Co-dispersion: A Windowless Approach to Lexical Association

Justin Washtell
University of Leeds
Leeds, UK
washtell@comp.leeds.ac.uk

Abstract

We introduce an alternative approach to extracting word pair associations from corpora, based purely on surface distances in the text. We contrast it with the prevailing window-based co-occurrence model and show it to be more statistically robust and to disclose a broader selection of significant associative relationships - owing largely to the property of scale-independence. In the process we provide insights into the limiting characteristics of window-based methods which complement the sometimes conflicting application-oriented literature in this area.

1 Introduction

The principle of using statistical measures of co-occurrence from corpora as a proxy for word association - by comparing observed frequencies of co-occurrence with expected frequencies - is relatively young. One of the most well known computational studies is that of Church & Hanks (1989). The method by which co-occurrences are counted, now as then, is based on a device which dates back at least to Weaver (1949): the context window. While variations on the specific notion of context have been explored (separation of content and function words, asymmetrical and non-contiguous contexts, the sentence or the document as context) and increasingly sophisticated association measures have been proposed (see Evert, 2007, for a thorough review) the basic principle – that of counting token frequencies within a context region – remains ubiquitous.

Herein we discuss some of the intrinsic limitations of this approach, as are being felt in recent research, and present a principled solution which does not rely on co-occurrence windows at all, but instead on measurements of the surface distance between words.

2 The impact of window size

The issue of how to determine appropriate window size (and shape) has often been glossed over in the literature, with such parameters being determined arbitrarily, or empirically on a per-application basis, and often receiving little more than a cursory mention under the description of method. For reasons that we will discuss however, the issue has been receiving increasing attention. Some have attempted to address it intrinsically (Sahlgren 2006; Schulte im Walde & Melinger, 2008; Hung et al, 2001); others no less earnestly in the interests of specific applications (Lamjiri, 2003; Edmonds, 1997; Wang 2005; Choueka & Lusignan, 1985) (note that this divide is sometimes subtle).

The 2008 Workshop on Distributional Lexical Semantics, held in conjunction with the European Summer School on Logic, Language and Learning (ESSLLI) – hereafter the ESSLLI Workshop - saw this issue (along with other "problem" parameters in distributional lexical semantics) as one of its central themes, and witnessed many different takes upon it. Interestingly, there was little consensus, with some studies appearing on the surface to starkly contradict one-another. It is now generally recognized that window size is, like the choice of corpus or specific association measure, a parameter which can have a potentially profound impact upon the performance of applications which aim to exploit co-occurrence counts.

One widely held (and upheld) intuition - expressed throughout the literature, and echoed by various presenters at the ESSLLI Workshop - is that whereas small windows are well suited to the detection of syntactico-semantic associations, larger windows have the capacity to detect broader "topical" associations. More specifically, we can observe that small windows are unavoidably limited to detecting associations manifest at very close distances in the text. For example, a

Proceedings of the 12th Conference of the European Chapter of the ACL, pages 861–869,
Athens, Greece, 30 March – 3 April 2009.

window size of two words can only ever observe bigrams, and cannot detect associations resulting from larger constructs, however ingrained in the language (e.g. "if … then", "ne … pas", "dear ... yours"). This is not the full story however. As, Rapp (2002) observes, choosing a window size involves making a *trade-off* between various qualities. So conversely for example, frequency counts within large windows, though able to detect longer-range associations, are not readily able to distinguish them from bigram style co-occurrences, and so some discriminatory power, and sensitivity to the latter, is lost. Rapp (2002) calls this trade-off "specificity"; equivalent observations were made by Church & Hanks (1989) and Church *et al* (1991), who refer to the tendency for large windows to "wash out", "smear" or "defocus" those associations exhibited at smaller scales.

In the following two sections, we present two important and scarcely discussed facets of this general trade-off related to window size: that of *scale-dependence*, and that concerning the specific way in which the *data sparseness problem* is manifest.

2.1 Scale-dependence

It has been shown that varying the size of the context considered for a word can impact upon the performance of applications (Rapp, 2002; Yarowsky & Florian, 2002), there being no ideal window size for all applications. This is an inescapable symptom of the fact that varying window size fundamentally affects *what* is being measured (both in the raw data sense and linguistically speaking) and so impacts upon the output qualitatively. As Church *et al* (1991) postulated, *"It is probably necessary that the lexicographer adjust the window size to match the scale of phenomena that he is interested in"*.

