Building a Portuguese Oenological Dictionary: from Corpus to Terminology via Co-occurrence Networks

***William Martinez*[1], *Sílvia Barbosa*[2]**
[1]*Université Sorbonne Nouvelle Paris 3,* [2]*NOVA FCSH - CLUNL*
E-mail: wmartinez68@gmail.com, silviabarbosa@fcsh.unl.pt

Abstract

This paper focuses on the elaboration of a dictionary of terms in the Portuguese language which describe the wine-tasting experience. We present a corpus-based analysis aimed at designing an electronic dictionary: on the basis of a compilation of approximately 21,000 wine descriptions downloaded from a dozen Portuguese websites, we estimated both by frequency analysis and lexicographical study which terms were recurrent, relevant and representative of the "hard to put into words" occupation that is oenology. From the results thus obtained, a list was made of words that describe the sensory analysis in its three main aspects: visual, olfactive and gustatory. An exhaustive co-occurrence analysis then identified those terms which contribute most to structuring the text by way of their tendency to attract other words against statistical odds. When displayed in a co-occurrence network, these anchors emerge from the mesh as the foundational lexicon for wine tasting, and can be evaluated as prime candidates for a distributional thesaurus.

Keywords: collocations, co-occurrences, word network, corpus linguistics, oenology, terminology

1 Objectives

Extracting relevant information from text in order to establish lexicographical lists of domain-specific terms, or as Kilgarriff (2005) says 'putting the corpus into the dictionary' is a complex operation. In the field of Information Science, the hierarchy known as the DIKW pyramid (from data to information to knowledge to wisdom) describes the processing chain that transforms dispersed raw facts into organized synthesized information. In this process, the most likely transition where value can be lost lies between data and information, because this transition is essentially achieved by inference: interrogative questions like *what?, when?* or *how*? are answered by data invested locally with meaning for a purpose. Typically, this data is selected in context, and relevant keywords are extracted by combining two things: a large volume of domain-related text and one or more domain experts, thus aiming for the highest representativity and the greatest degree of precision.

We believe that while higher levels of the DIKW pyramid require human cognition and judgment to reach understanding, the initial stage of information gathering can be greatly improved by a statistical approach to text. Indeed, an exhaustive analysis of all operating contexts can provide exact statistics regarding repeated word coincidence, which in turn can help describe word usage. According to the principle set out by J. R. Firth (1957) – "You shall know a word by the company it keeps" – co-occurrence analysis provides a contextually validated description of operative vocabulary for a given domain.

At the heart of the distributional hypothesis (Harris 1968) is the belief that the small number of realized combinations between words in context compared to the huge theoretical combinatorics that are possible between these elements must be indicative of strong linguistic relations at work. Whether

semantically or syntactically, words are bound and operate in a systematic manner that can be measured by their organized coincidences in context.[1]

2 Building a corpus

Cellars provide a time-tested environment for the preservation of wines. Temperature, humidity and light are carefully monitored and maintained for the optimal long-term aging. However, when it comes to harvesting wine vocabulary it is best to aim for quantity, diversity and topicality. This all-around representativity is achieved by collecting domain-specific information, where it is massively and archived in huge amounts: on the internet.

The corpus consisted of texts referred to as the *notas de prova* (wine tasting note) – a succinct text – presented in specific sections of newspapers or on producer's web pages. Those notes prototypically represent the descriptions of wines made by experts to describe the organoleptic sensations of each wine (color, aroma, type of fruit, degree of complexity, texture, final persistence, etc.) and evaluate the drink, using the descriptors available in the related repertoire, like a guide for those who read it.

By selecting a set of major websites relative to the Portuguese wine industry (with 1,480 notes of Portuguese producers and 20,011 notes collected from Portuguese online specialized wine magazines from several producers/wine companies) we were able to compile a set of over 21,000 wine tasting notes with 589,498 word tokens for 7,652 word types, among which there were 2,815 hapax legomena (Table 1).

Table 1: Corpus characteristics.

Tokens	589 498
Types	7 652
Hapax	2 815
Maximal frequency	40 548 (most frequent type: e)

Upon browsing the word-type dictionary extracted from the corpus (Table 2), it is interesting to note that there are very few tool-words among the most frequent types appearing in the text. The originality of layout in the frequency dictionary may find its explanation in the particular format of the text. Indeed, tasting notes are most often short and concise reports containing just one to three sentences.[2] Such a short text requires fewer tool words, which explains their absence at the top of the dictionary: connectors such as prepositions and conjunctions, anaphoric pronouns, etc.

Table 2: Most frequent word-types (frequency >=5000 occurrences).

