the attraction domain of the node needs to be not just limited but also constrained, though not necessarily fixed or universally predetermined: the extent of influence is likely to vary depending on grammatical category, semantic range, etc. This assumption significantly reduces the range of variability of the various possible attraction domains (optimal spans) of node words. In addition, this also fits in with the unification-based assumption of the *constellation* principle: ordered hierarchical elements forming autonomous structures, sub-structures or super-structures.

This explains the appearance of significant collocates for *MANO(layer of paint, etc.)* at positions –99, –89, –38, –28, –19 and +20, among others. If they are not attracted by the node word, then which items exert their influence on them and attract them? We might explain these facts as follows: the high *z-scores* categorise them as significant collocates within the sentence contexts of *MANO(layer of paint, etc.)*; however, they are likely to be attracted by other words, not necessarily by *mano*. That is, these

collocates are part of another unit, not the immediate unit or vicinity of the node word.

From the data, it becomes apparent that, for instance, *profesora* (−118) −a significant collocate within the concordance sentence of *MANO(layer of paint, etc.)*– might not be semantically attracted by *mano*, but by other lexical items within the same context, e.g. by *Universidad* (position −115). And, *Autónoma* (−114) is probably attracted by *Universidad* (position −115), too. The reason for these significant collocates, within the context sentences of *MANO(layer of paint, etc.)*, is not related to the influence the noun *mano* exert on them, but is due to the fact that they are attracted by other items. Indeed, they seem to be part of a different unit (i.e. part of a different phrase). This means that the direct influence of words on others is not just limited but also somehow structured.

To illustrate this, let us take one concordance sentence for *MANO(layer of paint, etc.)*: "Si algún tono no tiene la intensidad deseada, aplique otra {MANO} de pintura" ("If the shade hasn't got the desired intensity, apply another layer of paint"). This *mano* sentence contains just six statistically significant collocates: *algún*, *tono*, *intensidad*, *deseada*, *aplique* and *pintura*, distributed from –9 to +2 (a positively skewed influence distribution). This gives the following collocational distribution:

		Positioning					
		–10	–9	–8	–7	–6	–5
Freq.	1	Si	algún	tono	no	tiene	la

		Positioning					
		–4	–3	–2	–1	+1	+2
Freq.	1	intensidad	deseada	aplique	otra	de	pintura

A simple flat collocational analysis would just reveal that there are six collocates likely to be statistically significant for *mano*. Furthermore, if we had fixed the optimum span to –5 +5 or –4 +4, we would have missed at least: *algún* and *tono*.

To know more about the possible lexical hierarchy, we started by calculating the expected co-occurrence probability of all possible significant collocate tuples (combinations) within the same sentence, that is, the probability that event *x* occurs, given that the event *y* has already occurred. For example, taking the occurrence data for *tono* and *intensidad* we get:

	Tono
Intensidad	11
Without *intensidad*	1037
	1048

	Intensidad
Tono	11
Without *tono*	640
	651

P(intensidad□tono) = 11 / 1048 = 0.0105
P(tono□intensidad) = 11 / 651 = 0.0169

This shows that *tono* is lexically less tied to *intensidad* and more likely to appear without it, whereas *intensidad* is much more dependent on *tono* and less likely to occur without it. This information reveals that *tono* is lexically less tied to *intensidad* and more likely to appear without it, whereas *intensidad* is more dependent on *tono*. The probability figures indicate that the lexical attraction is mutual and bi-directional, but not equidistant: *tono* exerts a greater influence and attraction on *intensidad* than vice versa. Sinclair already realised this fact (1991: 115-116), coining the terms *upward* and *downward collocation*. These two types of collocation are explained using the term *node* (for the word that is being studied) and the term *collocate* (for any word that occurs in the specified environment of a node). Coming back to our previous example, if we take *tono* as the node and *intensidad* as the collocate, what we have is *downward collocation*: collocation of a node (*tono*) with a less frequent word (*intensidad*). On the contrary, if *intensidad* is the node and *tono* the collocate then we have *upward collocation*. The systematic difference between upward and downward collocation is that the former is the weaker pattern in statistical terms, and the words tend to be elements of grammatical frames, or subordinates. Downward collocation by contrast gives us a semantic analysis of words (Sinclair 1991: 116).