In the case of inferential lexical semantics, this puts strict limits on the interpretation of association scores derived from co-occurrence counts and, therefore, on higher-level features such as context vectors and similarity measures. As Wang (2005) eloquently observes, with respect to the application of word sense disambiguation, *"window size is an inherent parameter which is necessary for the observer to implement an observation … [the result] has no meaning if a window size does not accompany"*. More precisely, we can say that window-based co-occurrence counts (and any word-space models we may derive from them) are *scale-dependent*.

It follows that one cannot guarantee there to be an "ideal" window size within even a single application. Distributional lexical semantics often defers to human association norms for evaluation. Schulte im Walde & Melinger (2008) found that the correlation between co-occurrence derived association scores and human association norms were weakly dependent upon the window size used to calculate the former, but that certain associations tended to be represented at certain window sizes, by virtue of the fact that the best *overall* correlation was found by combining evidence from *all* window sizes. By identifying a single window size (whether arbitrary or apparently optimum) and treating other evidence as extraneous, it follows that studies may tend to distance their findings from one another.

As Church *et al* (1991) allude, in certain situations the ability to tune analysis to a specific scale in this way may be desirable (for example, when explicitly searching for statistically significant bigrams, only a 2-token window will do). In other scenarios however, especially where a trade-off in aspects of performance is found between scales, it can clearly be seen as a limitation. And after all, is Church *et al*'s notional lexicographer really interested in those features manifest at a specific scale, or is he interested in a specific *linguistic category* of features? Notwithstanding grammatical notions of scale (the clause, the sentence etc), there is as yet little evidence to suggest how the two are linked.

The existence of these trade-offs has led some authors towards creative solutions: looking for ways of varying window size dynamically in response to some performance measure, or simultaneously exploiting more than one window size in order to maximize the pertinent information captured (Wang, 2005; Quasthoff, 2007; Lamjiri *et a*l, 2003). When the scales at which an association is manifest are the quantity of interest and the subject of systematic study, we have what is known in scale-aware disciplines as *multi-scalar* analysis, of which fractal analysis is a variant. Although a certain amount has been written about the fractal or hierarchical nature of language, approaches to co-occurrence in lexical semantics remain almost exclusively mono-scalar, with the recent work of Quasthoff (2007) being a rare exception.

2.2 Data sparseness

Another facet of the general trade-off identified by Rapp (2002) pertains to how limitations in-

herent in the combination of data and co-occurrence retrieval method are manifest.

When applying a small window, the number of window positions which can be expected to contain a specific pair of words will tend to be low in comparison to the number of instances of each word type. In some cases, no co-occurrence may be observed at all between certain word pairs, and zero or negative association may be inferred (even though we might reasonably expect such co-occurrences to be feasible within the window, or know that a logical association exists). This is one manifestation of what is commonly referred to as the data sparseness problem, and was discussed by Rapp (2002) as a side-effect of *specificity*. It would of course be inaccurate to suggest that data sparseness itself is a response to window size; a larger window superficially lessens the sparseness problem by inviting more co-occurrences, but encounters the same underlying paucity of information in a different guise: as both the size and overlap between the windows grow, the available information is increasingly diluted both within and amongst the windows, resulting in an over-smoothing of the data. This phenomenon is well illustrated in the extreme case of a single corpus-sized window where - in the absence of any external information - observed and expected co-occurrence frequencies are equivalent, and it is not possible to infer any associations at all.

Addressing the sparseness problem with respect to corpus data has received considerable attention in recent years. It is usually tackled by applying explicit smoothing methods so as to allow the estimation of frequencies of unseen co-occurrences. This may involve applying insights on the statistical limitations of working from a finite sample (add-λ smoothing, Good-Turing smoothing), making inferences from words with similar co-occurrence patterns, or "backing off" to a more general language model based on individual word frequencies, or even another corpus; for example, Keller & Lapata (2003) use the Web. All of these approaches attempt to *mitigate* the data sparseness manifest in the observed co-occurrence frequencies; they do not presume to *reduce* data sparseness by improving the method of observation. Indeed, the general assumption would seem to be that the only way to minimize data sparseness is to use *more data*. However, we will show that, similarly to Wang's (2005) observation concerning windowed measurements in general, apparent data sparseness is as much a manifestation of the observation method as it is of the data itself; there may exist much pertinent information in the corpus which yet remains unexploited.