Rank	Type	Freq.		Rank	Type	Freq.
1	*e*	40 548		11	*bem*	7 377
2	*de*	24 747		12	*final*	7 013
3	*com*	23 712		13	*acidez*	7 006
4	*muito*	14 318		14	*fruta*	6 726
5	*boca*	11 815		15	*fruto*	6 494
6	*na*	11 091		16	*no*	5 429
7	*a*	10 887		17	*taninos*	5 401
8	*um*	10 315		18	*boa*	5 327
9	*aroma*	7 993		19	*mas*	5 315
10	*notas*	7 816				

1 The possible relationship between distributional and semantic similarities has been exploited for the generation of automatic thesauri in previous works, notably Lin 1998 and Curran & Moens 2002.

2 Based on three declared signs of punctuation (*.?!*), 39 939 sentences are identified in the corpus with an average length of 14.76 words for a standard deviation of 5.10.

3 Type frequency vs type valency

Word frequency is commonly used as a topic indicator because it is an obvious and efficient measure of themes developed in a corpus: among the most frequent word-types of a text are to be found the recurrent nouns and verbs[3] (and hopefully the subjects and predicates of the topic being discussed). However, as bricks are mixed with mortar, hierarchization by frequency mixes content words with tool words among the most frequently occurring types. Indeed, a text follows an inevitable organization of words according to their frequency. Word frequencies are conditioned by Zipf's Law (Zipf 1965) whereby a small number of words have a very high frequency and a large number of words have a very low frequency.

From this predictable structure, Luhn (1958) – whilst working on automatic summarization – derived a method for locating valuable information-loaded words in between too frequent and too rare elements in the dictionary. Contrary to the structure proposed by Luhn, as described by Deghani (2016), the very frequent word types (function words) and the very rare words (hapax) contain few if any information-loaded words. As consequence, automatic cut-off limits can usually be defined to isolate significant words based only on their frequency. Clearly, in the case of our corpus this filter cannot be automated. Moreover, many experiments prove that dictionary hierarchy is no measure of significance in context.[4]

Another strong indicator of lexical behavior in context is lexical valency[5] i.e. the propensity for a given keyword to attract other words in context. Typically, word valency is measured according to the number of co-occurrences[6] which are detected around a given keyword in a defined context. For example, in our corpus, around the 11,815 occurrences of the word *boca* (mouth) and within the limits of each of the 11,688 sentences (222 540 tokens or 37% of the corpus) where the word occurs, we tallied every occurrence of every other word appearing in these contexts.

Beyond the mere co-frequencies of these words alongside the keyword *boca*, the results we obtained can be normalized to take into account the volume of the corpus, the volume of the sample (all phrasal contexts of *boca*) and the global frequencies of the co-occurring words. These four parameters (noted T, t, F, f for global and local text volumes and frequencies) are the input data for a great number of statistical models. Our choice of method for calculating a probabilistic score is the Hypergeometric Model[7], because of its easy-to-read result which measures the degree of surprise when confronting

3 Another assumption is that these inevitable words are evenly distributed in the corpus even if overall high frequency is sometimes due to particularly loaded sections of a corpus.

4 More recently the suitability of word frequency as a criterion for vocabulary selection has been questioned in language teaching by Okamoto (2015).

5 Valency is a notion borrowed from chemistry, where it denominates the combining power or affinity of an element, especially as measured by the number of hydrogen atoms it can displace or combine with, all depending on the electrons present in the outermost shell.

6 Unlike the term 'collocation', which implies a number of two adjacent collocates, 'co-occurrence' alludes to attractions between words in a broad sense, without imposing constraints of contiguity, orientation or distance. As a consequence, the phenomena detected are numerous and varied, thus reflecting the richness of lexical activity in the corpus.

7 The Hypergeometric Model determines the probability for an observed word frequency (x occurrences of word w in the vicinity of keyword k) based on four parameters:

T : number of tokens in the corpus
t : number of tokens in keyword contexts
F : frequency of co-occurring word in corpus
f : frequency of co-occurring word in in keyword contexts

$$P[X = f] = \frac{\binom{F}{f}\binom{T-F}{t-f}}{\binom{T}{t}}$$

A numerical specificity summarizes the deviation between the theoretical value and observed value, which can be null, positive or negative. If the observed frequency is more or less what is expected in theory, then there is no specificity. If the observed frequency is higher than expected, then the specificity is positive. Inversely, if the observed frequency is lower than expected, then the specificity is negative. The value indicates the degree of probability of the occurrence, for example: +3 indicates a positive specificity (more occurrences than expected) and a likelihood of 1/1000 (3 zeroes). A negative specificity of -10 would indicate a negative co-occurrence between words (less coincidences in context than expected) with a probability of 1/10000000000 (10 zeroes).

the observed co-frequency with the expected co-frequency of a given word. From a pragmatic point a view, this model, albeit complex, yields a result that is very easy to interpret (also beyond the binary co-occurrence / no co-occurrence it sometimes indicates a negative result, which signals anti-co-occurrence or repulsion between words).

