The Poisson Collocation Measure and its Applications

Uwe QUASTHOFF, Christian WOLFF
Leipzig University, CS Institute, NLP Dept.
Augustusplatz 10/11
Leipzig, Germany, 04109
{quasthoff, wolff}@informatik.uni-leipzig.de

Abstract

In this paper we introduce a measure for calculating statistically significant collocation sets that is related to the Poisson distribution. We show that results calculated using this measure are comparable to well-known measures like the log-likelihood measure. Additionally, we discuss asymptotic behaviour and additivity as general properties of the measure which may be applied to the analysis of multi-word collocations as well and can be used for defining a relative collocation measure as well. Finally, we give a brief overview of various possibilities of postprocessing collocation sets.

1 Introduction

In this paper, the term collocation is used for two or more words with the following statistical property: In a given large corpus, they occur significantly often together within a predefined window. Useful windows are

- Next neighbours
- Sentences
- Fixed-size Windows (e. g. *n* word or character distances)
- Documents
- Collections of Documents

We will concentrate on the first two kinds of windows, i. e. next neighbours and sentences, and give only some remarks for very large windows. This selection is motivated by the trivial observation that word neighbourhood as well as sentences boundaries are restrictions that allow for a syntactic as well as semantic interpretation of some kind while fixed size windows impose a restriction that is merely technically motivated.

Collocations calculated using these kinds of window will often be found to carry different types of semantic relations. Having found collocations, the next challenging problem is to identify the corresponding relation. Here, both syntactic and semantic knowledge can be used.

2 Introduction to the Poisson collocation measure

We are interested in the joint occurrence of two given words *A* and *B* with probabilities p_a and p_b within a sentence. Let our corpus contain *n* sentences. For simplicity we will assume that both *A* and *B* occur at most *once* in any sentence. This is approximately correct if *A* and *B* are not high frequency words. To measure the surprise of a joint occurrence of *A* and *B* we first note that under the assumption of independence of *A* and *B* we get a probability of $p_a p_b$ for their joint occurrence in a sample sentence. The number *n* of sentences in the corpus can be considered as the number of repeated experiments. Using a Poisson distribution [cf. Chung 2000] we get the following approximation for *k* joint occurrences in the corpus of *n* sentences, where as usual $\lambda = n\, p_a p_b$:

$$p_k = \frac{1}{k!} \cdot \lambda^k \cdot e^{-\lambda}. \qquad (1)$$

We are interested in the case of at least k joint occurrences, i.e.

$$\sum_{l=k}^{\infty} \frac{1}{l!} \cdot \lambda^l \cdot e^{-\lambda}. \qquad (2)$$

As significance measure for collocations we choose the negative logarithm of this probability divided by the logarithm of the size of the corpus:

$$\mathrm{sig(A,B)} = \frac{-\log \sum_{l=k}^{\infty} \frac{1}{l!} \cdot \lambda^l \cdot e^{-\lambda}}{\log n}. \qquad (3)$$

For typical cases, λ is small. Hence, the above sum can be approximated by its first term and we get:

$$\mathrm{sig}(\mathrm{A},\mathrm{B}) = \frac{\lambda - k \cdot \log \lambda + \log k!}{\log n}. \qquad (4)$$

The above approximation gives good results for $(k+1) / \lambda > 10$, which is the typical case. If, moreover, $k \geq 10$ holds, we might use Stirling's formula to get

$$\mathrm{sig}(\mathrm{A},\mathrm{B}) = \frac{k \cdot (\log k - \log \lambda - 1)}{\log n}. \qquad (5)$$

The normalizing factor is mainly chosen to ensure the properties given in section 2.3. Notwithstanding the discussion in [Church, Gale 95, Church 00] the Poisson-based approached described above yields quite useful data, especially for *content* words not belonging to *extreme* frequency categories. It was used for calculating the online collocations of German, English, French and Dutch corpora up to 20 million of sentences at *www.wortschatz.uni-leipzig.de.*

2.1 Comparison to the log-Likelihood Measure

One of the most popular collocation measures in text analysis is the log likelihood (Lgl) measure as introduced to the analysis of large text corpora by [Dunning 93].
Translating the formula given in [Krenn 00a, 00b] into our notation and ignoring small terms we get

$$\mathrm{Lgl}(\mathrm{A},\mathrm{B}) = \frac{k \cdot (\log k - \log \lambda)}{n}. \qquad (6)$$

Up to the difference in the normalization factor both formulae are very similar. Consequently, the collocations calculated do only differ slightly. This can be seen comparing the results described above with the collocations of *http://www.ids-mannheim.de/kt/corpora.shtml*. While both the corpora and the calculation methods differ, the results are remarkably similar (see tables 1 and 2).