3 Proximity as association

Comprehensive multi-scalar analyses (such as applied by Quasthoff, 2007; and Schulte im Walde & Melinger, 2008) can be laborious and computationally expensive, and it is not yet clear how to derive simple association scores and suchlike from the dense data they generate (typically a separate set of statistics for each window size examined). There do exist however relatively efficient naturally *scale-independent* tools which are amenable to the detection of linguistically interesting features in text. In some domains the concept of *proximity* (or *distance* – we will use the terms somewhat interchangeably here) has been used as the basis for straightforward alternatives to various frequency-based measures. In biogeography, for example, the dispersion or "clumpiness" of a population of individuals can be accurately estimated by sampling the distances between them (Clark & Evans, 1954): a task more conventionally carried out by "quadrat" sampling, which is directly analogous to the window-based methods typically used to measure dispersion or co-occurrence in a corpus (see Gries, 2008, for an overview of dispersion in a linguistic setting). Such techniques are also been used in archeology. Washtell (2006) found evidence to suggest that distance-based approaches within the geographic domain can be both more accurate and more efficient than their window-based alternatives.

In the present domain, the notion of proximity has been applied by Savický & Hlavácová (2002) and Washtell (2007) - both in Gries (2008) - as an alternative to approaches based on corpus division, for quantifying the dispersion of words within the text. Hardcastle (2005) and Washtell (2007) apply this same concept to measuring word pair associations, the former via a somewhat ad-hoc approach, the latter through an extension of Clark-Evans (1954) dispersion metric to the concept of *co-dispersion*: the tendency of unlike words to gravitate (or be similarly dispersed) in the text. Terra & Clarke (2004) use a very similar approach in order to generate a probabilistic language model, where previously n-gram models have been used,

The allusion to *proximity* as a fundamental indicator of lexical association does in fact per-

meate the literature. Halliday (1966), for example, in Church *et al* (1991) talked not explicitly of frequencies within windows, but of identifying lexical associates via *"some measure of significant proximity, either a scale or at least a cut-off point"*. For one (possibly practical) reason or another, the "cut-off point" has been adopted and the intuition of proximity has since become entrained within a distinctly frequency-oriented model. By way of example, the notion of proximity has been somewhat more directly courted in some window-based studies through the use of "ramped" or "weighted" windows (Lamjiri et al, 2003; Bullinaria & Levy, 2007), in which co-occurrences appearing towards the extremities of the window are discounted in some way. As with window size however, the specific implementations and resultant performances of this approach have been inconsistent in the literature, with different profiles (even including those where words are discounted towards the *centre* of the window) seeming to prove optimum under varying experimental conditions (compare, for instance, Bullinaria, 2008, and Shaol & Westbury, 2008, from the ESSLLI Workshop).

Performance considerations aside, a problem arising from mixing the metaphors of frequency and distance in this way is that the resultant measures become difficult to interpret; in the present case of association, it is not trivially obvious how one might establish an expected value for a window with a given profile, or apply and interpret conditional probabilities and other well-understood association measures.[1] At the very least, Wang's (2005) observation is exacerbated.

3.1 Co-dispersion

By doing away with the notion of a window entirely and focusing purely upon distance information, Halliday's (1966) intuitions concerning proximity can be more naturally realized. Under the frequency regime, co-occurrence scores correspond directly to probabilities, which are well understood (providing, as Wang, 2005, observes, that a window size is specified as a reference-frame for their interpretation). It happens that similarly intuitive mechanics apply within a purely distance-oriented regime - a fact realised by Clark & Evans (1954), but not exploited by Hardcastle (2005). Co-dispersion, which is derived from the Clark-Evans metric (and more descriptively entitled "co-dispersion by nearest neighbour" - as there exist many ways to measure dispersion), can be generalised as follows:

$$CoDisp_{ab} = \frac{m \cdot n / (\max(freq_a, freq_b) + 1)}{M(dist_{ab_1}, \ldots, dist_{ab_n})}$$

Where, in the denominator, $dist_{abi}$ is the inter-word distance (the number of intervening tokens plus one) between the i^{th} occurrence of word-type *a* in the corpus, and the nearest preceding or following occurrence of word-type *b* (if one exists before encountering *(1)* another occurrence of *a* or *(2)* the edge of the containing document). *M* is the generalized mean. In the numerator, $freq_i$ is the total number of occurrences of word-type *i*, *n* is the number of tokens in the corpus, and *m* is a constant based on the expected value of the mean (e.g. for the arithmetic mean – as used by Clark & Evans - this is 0.5). Note that the implementation considered here does not distinguish word order; owing to this, and the constraint *(1)*, the measure is symmetric.[2]

Plainly put, co-dispersion calculates the ratio of the mean observed distance to the expected distance between word type pairs in the text; or how much closer the word types occur, on average, than would expected according to chance[3]. In this sense it is conceptually equivalent to Pointwise Mutual Information (PMI) and related association measures which are concerned with gauging how more *frequently* two words occur together (in a window), than would be expected by chance.