Our list of co-occurrences is then re-ordered according to this probabilistic measure. We counted a total of 4,538 statistically specific co-occurrences, of which 2,259 were positive (words over-represented around *boca*) and 2,279 negative (under-represented). In Table 3 an excerpt from our results shows that the strongest co-occurrences reach very high degrees of statistical unlikelihood. For example, the number of encounters between the keyword 'mouth' and co-occurring terms 'acidity', 'tannins', 'body' and 'soft' are given a specificity of +100. This means that the odds of these coincidences are 1 over 1 plus 100 zeroes (1^{-100}), so very unlikely to occur in context and thus worthy of our interest.

Table 3: Main positive and negative co-occurrences around *boca* (mouth).

Rank	Type	Positive		Rank	Type	Negative
1	*acidez* (acidity)	+100		1	*fruta* (fruit)	-100
2	*final* (end)	+100		2	*aroma*	-100
3	*taninos* (tannins)	+100		3	*madura* (mature)	-100
4	*corpo* (body)	+100		4	*frutos* (fruits)	-100
5	*bom* (good)	+100		5	*especiarias* (spices)	-100
6	*macio* (soft)	+100		6	*cor* (color)	-100
7	*volume*	+100		7	*minerais* (minerals)	-100
8	*redondo* (round)	+82		8	*flores* (flowers)	-100
9	*frescura* (freshness)	+68		9	*preta* (black)	-100
10	*doçura* (sweetness)	+63		10	*vermelha* (red)	-87
11	*estrutura* (structure)	+62		11	*floral*	-86
12	*sabor* (taste, flavor)	+60		12	*citrinos* (citruses)	-85
13	*textura* (texture)	+52		13	*vegetais* (vegetables)	-84
14	*secura* (dryness)	+48		14	*nariz* (nose)	-73
15	*viva* (bright)	+47		15	*barrica* (barrel)	-72
16	*longo* (long)	+46		16	*tostados* (toasted)	-70
17	*mediano* (average)	+46		17	*aromática* (aromatic)	-65
18	*equilíbrio* (balance)	+40		18	*silvestres* (wild, sylvan)	-64
19	*sabroso* (tasty)	+35		19	*folhas* (leaves)	-60
20	*cheia* (full)	+32		20	*chocolate*	-57

With these results we are able to build an understanding of the buccal experience in wine tasting: acidity, tannins, body, soft, round, freshness and sweetness are part of the tasting experience. Inversely, other words are given negative specificities to indicate their absence in the contexts of 'mouth': fruit, mature, spices , flowers, citruses and vegetables or chocolate for example do not come to the mind of wine tasters when describing their 'in mouth' appreciation.

Regarding these probability scores, it should be noted that they reward high degrees of coincidence in context, regardless of individual word frequency. This means that low-frequency terms can be promoted to the top of our statistical ranking dictionary to provide an order of importance quite different to that in our initial frequency dictionary. The results in Table 4 show how words ranked according to their valency reveal an unexpected order. From a linguistic perspective, keywords with a high valency can be interpreted as elements of disruption. Indeed, as statistical measures point out, whenever a high-valency word appears in context it seems to trigger the appearance of one or several other words which, according to the laws of probability, were not expected to arrive at this moment in the flow

of text. Therefore, we can consider high-valency words to play a particular role in context as they contribute strongly to structuring the entire text.

Table 4: Highest valency types.

Rank	Type	Valency		Rank	Type	Valency
1	*fruto*	49		11	*longo*	30
2	*notas*	41		12	*corpo*	30
3	*frutos*	39		13	*acidez*	29
4	*cor*	38		14	*final*	28
5	*leve*	37		15	*fruta*	27
6	*taninos*	36		16	*médio*	25
7	*especiarias*	35		17	*boca*	24
8	*preta*	33		18	*vermelha*	24
9	*citrinos*	33		19	*flores*	23
10	*aroma*	31		20	*encorpado*	23

To fully appreciate this lexical dynamic, we extend co-occurrence to all word-types[8] with a frequency >=10, which yields a table of 2,170 x 2,170 words (excerpt in Table 5) identifying all co-occurring types. This huge table requires some form of filtering to reduce its size to the essential statistical data; for example: words retained after analysis should co-occur at least *x* times at no greater distance than *y* words. A more stringent criterion would be mutual co-occurrence, whereby *A* is in co-occurrence with *B* if and only if *B* is in co-occurrence with *A* (indeed depending on the method for measuring, a co-occurrence relation may not be reciprocal[9]).