	IDS Cosmas I (W-PUB)	*Wortschatz (German)*
Corpus Size	374 Mio	255 Mio
Sources	Mainly Newspapers	Mainly Newspapers
Window size	Fixed size (here: ±5 words	Sentence
Collocation Measure	Log Likelihood	Poisson Distribution

Table 1: Comparison of Collocation Resources in Different Corpora

Rank	*IDS Cosmas I*	*Cosmas Rating*	*Wortschatz (German)*	*Sig-Rating*
1	Wein	4351	trinken	1234
2	trinken	2745	Wein	648
3	getrunken	1715	getrunken	478
4	kühles	1627	Liter	460
5	Glas	1379	trinkt	428
6	Liter	1318	Glas	348
7	Faß	1236	Schnaps	318
8	Fass	1139	Hektoliter	300
9	Flasche	1071	Flaschen	272
10	Hektoliter	899	gebraut	269
11	Trinkt	881	Wein	244
12	Flaschen	873	Kaffee	242

Table 2: Most Significant Collocations for "Bier"

2.2 Multi-Word Collocations

The above Poisson approach can easily be adopted to multi-word collocations as well: Calculating the Poisson collocation measure for s words $A_1, A_2, \ldots, A_s$ with probabilities $p_1, p_2, \ldots, p_s$ we set $\lambda = n\, p_1 p_2 \ldots p_s$. For k joint occurrences ($k \geq 10$) we define

$$\mathrm{sig}(\mathrm{A}_1,\ldots,\mathrm{A}_s) = \frac{k \cdot (\log k - \log \lambda - 1)}{(s-1)\log n}. \qquad (7)$$

The additional normalizing factor ensures a nice asymptotic behaviour as can be seen in the next section. Multi-word collocations again can be calculated for next neighbours or at sentence level.

Because of the variable number of words in such a collocation set, it is difficult to decide whether they represent just one relation (for instance, co-hyponymy in a collocation set like the set *nickel, cadmium, copper, iron, zinc, chromium*) or a mixture of different types of relations. In the case of longer next neighbour collocations their boundaries are

not always visible. Here we have to add additional filtering techniques (see ch. 5 below) to identify linguistically useful collocations.
Some examples for German next neighbour collocations of length 3-5 are given in appendix A (ch. 8). Note that in the last three examples it is not clear from the numbers whether a leading determiner belongs to the phrase.

2.3 General Properties of our Collocation Measure

The following properties can easily be verified using the approximation formula (see equ. 4 above).

2.3.1 Asymptotic Properties

In order to describe basic properties of our measure, we write sig(n, k, a, b) instead of sig(A, B) where n, k are defined as above and a, and b are individual total frequencies of A and B, resp. Analogously, we write sig(n, k, a_1, a_2, ..., a_s) in the case of multi-words. The following asymptotic relations hold:

Simple co-occurance: If the words A_1, A_2, …, A_s occur only once, and they occur together:

$$\text{sig}(n,1,1,1,\ldots,1) \to 1 \quad (\text{for } n\to\infty). \quad (8)$$

Independence: A and B occur statistically independently with probabilities p and q:

$$\text{sig}(n,npq,np,nq) \to 0 \quad (\text{for } n\to\infty). \quad (9)$$

Hence, the collocation measure is scalable in a way that absolute values are comparable for multi-word collocations of different size. From a more practical point of view one can say that a collocation measure > 3 usually leads to collocations for which a meaningful interpretation can be given.

2.3.2 Additivity

The unification of the words B and B' just adds the corresponding significances. For $k / b \approx k' / b'$ we have

$$\text{sig}(n,k,a,b) + \text{sig}(n,k',a,b') \approx \text{sig}(n,k+k',a,b+b') \quad (10)$$

The same is, of course, true for more than two objects. This property has several implications for collocation processing, depending of the type of unified words: Using this we can unify several words to form a concept. Additivity as introduced above ensures the same results for the following two operations: First, unify several words to some kind of (virtual) concept, and second, calculate the collocation measure for this concept. Alternatively, calculate the collocation measure of this concept *a posteriori* by adding the corresponding collocation measures of the words contained in the concept. The latter is much more convenient because our collocation calculation tools yield a *complete* database of collocation pairs for any given corpus. This information may be subject to this unification process at a later stage.