Like many of its frequency-oriented cousins, co-dispersion can be used directly as a measure of association, with values in the range $0 >= CoDisp <= \infty$ (with a value of 1 representing no discernible association); and as with these measures, the logarithm can be taken in order to present the values on a scale that more meaningfully represents relative associations (as is the default with PMI). Also as with PMI *et al*, co-dispersion can have a tendency to give inflated estimates where infrequent words are involved. To address this problem, a simple significance-

[1] Existing works do not go into detail on method, so it is possible that this is one source of discrepancies.

[2] This constraint, which was independently adopted by Terra & Clarke (2004), has significant computational advantages as it effectively limits the search distance for frequent words.

[3] The expected distance of an independent word-type pair is assumed to be half the distance between neighbouring occurrences of the more frequent word-type, were it uniformly distributed within the corpus.

corrected measure, more akin to a Z-Score or T-Score (Dennis, 1965; Church *et al*, 1991) can be formed by taking (the root of) the number of word-type occurrences into account (Sackett, 2001). The same principal can be applied to PMI, although in practice more precise significance measures such as Log-Likelihood are favoured.[4]

These similarities aside, co-dispersion has the somewhat abstract distinction of being effectively based on degrees rather than probabilities. Although it is windowless (and therefore, as we will show, scale-independent), it is not without analogous constraints. Just as the concept of mean frequency employed by co-occurrence requires a definition of distance (window size), the concept of distance employed by co-dispersion requires a definition of frequency. In the case presented here, this frequency is 1 (the nearest neighbour). Thus, whereas the assumption with co-occurrence is that the linguistically pertinent words are those that fall within a fixed-sized window of the word of interest, the assumption underpinning co-dispersion is that the relevant information lies (if at all) with the closest neighbouring occurrence of each word type. Among other things, this naturally favours the consideration of nearby function words, whereas (generally less frequent) content words are considered to be of potential relevance at some distance. That this may be a desirable property - or at least a workable constraint - is borne out by the fact that other studies have experienced success by treating these two broad classes of words with separately sized windows (Lamjiri *et al*, 2003).

4 Analyses

4.1 Scale-independence

Table 1 shows a matrix of agreement between word-pair association scores produced by co-occurrence and co-dispersion as applied to the unlemmatised, untagged, Brown Corpus. For co-occurrence, window sizes of ±1, ±3, ±10, ±32, and ±100 words were used (based on to a - somewhat arbitrary - scaling factor of $\sqrt{10}$).

The words used were a cross-section of stimulus-response pairs from human association experiments (Kiss *et al*, 1973), selected to give a uniform spread of association scores, as used in the ESSLLI Workshop shared task. It is not our purpose in the current work to demonstrate competitive correlations with human association norms (which is quite a specific research area) and we are making no cognitive claims here. Their use lends convenience and a (limited) degree of relevance, by allowing us to perform our comparison across a set of word-pairs which are deigned to represent a broad spread of associations according to some independent measure. Nonetheless, correlations with the association norms are presented as this was a straightforward step, and grounds the findings presented here in a more tangible context.

Because the human stimulus-response relationship is generally asymmetric (favouring cases where the stimulus word evokes the response word, but not necessarily vice-versa), the conditional probability of the response word was used, rather than PMI which is symmetric. For the windowless method, co-dispersion was adapted equivalently - by multiplying the resultant association score by the number of word pairings divided by the number of occurrences of the cue word. These association scores were also corrected for statistical significance, as per Sackett (2001). Both of these adjustments were found to improve correlations with human scores across the board, but neither impacts directly upon the comparative analyses performed herein. It is also worth mentioning that many human association reproduction experiments employ higher-order paradigmatic associations, whereas we use only syntagmatic associations.[5] This is appropriate as our focus here is on the information captured at the base level (from which higher order features – paradigmatic associations, semantic categories etc - are invariably derived). It can be seen in the rightmost column of table 1 that, despite the lack of sophistication in our approach, all window sizes and the windowless approach generated statistically significant (if somewhat less than state-of-the-art) correlations with the subset of human association norms used.

Owing to the relatively small size of the corpus, and the removal of stop-words, a large portion of the human stimulus-response pairs used as our basis generated no association (no smoothing was used as we are concerned at this level in raw evidence captured from the corpus). All correlations presented herein therefore consider only those word pairs for which there was *some evidence* under the methods being com-

[4] Although the heuristically derived MI^2 and MI^3 (Daille, 1994) have gained some popularity.