Table 5: Table of co-occurrences (excerpt).

Keyword	Co-oc. 1	Co-oc. 2	Co-oc. 3	Co-oc. 4	Co-oc. 5
boca	*prova*	*final*	*na*	*acidez*	*fácil*
na	*final*	*boca*	*acidez*	*firme*	*bom*
acidez	*final*	*na*	*boca*	*viva*	*cremoso*
final	*na*	*boca*	*acidez*	*longo*	*boa*
taninos	*final*	*na*	*boca*	*firme*	*cheio*
corpo	*prova*	*final*	*boca*	*acidez*	*viva*
prova	*boca*	*uma*	*fácil*	*corpo*	*ampla*
bom	*na*	*acidez*	*conjunto*	*nervo*	*corpo*
macio	*prova*	*e*	*acidez*	*fácil*	*corpo*
volume	*bom*				
tanino	*algum*	*maduro*	*alguma*	*meio*	*secura*

The description provided on the Table 5 is total: all statistically remarkable encounters in context are recorded. Because it describes all relations of attraction between words in the corpus, this 'adjacency matrix' is akin to a lexical mesh that supports the entire corpus. Every single binary co-occurrence contributes to building a complex network and forming, link by link, the lexical backbone of the text. However, once this information is ordered in tabular form, the problem is to interpret it, or at least read it. This is why, given their density, such matrices are usually visualized in the form of a co-occurrence network.

8 Textual statistics requires a minimal number of phenomena to rule on the over or under-representation of words in a given context. For this reason, hapax legomena and low frequency words are typically excluded from analysis. A common threshold would be F>=5 for an average sized corpus or F>=10 or higher for bigger textual compilations.

9 With the TtFf parameters for the Hypergeometric Model, measuring co-occurrence from *A* to *B* can provide a different score than from B to A.

4 Co-occurrence networks

Because of their logic of construction (in context, every word co-occurs – directly or not – with every other word) co-occurrence graphs can build up exponentially and produce unreadable results. Choices must therefore be made with a view to filtering out less important word attractions: minimal frequency and minimal specificity are the basic filters available. Even at co-frequency and specificity thresholds of 50 and 25, a total of 25 networks are detected in the corpus. [10] Most of these have two or three components: [*concentrado*, *rico*] (concentrated, rich), [*como*, *aperitivo*] (as, aperitif), [*lote*, *castas*] (lot, varieties), [*sempre*, *presente*] (always, present), [*são*, *os*] (are, the), [*sauvignon*, *blanc*, *cabernet*], [*framboesa*, *groselha*, *morango*] (raspberry, currant, strawberry), [*muita*, *bela*, *frescura*] (very, beautiful, freshness). Others, more elaborate describe semantic fields: [*região*, *apesar*, *marca*, *da*, *as*, *onde*, *aqui*] (region, despite, brand, of the, the, where, here).

Of the 25 networks detected in our corpus, one represents the major structure in a text with over 230 lexical components. Despite the strict thresholds for network extraction, the graph hereafter (Figure 1) is huge and as a result does not lend itself to close-up analysis but rather calls for observation from a distance. What should be noted in the following network is the gross organization of nodes: some are strongly connected to the graph by several links, others are attached by one relation. This topological view helps identify nodes of great importance whose absence from the graph would considerably alter its structure whilst others could be removed from the graph without any consequence. From a linguistic perspective, these opposite profiles correspond to words that either have enormous consequences on their contexts whenever they appear, or very little influence on their surroundings. Some words trigger a cohort of co-occurring words, others entail little, if anything, in context.

On closer inspection of the graph, the salience of over a dozen nodes corresponds to the identification by statistics of the essential vocabulary in wine tasting. Some words are outstandingly magnetic and form constellations around themselves: *taninos*, *fruta*, *fruto*, *frutos*, *leve*, *notas*, *cor*, *boa*, *corpo*, *longo*, *boca*, *final*, *acidez* and *preta* (*tanins, fruit(s), light, notes, color, good, body, long, mouth, end, acidity and black*). By altering the statistical thresholds, the graph can be reduced to under 150 word-types, which makes the figure easier to read close-up (see appendix, Figure 2).

This first glance at the figure shows that from a topological angle the underlying structure of the co-occurrence network is clear and meaningful: co-occurrence activity is uneven among the lexical components of our corpus. Some essential words contribute fundamentally to structuring the text. How does this contextual prominence translate to lexical importance?

5 From map to dictionary

Once the co-occurrence network is extracted from the corpus, what (type of) information does it provide?