2.4 Corpus Size

We are surprised if we observe a rare event much more often than expected. If, in a much longer observation, we still observe this event at a higher rate, we are more surprised because we get convinced to see a regularity. Here, the rare event is the joint occurrence of two words and a longer observation corresponds to a larger corpus.
Enlarging the corpus by a factor m gives:

$$\text{sig}(mn, mk, ma, mb) \approx m\ \text{sig}(n, k, a, b). \quad (11)$$

In other words, for a pair of words with a low collocation measure we can test a larger corpus. If their joint occurrence is not by chance, the collocation measure should increase with the corpus.

2.5 A Relative Collocation Measure

The above phenomenon of additivity suggests the introduction of a *relative collocation measure*. Here we calculate the collocation measure of a fixed word C. We define

$$\text{sig}_C(A) = \frac{sig(A,C)}{\sum_B sig(B,C)}, \quad (12)$$

where the sum is taken over all collocates B of C.

While the collocation measure sig(A,B) is symmetric in A and B, this is no longer true for the relative collocation measure.
In the above definition, the sum over all collocates B might be difficult to calculate. As a crude approximation for this sum we might use the maximum possible collocation measure for a given word C. This maximum is achieved for a word Y which always appears together with C. Hence, we get $\lambda = k^2 / n$, and, for $k << n$:

$$\max_B(\mathrm{sig}_C(B)) = \mathrm{sig}_C(Y) \approx k. \qquad (13)$$

If c is the frequency of C, we get the following approximation for $\mathrm{sig}_C(A)$:

$$\mathrm{sig}_C(A) = \frac{sig(A,C)}{c}. \qquad (14)$$

Two immediate implications of this measure appear to be obvious:

- First we can use this to decide whether the most significant collocation of a given word is so strongly connected that we can expect a *fixed multiword construct* like in fixed proper name phrases.
- Second, we can use the result for deviding collocation sets into subsets corresponding to subject area, part of speech, or semantic type to get a frame-like representation of the collocation set. Some exmaples are given below.

3 Applications

In the following subsections, we give some examples for postprocessing of collocation sets calculated using the above mentioned approach. For a more detailed review of technical aspects see [Quasthoff & Wolff 00]. The basic idea behind the various postprocessing operations is the observation that significant collocations represent a *universal principle of relatedness* between two informational items which has to be further analysed for meaningful interpretation - natural language texts are only one type of information items that may be subjected to collocation analysis; other applications exist, e. g. in genome analysis. All examples given below are derived from our collocation databases.

3.1 Collocations of Basic Forms

We can unify all the inflected forms of a basic form. Here, additivity ensures the same result whether calculating the collocation measure directly for basic forms (*basic layer of analysis,* see ch. 6 below) or summing up the corresponding values for the inflected forms (*postprocessing layer* (see ch. 6 and fig. 1 in the appendix)).

While it is possible that the interpretation of collocations differs significantly for inflected forms (e. g. in German, the word *Schwein* (pig) in its singular form has significantly different types of collocations in comparison with its plural form *Schweine* (pigs)), in general the unification of inflected forms makes sense. Leaving the unification to the pastprocessing keeps the detailed information, if necessary.

3.2 Collocations of Semantic Types and the Separation of Collocation Sets

Another application of additivity of collocation measures is to estimate the frequencies for the different meanings of polysemous words. For instance, the collocations of *space* taken from our general language corpus of English fall mainly into three classes: The subject areas *computer*, *real estate* and *outer space*. The corresponding senses of *space* are denoted with $space_1$, $space_2$, and $space_3$. Assigning the top 30 collocations of space (*disk, shuttle, square, station, NASA, feet, ...)* to these three senses we get an qualitative estimate of these senses:

space$_1$ 28.2%: *disk (2629), memory (718), storage (479), program (308), RAM (307), free (300), hard (336)*

space$_2$ 53.2%: *shuttle (2618), station (991), NASA (920), Space (602), launch (505), astronauts (473), Challenger (420), manned (406), NASA's (297), flight (293), Atlantis (291) Mir (335), rocket (329), orbit (326), Discovery (341), mission (385)*

space$_3$: 18.6%: *square (1163), feet (822), leased (567), office (382), lessor (390)*

With a complete database of collocation sets for any given corpus being available, such numbers are much easier to calculate than manually counting classified sentences containing *space*.