[5] Though interestingly, work done by Wettler *et al* (2005) suggests that paradigmatic associations may not be necessary for cognitive association models.

pared from which to generate a non-zero association score (however statistically insignificant). This number of word pairs, shown in square brackets in the leftmost column of table 1, naturally increases with window size, and is highest for the windowless methods.

Wind'd	±100	±32	±10	±3	±1	Human (*r*, *p*)	
±1 [34]	0.18	0.34	0.59	0.81	1	0.35	4.15%
±3 [47]	0.32	0.49	0.70	1		0.41	0.44%
±10 [64]	0.52	0.76	1			0.36	0.31%
±32 [78]	0.83	1				0.50	<0.01%
±100 [89]	1					0.53	<0.01%
W'less [103]	0.46	0.45	0.50	0.55	0.58	0.42	<0.01%

Table 1: Matrix of agreement (*corrected* r^2) between association retrieval methods; and correlations with sample association norms (r, and *p*-value).

The coefficients of determination (*corrected* r^2 values) in the main part of table 1 show clearly that, as window sizes diverge, their agreement over the apparent association of word pairs in the corpus diminishes - to the point where there is almost as much disagreement as there is agreement between windows whose size differs by a decimal order of magnitude. While relatively small, the fact that there remains a degree of information overlap between the smallest and largest windows in this study (18%), illustrates that some word pairs exhibit associative tendencies which markedly transcend scale. It would follow that single window sizes are particularly impotent where such features are of holistic interest.

The figures in the bottom row of table 1 show, in contrast, that there is a more-or-less constant level of agreement between the windowless and windowed approaches, *regardless* of the window size chosen for the latter.

Figure 1 gives a good two-dimensional schematic approximation of these various relationships (in the style of a Venn diagram). Analysis of partial correlations would give a more accurate picture, but is probably unnecessary in this case as the areas of overlap between methods are large enough to leave marginal room for misrepresentation. It is interesting to observe that co-dispersion *appears* to have a slightly higher affinity for the associations best detected by small windows in this case. Reassuringly nonetheless, the relative correlations with association norms here - and the fact that we see such significant overlap – do indeed suggest that co-dispersion is sensitive to useful information present in each of the various windowed methods. Note that the regions in Figure 1 necessarily have similar areas, as a correlation coefficient describes a symmetric relationship. The diagram therefore says nothing about the *amount* of information captured by each of these methods. It is this issue which we will look at next.

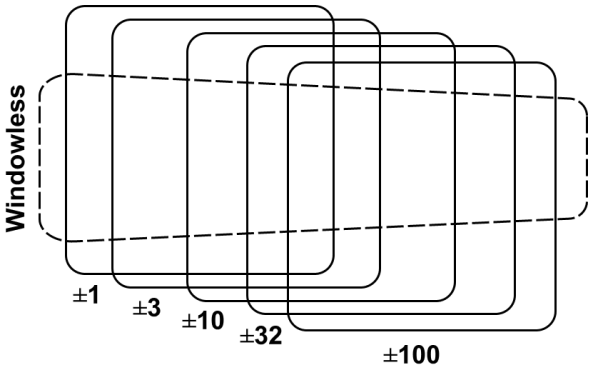

Figure 1: Approximate Venn representation of agreement between windowed and windowless association retrieval methods.

4.2 Statistical power

To paraphrase Kilgariff (2005), language is anything but random. A good language model is one which best captures the non-random structure of language. A good measuring device for any linguistic feature is therefore one which strongly differentiates real language from random data.

The solid lines in figures 2a and 2b give an indication of the relative confidence levels (*p*-values) attributable to a given association score derived from windowed co-occurrence data. Figure 2a is based on a window size of ±10 words, and 2b ±100 words. The data was generated, Monte Carlo style, from a 1 million word randomly generated corpus. For the sake of statistical convenience and realism, the symbols in the corpus were given a Zipf frequency distribution roughly matching that of words found in the Brown corpus (and most English corpora). Unlike with the previous experiment, *all* possible word pairings were considered. PMI was used for measuring association, owing to its convenience and similarity to co-dispersion, but it should be noted that the specific formulation of the association measure is more-or-less irrelevant in the present context, where we are using *relative* association levels between a real and random corpus as a proxy for how much structural information is captured from the corpus.