Reading the graph in Figure 2 is an unpredictable experience. Any node can be a point of entry into the mesh. Any group of words can be read to form a meaningful set in one's mind which corresponds – or not – to an attested sequence in context. Here lies the complexity of the virtual network: all associations are presented simultaneously. Whilst multi-word expressions, grammatical constructions and all kinds of dependencies are suggested in the graph, only some really exist in the corpus. Consider the aforementioned example [*muita*, *bela*, *frescura*] (very, beautiful, freshness). Several constructs

10 This implies that only words which coincide in context at least 50 times with a hypergeometric specificity of +25 will qualify for the network. Depending on the volume of the corpus, only words with the strongest power of co-occurrence should appear on the graph.

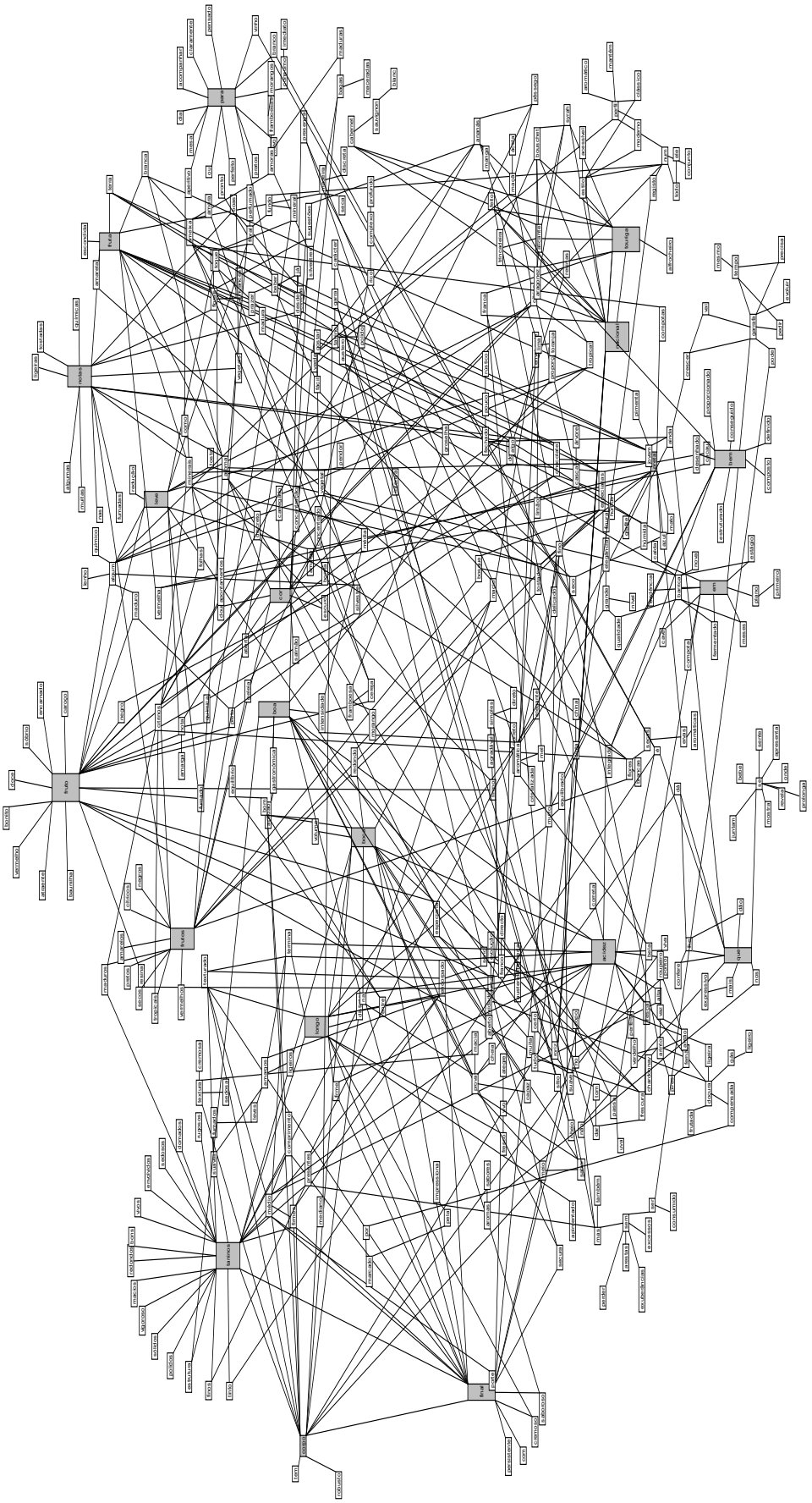

Figure 1: Abstract view of a co-occurrence network graph.

are possible in theory ("*muita bela*", "*muita bela frescura*", "*frescura muita bela*") yet this network actualizes in the corpus in only two forms: "*muita frescura*" 222 occ. and "*bela frescura*" 71 occ.