Once the attraction direction procedure was established (upward or downward collocation), we determined all possible tuples of *mano* by means of the given statistically significant collocates (namely, *tono*, *intensidad*, *deseada*, *aplique* and *pintura*). We discarded *algún* as it is not a content word or open-class lexical item. Potentially, the possible tuples or combinations within the same context sentence range for *mano* from single-word ones up to six-word clusters (the total number of

statistically significant collocates plus the node word *mano*), that is, the sum of all combinations of *r* objects (number of co-occurring items) from *n* (set of statistically significant collocates). Thus, the combination of *r* objects from *n* is calculated by means of: $C(n, r) = n! / (n – r)!r!$

This totals an overall of 63 possible order-independent combinations for *mano* (6 + 15 + 20 + 15 + 6 + 1). However, the singletons (single-word combinations) are practically irrelevant as they just indicate their own probability of occurrence in the corpus, irrespective of co-occurrence with other items.

According to the data and probability calculations for *tono* and *intensidad* above, we believe that the attraction direction is somehow inherently determined by the observed frequency (consider also Sinclair's notion of *downward* and *upward collocation* (Sinclair 1991: 115-116)). This means that, between two items, the one that occurs more often in a language model (i.e. corpus) is likely to exert a greater influence on the less occurring one(s). This explains why we do not care about the actual order of the combinations as this is to some extent predetermined by the observed frequencies of their constituents.

In order to concentrate on the really significant combinations, we took a number of preliminary decisions. First, all singletons were discarded as they do not give any information as to which other items they combine with. Next, we discarded all combinations with an occurrence frequency = 1. The reason is that all significant collocates occur, at least once, that is, whenever they co-occur in the concordance sentence itself. Consequently, combinations with a frequency lower than 2 are irrelevant, as they do not contribute anything to their association power with other items. So, among all potential combinations, we discarded the hapax tuples (combinations with an occurrence frequency = 1) and the singletons (single words). This resulted in the following significant combinations for *mano*:

TOKEN(S)	FREQ(corpus)	PROB(corpus)
mano,pintura	17	6.19820E-14
tono,intensidad	11	4.01060E-14
pintura,tono	8	2.91680E-14
mano,tono	7	2.55220E-14
mano,intensidad	6	2.18760E-14
intensidad,deseada	3	1.09380E-14
mano,aplique	2	7.29200E-15
tono,deseada	2	7.29200E-15
tono,intensidad,deseada	2	4.40306E-22

We find that the most likely tuple for *mano* is *mano-pintura*. This means that, in fact, *mano* is responsible for its co-occurrence with *pintura*, but not necessarily the other way round, as *mano* is more likely to be found than *pintura*. This also applies to the tuple *tono-intensidad*, producing these two ordered pairs:

```
mano    (pintura
        )

tono    (intensidad
        )
```

This means that the collocate *intensidad*, though significant within the context of *MANO(layer of paint, etc.)*, is not directly attracted by *mano* but by another significant collocate (*tono*) within the same context. The third most likely combination is *pintura-tono*. Here we have now the missing link between *mano-pintura* and *tono-intensidad*. The two previous combinations can be merged and reduced into a single structure, giving:

```
mano    (pintura
                (tono
                        (intensidad
                        )
                )
        )
```

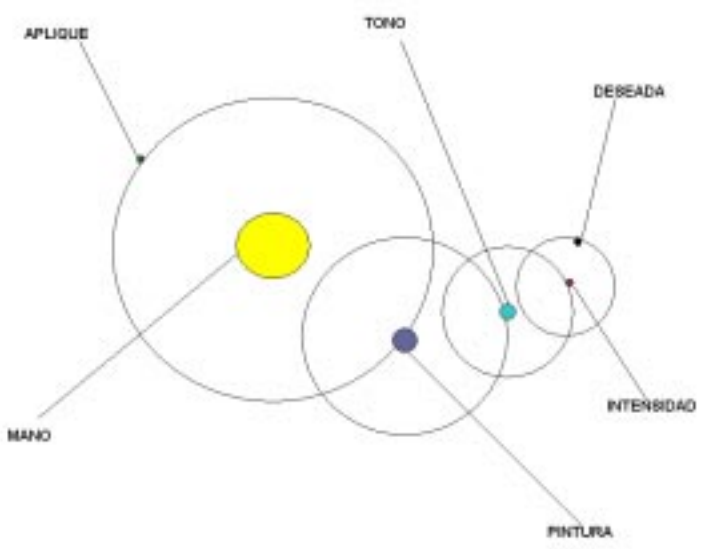
The next tuple *mano-tono* is already entailed in the structure above, and indicates that *mano* does indeed attract *tono*. However, evidence shows that it is actually *pintura* that exerts a greater influence on *tono*, or, in other words, *mano* attracts *tono* by means of *pintura*. If it were not for *pintura*, it is unlikely that *tono* would be a collocate of *mano* or, at least, directly attracted by it. A similar case is that of *mano-intensidad*. The next combination, *intensidad-deseada*, enlarges the structure into:

mano (pintura
(tono
(intensidad
(deseada
)
)
)
)
Mano-aplique brings along a new directly attracted collocate of *mano*:
mano (pintura
(tono
(intensidad
(deseada
)
)
)
)
(aplique
)
and finally, *tono-deseada* and *tono-intensidad-deseada* which are already entailed in the structure above.

We conclude that, in this example sentence, within the context of *MANO(layer of paint, etc.)*, *mano* is the dominant collocate or node as it directly attracts *pintura* and *aplique* and indirectly attracts most of collocates (*tono*, *intensidad* and *aplique*), except for *deseada*. The resulting collocational hierarchy is:

MANO (pintura
(tono
(intensidad
(deseada
)
)
)
)
(aplique
)

Or in a more visual mode, very similar to a 3D star constellation:



4. Conclusions

Constellations represent lexical relations among collocates in a structured hierarchical way *versus* the standard simple representation of most collocational analyses. It is precisely the plain vision of standard collocational work that misses two main features of collocational behaviour: (1) their non-

linear dependency: not all collocates are necessarily attracted by the same node word and, furthermore; (2) those that co-occur within a determined unit (phrase, sentence, etc.) form a kind of complex lexical hierarchy.

It is also interesting that by means of constellational analysis we might also get rid of the optimum span dilemma, as this analysis always takes the sentence unit. What matters is not the span, but the lexical hierarchy collocates form within a linguistic unit (sentence, phrase, concordance line, etc.). The lexical hierarchies are neither predetermined nor assumed (as happens in most semantic networks or word sense hierarchies); they result from the co-occurrence probabilities and the attraction calculations, which might lead to divergences depending on the corpus or language samples analysed.

The potential implementation of this constellation analysis could, of course, go much further by means of incorporating lemmatisation and part-of-speech tagging. This would prevent duplicate lexical relations (i.e. homonymy, polysemy) and part-of-speech ambiguity, and it would refine frequencies, producing more accurate constellational models.

Notice that we have not tackled here the problem of which statistical collocation model or method to use, as this is not a basic issue for constellational analysis. Any statistical model would, in principle, do; this is up to the researcher's preferences. Of course, much research is still needed to totally configure this new trend in collocational analysis and to sort out hierarchical discrepancies/similarities (e.g. very close statistical significances).

We are confident that this new view within collocational analysis can be a positive contribution. The approach presented is transparent in nature and empirically testable, producing a good description of the storehouse of vocabulary items. That is, how words are connected through their associations and how lexical knowledge is schematic and associative. This could lead to improved results in computational lexicography (development of collocational dictionaries, automatic construction and typification of proto-senses, etc.), language pedagogy (the teaching/learning of collocational knowledge for L2 learners) and Machine Translation (particularly to EBMT), to mention some important applications.

Similarly, there are also various research interests within the lexical constellation framework. This includes the capturing and/or mapping of syntactic structures (i.e. exploring whether lexical constellations capture the representation of syntax) by means of various linguistic models (Compositional Grammars, Construction Grammars, etc.) or specific parsing strategies (e.g. Data Oriented Parsing).

References

Allen J 1995 *Natural Language Understanding.* Redwood: The Benjamin/Cummings Publishing Company.

Berber Sardinha AP 1997 Lexical Co-occurrence: A Preliminary Investigation into Business English Phraseology. *Letras & Letras*, 13/1: 15-23.

Berry-Rogghe GLM 1973 The Computation of Collocations and their Relevance in Lexical Studies. In AJ Aitken, R Bailey, N Hamilton-Smith (eds) *The Computer and Literary Studies*. Edinburgh: Edinburgh University Press.

Berry-Rogghe GLM 1974 Automatic Identification of Phrasal Verbs. In J.L. Mitchell (ed) *Computers in the Humanities*. Edinburgh: Edinburgh University Press.