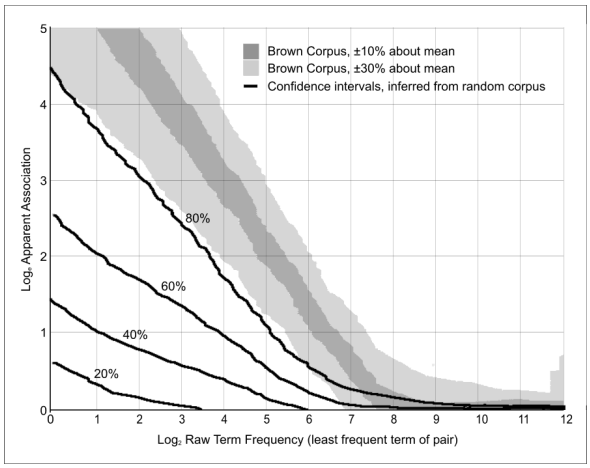

Figure 2a: Co-occurrence significances for a moderate (±10 words) window.

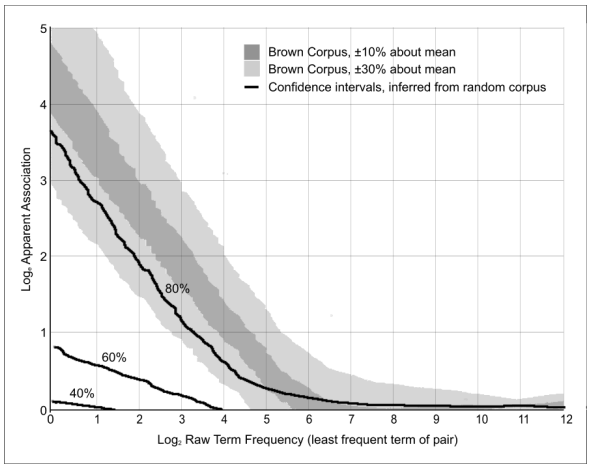

Figure 2b: Co-occurrence significances for a large (±100 words) window.

Precisely put, the figures show the percentage of times a given association score or lower was measured between word types in a corpus which is known to be devoid of any actual syntagmatic association. The closer to the origin these lines, the fewer word instances were required to be present in the random corpus before high levels of apparent association became unlikely, and so the fewer would be required in a real corpus before we could be confident of the import of a measured level of association. Consequently, if word pairs in a real corpus exceed these levels, we say that they show *significant association*.

The shaded regions in figures 2a and 2b show the typical range of apparent association scores found in a *real* corpus – in this case the Brown corpus. The first thing to observe is that both the spread of raw association scores and their significances are relatively constant across word frequencies, up to a frequency threshold which is linked to the window size. This constancy exists in spite of a remarkable variation in the raw association scores, which are increasingly inflated towards the lower frequencies (indeed illustrating the importance of taking statistical significance into account). This observed constancy is intuitive where long-range associations between words prevail: very infrequent words will tend to co-occur within the window less often than moderately frequent words - by simple virtue of their number - yet when they do co-occur, the evidence for association is that much stronger owing to the small size of the window relative to their frequency. Beyond the threshold governed by window size, there can be seen a sharp levelling out in apparent association, accompanied by an attendant drop in overall significance. This is a manifestation of Rapp's *specificity*: as words become much more frequent than window size, the kinds of tight idiomatic co-occurrences and compound forms which would otherwise imply an uncommonly strong association can no longer be detected as such.

A related observation is that, in spite of the lower random baseline exhibited by the larger window size, the actual *significance* of the associations it reports in a real corpus are, for all word frequencies, lower than those reported by the smaller window: i.e. *quantitatively* speaking, larger windows seem to observe less! Evidently, apparent association is as much a function of window size as it is of actual syntagmatic association; it would be very tempting to interpret the association profiles in figures 2a or 2b, in isolation of each other or their baseline plots, as indicating some interesting scale-varying associative structure in the corpus, where in fact they do not.

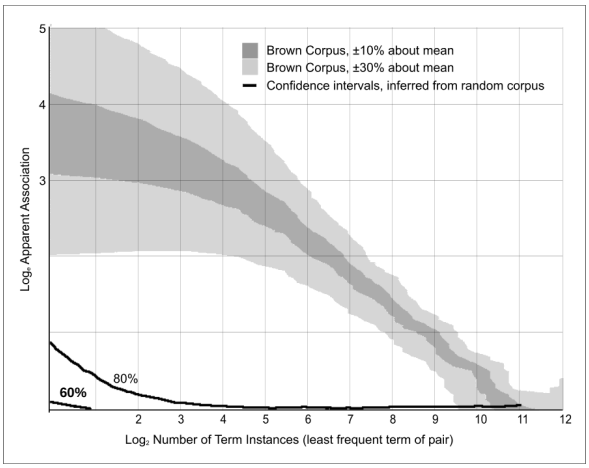

Figure 3: Significances for windowless co-dispersion.