Is this misleading? Not if one accepts that the graph shows words according to their statistical importance. If we bear this in mind, our interpretation of the network is cautious: all co-occurrences are significant but not necessarily meaningful. Working on the map as a foundation, a lexicographer is shown the entire backbone of the *corpus*. In this hierarchized schema, he can consciously choose which nodes to investigate and which to ignore.

The relation between an object and its representation is often described by the semiotician Alfred Korzybski's famous words "the map is not the territory", which he extended to a more domain-specific "the word is not the thing". The general idea is that perception always intercedes between the observed and observer. This inevitable distortion rears its head in any human-based enterprise, including lexicography.

Let's consider WordNet (Miller 1991), a (mostly) handwritten lexical database that was started by psycholinguists in the 1980s and is now emulated in many different languages. Its 1,7000 synonymy sets are interlinked by conceptual relations all determined by lexicographers. As Maziarz (2013) points out, these synsets are *de facto* the building blocks of the thesaurus, not words. Thus, pre-imposed synonymy becomes the norm and poses a problem of circularity in database construction and maintenance: WordNet presents words as synonyms because someone upstream deemed them to be.

When we set out to build a Portuguese terminology for wine we wanted to avoid, as much as possible, all initially built-in flaws. Therefore, our main objective was to implement an unsupervised method and apply it to a representative corpus. We consider the extracted data – filtered by strict statistical thresholds – to be exhaustive and objective, thus presenting a representative view of lexical phenomena in our wine-tasting compilation.

However, while typical problems do exist in our database they appear under a new light. Take circularity, for example. Whatever the thresholds we set for contextual exploration, small subsets of words mutually defining each other in very closed systems continue to emerge from co-occurrence network analysis. Yet, when these isolates appear beside major word networks, we are able to visually appreciate their importance vis-à-vis the main co-occurrence structures. Indeed, in network theory, these cliques are a well-documented phenomena, and their integration to the general structure is a matter of graph algebra not a decision made by the lexicographer.

Our main preoccupation is to preserve, as honestly as possible, the structure of our corpus, both syntactically and semantically. Indeed, as Korzybski underlined, the single most important item of information is the structure: a map *is not* the territory it represents, but, if correct, it has a *similar structure* to the territory, which accounts for its usefulness – Korzybski (1933).

After accumulating knowledge produced by total co-occurrence analysis for all word-types in the corpus provides, we consider this to be a statistically validated basis for the production of a distributional thesaurus. Indeed, distributed representations of words learned from text have proven to be successful in various Natural Language Processing tasks, such as word sense disambiguation, information retrieval or document summarization. In recent applications, distributional similarity has successfully been exploited as an approximation to semantic similarity. Kilgarriff and Rychly (2007) present an automatically produced thesaurus which identifies words which occur in similar contexts as the keyword, and draws on the hypothesis of distributional semantics. Ziai *et al.* (2016) use distributional semantics to support qualitative insights into the data and identify phenomena at the lexical level. Notably, Maziarz *et al.* (2013) recenter their Polish WordNet on lexical units in order to automatically construct synsets out of words with similar connectivity.

6 Conclusion

In this paper we have shown how co-occurrence analysis can be applied in the process of information retrieval for deriving lexicons from a domain-specific corpus. The results of generalized co-occurrence analysis for all word-types show circumscribed lexical systems that operate in the text. These structures display semantic homogeneity and can be interpreted as sense clusters, where linked words all serve to complete and precise their meaning in context.

After looking at a selection of 21,000 wine tasting notes, our experiment made it possible to extract from a word-type dictionary of over 7,500 terms a list of 300 words with high co-occurrence activity and establish evidence for repeated meaning building in context by association or dissociation of words as measured by positive and negative co-occurrence. Where typical distributional *thesauri* identify only words that occur in similar contexts to the keyword to posit their synonymy, our co-occurrence network provides data for the production of a more precise thesaurus. Following Greffenstette (1994), the analysis of second order co-occurrences (co-occurrences of co-occurrences) can identify sets of synonyms (words that appear in the same type of context as the keyword). An extension of this logic would be to exploit negative co-occurrences for a given keyword and detect their second order co-occurrences so as to build sets of antonyms within a domain vocabulary.

Future work involves expanding the corpus to provide a larger image of the list of words and co-occurrences used to describe the main aspects: visual, olfactive and gustatory, in order to understand, in more detail, the apparent non-complex verbalization of the wine-tasting experience.