Church KW, Hanks P 1990 Word Association Norms, Mutual Information and Lexicography. *Computational Linguistics*, 16/1: 22-9.

Church KW, Gale W, Hanks P, Hindle D 1991 Using Statistics in Lexical Analysis. In U. Zernik (ed) *Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon*. Hillsdale, NJ: Lawrence Erlbaum Associates, 115-164.

Clear J 1993 From Firth Principles: Computational Tools for the Study of Collocation. In M. Baker et al. (eds) *Text and Technology*. Amsterdam: Benjamins, 271-292.

Coulmas F 1981 Introduction: Conversational Routine. In F Coulmas (ed) *Conversational Routine: Explorations in Standardized Communication Situations and Pre-patterned Speech*. The Hague: Mouton, 1-17.

Cowie AP 1981 The Treatment of Collocations and Idioms in Learners's Dictionaries". *Applied Linguistics*, 2: 223-235.

Daille B 1995 Combined Approach for Terminology Extraction: Lexical Statistics and Linguistic Filtering. *UCREL Technical Papers. Volume 5*. Lancaster: University of Lancaster Press.

Dunning T 1993 Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19/1: 61-74.

Firth J 1968 A Synopsis of Linguistic Theory 1930-1955. In FR Palmer (ed) *Selected Papers of J.R.*

Firth 1952-59. Bloomington: Indiana University Press, 1-32.
Geffroy A, Lafon P, Seidel G, Tournier M 1973 Lexicometric Analysis of Co-occurrences. In AJ Aitkien, R Bailey and N Hamilton-Smith (eds) *The Computer and Literary Studies*. Edinburgh: Edinburgh University Press.
Greenbaum S 1974 Some Verb-Intensifier Collocations in American and British English. *American Speech*, 49: 79-89.
Jones S, Sinclair J 1973 English Lexical Collocations. A Study in Computational Linguistics. *Cahiers de Lexicologie*, 23/2: 15-61.
Kita K, Kato Y, Omoto T, Yano Y 1994 Automatically Extracting Collocations from Corpora for Language Learning. In A Wilson, T McEnery (eds) *UCREL Technical Papers. Volume 4. Corpora in Language Education and Research. A Selection of Papers from TALC94*. University of Lancaster. 53-64.
Lafon P 1984 *Dépouillements et statistiques en lexicométrie*. Geneva: Slatkine-Champion.
Martin W, Al B, van Sterkenburg P 1983 On the Processing of a Text Corpus: From Textual Data to Lexicographical Information. In R Hartmann (ed) *Lexicography: Principles and Practice*. London: Academic Press.
Mason O 1997 The Weight of Words: An Investigation of Lexical Gravity. In *Proceedings of PALC'97*, 361-375.
Nattinger J 1980 A Lexical Phrase Grammar for ESL. *TESOL Quarterly*, 14: 337-344.
Nattinger J 1988 Some Current Trends in Vocabulary Teaching. In C. Carter and M. McCarthy (eds) *Vocabulary and Language Teaching*. London: Longman, 62-82.
Phillips M 1989 *Lexical Structure of Text*. Birmingham: ELR/University of Birmingham.
Sinclair J 1991 *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
Sinclair J, Mason O, Ball J, Barnbrook G 1998 Language Independent Statistical Software for Corpus Exploration. *Computers and the Humanities*, 31: 229-255.
Smadja FA 1989 Lexical Co-occurrence: The Missing Link. *Literary and Linguistic Computing*, 4/3: 163-168.
Smadja FA 1992 XTRACT: An Overview. *Computers and the Humanities*, 26/5-6: 399-414.
Smadja FA 1991 Macro-coding the Lexicon with Co-occurrence Knowledge. In U Zernik (ed) *Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon*. Hillsdale, NJ: Lawrence Erlbaum Associates, 165-189.
Smadja FA, McKeown KR 1990 Automatically Extracting and Representating Collocations for Language Generation. *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, 252-259.
Smadja FA, McKeown KR, Hatzivassiloglou V 1996 Translating Collocations for Bilingual Lexicons: A Statistical Approach. *Computational Linguistics*, 22/1: 1-38.
Stubbs M 1995 Collocations and Semantic Profiles: On the Cause of the Trouble with Quantitative Methods. *Functions of Language*, 2/1: 1-33.
Wolff JG 1991 *Towards a Theory of Cognition and Computing*. Chichester: Ellis Horwood.