Figure 3 is identical to figures 2a and 2b (the same random and real world corpora were used) but it represents the windowless co-dispersion method presented herein. It can be seen that the random corpus baseline comprises a smooth power curve which gives low initial association levels, rapidly settling towards the expected value of zero as the number of token instances increases. Notably, the bulk of apparent association scores reported from the Brown Corpus are, while not necessarily greater, orders of magnitude more significant than with the windowed examples for all but the most frequent words (ranging well into the 99%+ confidence levels). This gain can only follow from the fact that more information is being taken into account: not only do we now consider relationships that occur at all scales, as previously demonstrated, but we consider the exact distance between word tokens, as opposed to low-range ordinal values linked to window-averaged frequencies. There is no observable threshold effect, and without a window there is no reason to expect one. Accordingly, there is no specificity trade-off: while word pairs interacting at very large distances are captured (as per the largest of windows), very close occurrences are still rewarded appropriately (as per the smallest of window).

5 Conclusions and future direction

We have presented a novel alternative to co-occurrence for measuring lexical association which, while based on similar underlying linguistic intuitions, uses a very different apparatus. We have shown this method to gather more information from the corpus overall, and to be particularly unfettered by issues of scale. While the information gathered is, by definition, linguistically relevant, relevance to a given task (such as reproducing human association norms or performing word-sense disambiguation), or superior performance with small corpora, does not necessarily follow. Further work is to be conducted in applying the method to a range of linguistic tasks, with an initial focus on lexical semantics. In particular, properties of resultant word-space models and similarity measures beg a thorough investigation: while we would expect to gain denser higher-precision vectors, there might prove to be overriding qualitative differences. The relationship to grammatical dependency-based contexts which often out-perform contiguous contexts also begs investigation.

It is also pertinent to explore the more fundamental parameters associated with the windowless approach; the formulation of co-dispersion presented herein is but one interpretation of the specific case of association. In these senses there is much catching-up to do.

At the present time, given the key role of window size in determining the selection and apparent strength of associations under the conventional co-occurrence model - highlighted here and in the works of Church *et al* (1991), Rapp (2002), Wang (2005), and Schulte im Walde & Melinger (2008) - we would urge that this is an issue which window-driven studies continue to conscientiously address; at the very least, scale is a parameter which findings dependent on distributional phenomena must be qualified in light of.

Acknowledgements

Kind thanks go to Reinhard Rapp, Stefan Gries, Katja Markert, Serge Sharoff and Eric Atwell for their helpful feedback and positive support.

References

John A. Bullinaria. 2008. *Semantic Categorization Using Simple Word Co-occurrence Statistics*. In: M. Baroni, S. Evert & A. Lenci (Eds), Proceedings of the ESSLLI Workshop on Distributional Lexical Semantics: 1 - 8

John A. Bullinaria and Joe P. Levy. 2007. *Extracting Semantic Representations from Word Co-occurrence Statistics: A Computational Study*. Behavior Research Methods, 39:510 - 526.

Yaacov Choueka and Serge Lusignan. 1985. *Disambiguation by short contexts*. Computers and the Humanities. 19(3):147 - 157

Kenneth W. Church and Patrick Hanks. 1989. *Word association norms, mutual information, and lexicography*. In Proceedings of the 27th Annual Meeting on Association For Computational Linguistics: 76 - 83

Kenneth W. Church, William A. Gale, Patrick Hanks and Donald Hindle. 1991. *Using statistics in lexical analysis*. In: Lexical Acquisition: Using Online Resources to Build a Lexicon, Lawrence Erlbaum: 115 - 164.

P. J. Clark and F. C. Evans. 1954. *Distance to nearest neighbor as a measure of spatial relationships in populations.*Ecology. 35: 445 - 453.

Béatrice Daille. 1994. *Approche mixte pour l'extraction automatique de terminologie: statistiques lexicales et filtres linguistiques*. PhD thesis, Université Paris.

Sally F. Dennis. 1965. *The construction of a thesaurus automatically from a sample of text*. In Proceedings of the Symposium on Statistical Association Methods For Mechanized Documentation, Washington, DC: 61 - 148.

Philip Edmonds. 1997. *Choosing the word most typical in context using a lexical co-occurrence network*. In Proceedings of the Eighth Conference on European Chapter of the Association For Computational Linguistics: 507 - 509

Stefan Evert. 2007. *Computational Approaches to Collocations: Association Measures*, Institute of Cognitive Science, University of Osnabruck, <http://www.collocations.de>.