References

Abhik J. & Pawan G. (2018). Can Network Embedding of Distributional Thesaurus be Combined with Word Vectors for Better Representation? In *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2018).* New Orleans. June 2018.

Čermák F. (2006). Collocations, Collocability and Dictionary. In *Proceedings of the 12th Euralex International Congress*, 2006. Turin, publisher Edizioni dell'Orso, pp 929-937, 2006.

Curran J. & Moens M. (2002). Improvements in Automatic Thesaurus Extraction. In Proceedings of the Workshop of the ACL Special Interest Group on the Lexicon (SIGLEX): *Unsupervised Lexical Acquisition*. pp 59-66. 2002.

Deghani M. et al. (2016). Luhn Revisited. Significant Words Language Models. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management.* Indiana. pp 1301-1310, 2016.

Firth, J. R. (1957). A synopsis of linguistic theory 1930-1955. In *Special Volume of the Philological Society*. Oxford, Oxford University Press. 1957.

Grefenstette G. (1994). Corpus-derived first-, second- and third-order word affinities. In *Proceedings of Euralex*, pp 279–290, Amsterdam, Holland. 1994.

Guthrie D., Allison B., Liu W., Guthrie L., Wilks Y. (2006). A closer look at skip-gram modelling, In *Proceedings of the 5th international Conference on Language Resources and Evaluation (LREC-2006).* Genoa, Italy, 2006.

Harris Z. S. (1968). Mathematical Structures of Language. John Wiley, New York. 1968.

Kilgarriff A. (2005). Putting the Corpus into the Dictionary, In *Proceedings of Meaning Workshop*, Trento.

Kilgarriff A. & Rychly P. (2007). An efficient algorithm for building a distributional thesaurus (and other Sketch Engine developments). In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, Prague. June 2007, pp. 41- 44. 2007.

Kim W., Wilbur J. (2000). Corpus-based Statistical Screening for Phrase Identification, In *Journal of the American Medical Informatics Association*, Volume 7 Number 5 Sep / Oct 2000.

Korzybski A. (1995). Science and Sanity: an introduction to non-Aristotelian systems and general semantics, Institute of General Semantics; 5th edition, (1st edition 1933) 1995.

Lin D. (1998). Automatic Retrieval and Clustering of Similar Words. In *Proceedings of the 17th International Conference on Computational Linguistics (COLING98).* pp 768-774.Montreal, Canada. 1998.

Luhn H. P. (1958). *A Business Intelligence System*. IBM Journal. October 1958.

Luísa & Pereira A., Santos & Mendes, A. (2002). An Electronic Dictionary of Collocations for European Portuguese: Methodology, Results and Applications.

Marziarz, M., Piasecki M., Szpakowicz S. (2013). The Chicken and Egg Problem in WordNet Design: Synonymy, Synsets and Constitutive Relations, In *Language Resources and Evaluations*. Volume 47, Issue 3, September 2013.

Miller, G., Beckwith R., Fellbaum, C., Gross D., Miller K. (1991). Introduction to WordNet: An On-line Lexical Database, In *International Journal of Lexicography*. Volume 3, January 1991.

Okamoto M. (2015). Is corpus word frequency a good yardstick for selecting words to teach? Threshold levels for vocabulary selection. *System*. Volume 51, July 2015.

Periñán-Pascual C. (2015). The underpinnings of a composite measure for automatic term extraction. The case of SRC. In *Terminology across Languages and Domains*. Special Issue of Terminology. Volume 21. Issue 2. Edited by Drouin P., Grabar N., Hamon T. & Kaguera K. pp 151-179. 2015.

Sinclair J. (1991). *Corpus, Concordance, Collocation*, OUP, Oxford. 1991.

Tsegaye R., Wartena C., Drumond L. & Schmidt-Thieme L. (2016). Learning Thesaurus Relations from Distributional Features. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC 2016). pp 442–446, Portorož, Slovenia. 2016.

Verdaguer I. & González E. (2004). A lexical database of collocations in scientific English, In *Proceedings of the 11th Euralex International Congress*, 2004.

Wanner L., Bohnet B., Giereth M., (2006). What is beyond collocations? Insights from Machine Learning experiments, In *Proceedings of the 12th Euralex International Congress*, 2006, Turin, publisher Edizioni dell'Orso.

Wanner L., Bohnet B., Giereth M., (2006). Making sense of collocations. *Computer Speech & Language*, volume 20, number 4, pp. 609-624. 2006.

Weeds J. & Weir D. (2005). Co-occurrence Retrieval: A Flexible Framework for Lexical Distributional Similarity. N *Journal of Computational Linguistics*. Volume 31. Issue 4. December 2005. MIT Cambridge, MA, USA. 2005.