3.3 Identification of Proper Names and Phrases

A large relative collocation measure $sig_C(A)$ indicates that a reasonable part of all occurrences of the word *C* is together with *A*. The opposite need not be true, as *A* can be much more frequent then *C*. Such pairs are often good candidates for proper names or phrases as can be seen in Table 3. The "head" denotes the word *C*.

Left Word	*Right Word*	*"head"*
Alzheimersche	Krankheit	left
AQA	total	left
Anorexia	nervosa	left and right
Algighiero	Boetti	left and right
30jährige	US-Bond	right
André	Lussi	right

Table 3: Pairs with Large Relative Collocation Measure

3.4 Compound Decomposition

Multi-word collocations as described above can be employed for the decomposition and the semantic interpretation of compounds which are a notorious problem for text analysis in languages like German. Table 4 below is constructed the following way: We first we try to decompose a compound like *Geschwindigkeitsüberschreitung* into the parts *Geschwindigkeit* and *Überschreitung*. Next we look for multi-word collocations containing the above parts as borders. If the multi-word collocation is of some predefined form (here: *A der B*), we accept this collocation as a semantic description. In our example, we get *Überschreitung der Geschwindigkeit*. Using only a few patterns will produce many descriptions for compounds. Patterns are selected using syntactic (prepositional phrases) as well as semantic considerations (the phrase structure should represent some type of explanation for a meaningful relationship between two concepts).

Pattern	*Word A*	*Word B*	*Compound*
A aus B	Orgie	Farben	Farbenorgie
A der B	Bebauung	Insel	Inselbebauung
A mit B	Feld	Getreide	Getreidefeld
A in der B	Feldbau	Regenzeit	Regenzeitfeldbau
A für B	Übung	Anfänger	Anfängerübung
A für die B	Gebäude	Flugsicherung	Flugsicherungsgebäude
A von B	Anbau	Kaffee	Kaffeeanbau
A zur B	Andrang	Eröffnung	Eröffnungsandrang

Table 4: Examples for Multiword Collocations Used for Segmenting and Identifying Compounds

3.5 Filtering of Collocation Sets

An obvious postprocessing step for collocation sets is filtering using categorical information. Without postprocessing, collocation sets not only contain word pair relations which may be attributed with different semantic relations, they also contain words of different POS categories. Given category information for the members of a collocation set, typical combinations of nouns and verbs or nouns and adjectives may be extracted for a given corpus. The following example gives adjectives as significant left neighbours of *Husten* (cough):

bellender, trockener, verschleimten, heiseres, trockenen, trockenem, blutiger, heftiges, leichtem, anhaltender, kleiner, schrecklichen, heftigen.

While it should be immediately obvious that results like these are interesting for lexicography or language learning, other areas of application additionally draw domain-specific corpora into account: Either by simply analysing a domain-specific corpus or by comparing analysis results for a general-purpose corpus with a domain-specific corpus, category filtering can be applied in areas like software reengineering (extraction of class – instance relationships, finding typical attributes and methods given descriptive texts on a specific software project) or knowledge engineering (using category information so segment collocation sets which are then used as seed information for the generation of semantic networks or Topic Maps). For a detailed discussion of category filtering, see [Heyer et al. 01b].

4 Concluding Remarks

Ch. 1 - 4 of this paper dealt with the basis of our approach towards corpus, i. e. the statistical layer of collocation analysis, while ch. 5 discussed collocation postprocessing using additional information. As should have become obvious, different *methodical approaches* may be applied in postprocessing:

- Changing the *parameters of collocation analysis* like grouping of inflected forms for collocation analysis
- *Filtering by introducing additional* (syntactic and / or semantic) *knowledge*.
- *Separation* of collocation sets by introducing *comparative* corpus analysis methods.

We have practically applied our collocation measure to areas as diverse as information retrieval, knowledge management, knowledge extraction / named entity recognition, analysis of time-related semantic trends or document and text classification.