Manfred Wettler, Reinhard Rapp and Peter Sedlmeier. 2005. *Free word associations correspond to contiguities between words in texts*. Journal of Quantitative Linguistics, 12:111 - 122.

Michael K. Halliday. 1966 *Lexis as a Linguistic Level*, in Bazell, C., Catford, J., Halliday, M., and Robins, R. (eds.), In Memory of J. R. Firth, Longman, London.

David Hardcastle. 2005. *Using the distributional hypothesis to derive cooccurrence scores from the British National Corpus*. Proceedings of Corpus Linguistics. Birmingham, UK

Kei Yuen Hung, Robert Luk, Daniel Yeung, Korris Chung and Wenhuo Shu. 2001. *Determination of Context Window Size*, International Journal of Computer Processing of Oriental Languages, 14(1): 71 - 80

Stefan Gries. 2008. *Dispersions and Adjusted Frequencies in Corpora*. International Journal of Corpus Linguistics, 13(4)

Frank Keller and Mirella Lapata. 2003. *Using the web to obtain frequencies for unseen bigrams*, Computational Limguistics, 29:459 – 484

Adam Kilgarriff. 2005. Language is never ever ever random. Corpus Linguistics and Linguistic Theory 1: 263 - 276.

George Kiss, Christine Armstrong, Robert Milroy and James Piper. 1973. *An associative thesaurus of English and its computer analysis*. In Aitken, A.J., Bailey, R.W. and Hamilton-Smith, N. (Eds.), The Computer and Literary Studies. Edinburgh University Press.

Abolfazl K. Lamjiri, Osama El Demerdash and Leila Kosseim. 2003. *Simple Features for Statistical Word Sense Disambiguation*, Proceedings of Senseval-3:3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text: 133 - 136.

Uwe Quasthoff. 2007. *Fraktale Dimension von Wörtern*. Unpublished manuscript.

Reinhard Rapp. 2002. *The computation of word associations: comparing syntagmatic and paradigmatic approaches.* In Proceedings of the 19th international Conference on Computational Linguistics.

D. L. Sackett. 2001. *Why randomized controlled trials fail but needn't: 2. Failure to employ physiological statistics, or the only formula a clinician-trialist is ever likely to need (or understand!)*. CMAJ, 165(9):1226 - 37.

Magnus Sahlgren. 2006. *The Word-Space Model: using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector space*, PhD Thesis, Stockholm University.

Petr Savický and Jana Hlavácová. 2002. *Measures of word commonness.* Journal of Quantitative Luiguistics, 9(3): 215 – 31.

Cyrus Shaoul, Chris Westbury. 2008. *Performance of HAL-like word space models on semantic clustering.* In: M. Baroni, S. Evert & A. Lenci (Eds), Proceedings of the ESSLLI Workshop on Distributional Lexical Semantics: 1 – 8.

Sabine Schulte im Walde and Alissa Melinger, A. 2008. *An In-Depth Look into the Co-Occurrence Distribution of Semantic Associates*, Italian Journal of Linguistics, Special Issue on From Context to Meaning: Distributional Models of the Lexicon in Linguistics and Cognitive Science.

Egidio Terra and Charles L. A. Clarke. 2004. *Fast Computation of Lexical Affinity Models*, Proceedings of the 20th International Conference on Computational Linguistics, Geneva, Switzerland.

Xiaojie Wang. 2005. *Robust Utilization of Context in Word Sense Disambiguation*, Modeling and Using Context, Lecture Notes in Computer Science, Springer: 529-541.

Justin Washtell. 2006. *Estimating Habitat Area & Related Ecological Metrics: From Theory Towards Best Practice*, BSc Dissertation, University of Leeds.

Justin Washtell. 2007. *Co-Dispersion by Nearest Neighbour: Adapting a Spatial Statistic for the Development of Domain-Independent Language Tools and Metrics*, MSc Thesis, University of Leeds.

Warren Weaver. 1949 *Translation*. Repr. in: Locke, W.N. and Booth, A.D. (eds.) *Machine translation of languages: fourteen essays* (Cambridge, Mass.: Technology Press of the Massachusetts Institute of Technology, 1955), 15-23. Association for Computing Machinery, 28(1):114-133.

David Yarowsky and Radu Florian. 2002. *Evaluating Sense Disambiguation Performance Across Diverse Parameter Spaces.* Journal of Natural Language Engineering, 8(4).