Ziai R., De Kuthy K. & Meurers D. (2016). Approximating Givenness in Content Assessment through Distributional Semantics. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016).* pp 209–218, Portorož, Slovenia. 2016.

Zipf G. K. (1965). The Psycho-Biology of Language. An Introduction to Dynamic Philology. The MIT Press. Cambridge, MA, USA. 1965.

Acknowledgements

The second author gratefully acknowledges the financial support of the Fundação para a Ciência e Tecnologia PhD grant (PD/BD/52261/2013) and NOVA FCSH - CLUNL for this research.

Appendix 1

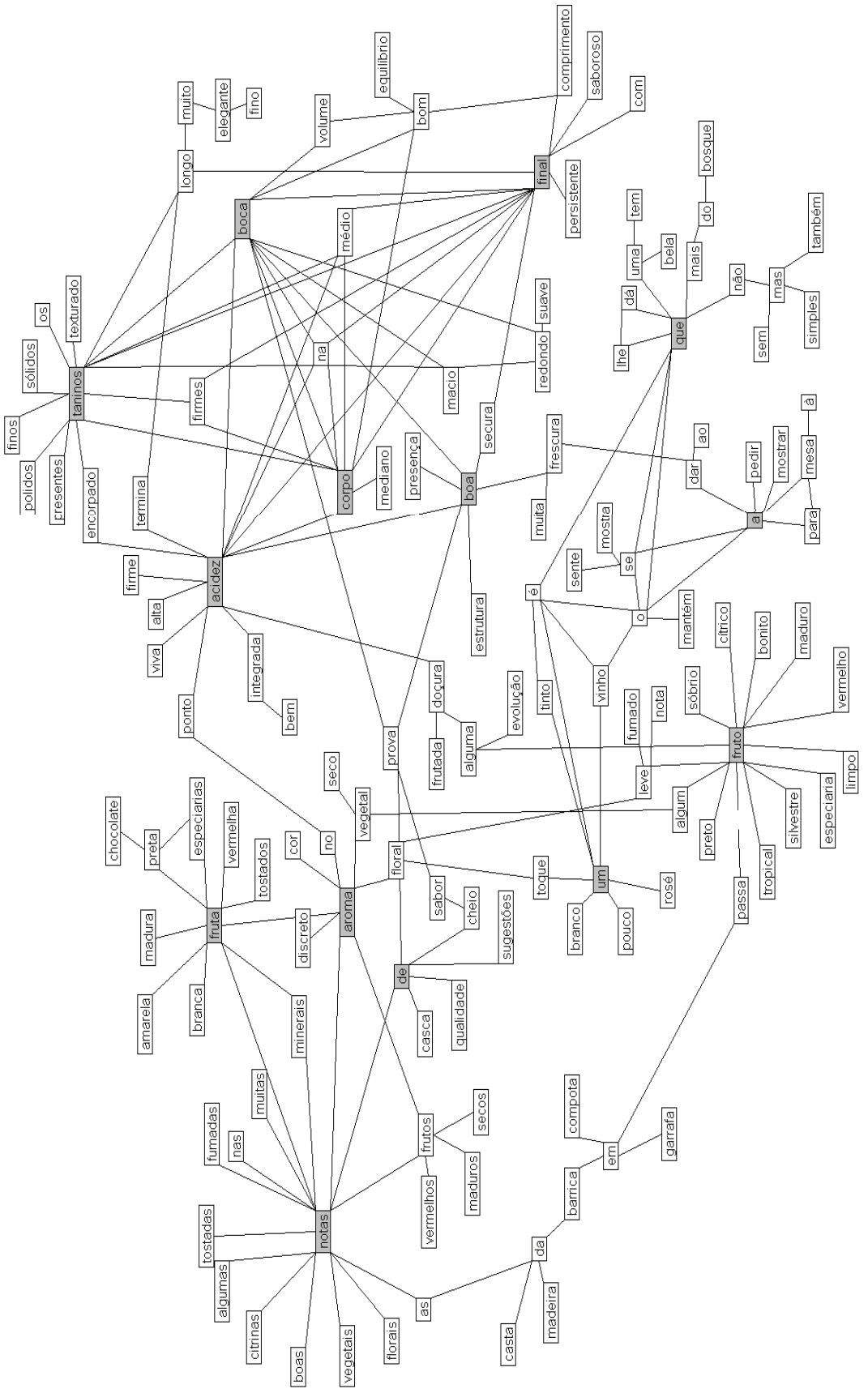

Figure 2: Graph of co-occurrence network: 147 nodes. Thresholds: min. co-frequency 200, min. specificity +50.