In general, it has become obvious that a robust statistical measure can only be a *starting point* for further applications. In more general terms, a four layered system architecture has evolved for applied collocation analysis:

- *Corpus Preprocessing*: This layer comprises all necessary tools for text and document import and conversion, text and sentence segmentation, word tokenisation etc. Results are stored in a relational database system. Preprocessing can be done in the same way for arbitrary types of corpora (like text sets differing in domain, language, time or any other generic attribute).
- *Base layer*: robust and complete analysis of large corpora, generating a comprehensive database of collocations for a given corpus (ch. 1-4). We have developed an infrastructure for corpus analysis that works for very large corpora as well as for different languages and comprises necessary processing steps like text and sentence segmentation as well [cf. Quasthoff & Wolff 00].
- *Postprocessing layer*: Integrating collocation analysis with other knowledge sources like frequency information, linguistic features or subject categories [ch. 5 above; cf. Heyer et al. 01a, Heyer, Quasthoff, Wolff 02].
- *Application layer*: Practical application of results from layers 1 and 2 for specific text analysis problems like the areas mentioned above [cf. Heyer, Quasthoff, Wolff 00].

Figure 1 in the appendix gives a schematic overview of this system architecture. It should be noted that the application of additional knowledge in the postprocessing layer is symbolized by taking information from a

generic and / or domain-specific database for postprocessing of collocation sets. For a further illustration of the contents of our reference database, see [Quasthoff & Wolff 00] or http://wortschatz.uni-leipzig.de.

5 References

[Armstrong 93]. Armstrong, S. (ed.); "Using Large Corpora"; Computational Linguistics 19, 1/2 (1993) [Special Issue on Corpus Processing, repr. MIT Press 1994].

[Chung 00]. Chung, K. L. "A Course in Probability Theory", Academic Press 2000.

[Church 00]. Church, K. W. (2000), "Empirical Estimates of Adaptation: The chance of Two Noriega's is closer to p/2 than p2," Proc. Coling 2000, 173-179.

[Church, Gale 95] Church, K. W.; Gale, W. "Poisson Mixtures". In: Journal of Natural Language Engineering 1, 2, 163-190.

[Dunning 93]. Dunning, T. "Accurate Methods for the Statistics of Surprise and Coincidence". In: Computational Linguistics 19, 1 (1993), 61-74.

[Heyer et al. 01a] Heyer, G.; Läuter, M.; Quasthoff, U.; Wittig, Th.; Wolff, Ch.; "Learning Relations using Collocations"; In: Proc. IJCAI Workshop on Ontology Learning, Seattle/WA, August 2001, 19-24.

[Heyer et al. 01b] Heyer, G.; Läuter, M.; Quasthoff, U.; Wolff, Ch. „Wissensextraktion durch linguistisches Postprocessing bei der Corpusanalyse". In: Lobin, H. (ed.) (2001). Sprach- und Texttechnologie in digitalen Medien. Proc. GLDV-Jahrestagung 2001, Universität Gießen, 71-83

[Heyer, Quasthoff, Wolff 00] Heyer, G.; Quasthoff, U.; Wolff, Ch.; "Aiding Web Searches by Statistical Classification Tools. " Proc. Proc. 7. Intern. Symposium f. Informationswissenschaft ISI 2000, UVK, Konstanz (2000), 163-177.

[Heyer, Quasthoff, Wolff 02] Heyer, G.; Quasthoff, U.; Wolff, Ch.; "Knowledge Extraction from Text: Using Filters on Collocation Sets." Accepted Paper for LREC 2002.

[Krenn 00a] Brigitte Krenn. 2000. Empirical Implications on Lexical Association Measures. Proceedings of the Ninth EURALEX International Congress. Stuttgart, Germany.

[Krenn 00b] Krenn, B.; "Distributional and Linguistic Implications of Collocation Identification." Proc. Collocations Workshop, DGfS Conference, Marburg, March 2000.

[Läuter & Quasthoff 99] Martin Läuter and Uwe Quasthoff. 1999. Kollokationen und semantisches Clustering. In *11. Jahrestagung der GLDV,* Enigma Corporation, Prag.

[Lemnitzer 98] Lemnitzer, L.; "Komplexe lexikalische Einheiten in Text und Lexikon." In: Heyer, G.; Wolff, Ch. (edd.). Linguistik und neue Medien. Wiesbaden: Dt. Universitätsverlag, 1998, 85-91.

[Maedche & Staab 01] Maedche, A.; Staab, St.; „Ontology Learning for the Semantic Web"; IEEE Intelligent Systems 16, 2 (2001), 72-79.

[Manning & Schütze 99]. Manning, Ch. D.; Schütze, H.; Foundations of Statistical Language Processing; Cambridge/MA, London: The MIT Press 1999.

[Quasthoff & Wolff 00] Quasthoff, U.; Wolff, Ch.; "An Infrastructure for Corpus-Based Monolingual Dictionaries." Proc. LREC-2000. Second International Conference on Language Resources and Evaluation. Athens, May / June 2000, Vol. I, 241-246.

[Schatz 02] Schatz, B.; "The Interspace: Concept Navigation across Distributed Communities"; IEEE Computer 35, 1 (2002), 54-62.

[Smadja 93] Smadja, F.; "Retrieving Collocations from Text: Xtract"; Computational Linguistics 19, 1 (1993), 143-177.

6 Appendices

6.1 Appendix A: Examples for German Next Neighbour Collocations of Length 3-5

Multi Word Example	a_1	a_2	a_3	a_4	a_5	k	$\lambda = a_1 \ldots a_s / n^{s-1}$	$(k+1)/\lambda$	*sig*
Haut und Haar	4384	3617417	3366	-	-	73	0.0534	1368	33.94
Heraufsetzung des Rentenalters für Frauen	132	1222152	149	1180668	68151	11	0.000000028	$4.23 \cdot 10^8$	15.68
arm wie eine Kirchenmaus	1183	480352	913563	15	-	11	0.000000007	$1.54 \cdot 10^9$	12.10
Ausstieg aus der Atomenergie	2152	694802	5463574	1055	-	170	0.00000862	$1.98 \cdot 10^7$	145.98
dem Ausstieg aus der Atomenergie	1084573	2152	694802	5463574	1055	8	0.000000093	$9.62 \cdot 10^7$	7.60
der Ausstieg aus der Atomenergie	5463574	2152	694802	5463574	1055	14	0.00000047	$3.19 \cdot 10^7$	12.44

Table 5: Some Results for Multi-word Collocation Analysis

6.2 Appendix B: System Architecture Overview

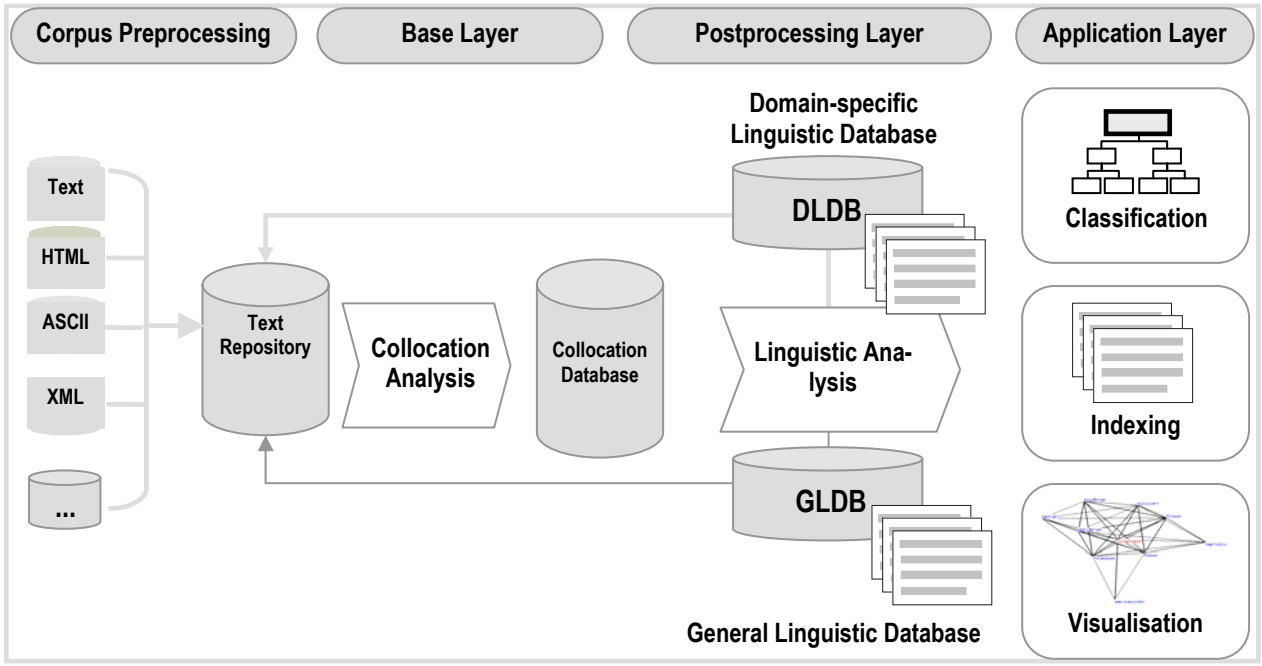

Figure 1: Corpus and Collocation Analysis System Architecture Overview