Investigating semantic issues in Latin by means of network analysis

Paola Marongiu[1]

[1]Université de Neuchâtel

1 Introduction

This paper will illustrate how network analysis can be employed in order to explore a specific area of semantics, that is, the expression of modality. The study focuses in particular on the co-occurrence of modal markers. This work is meant to provide a quantitative and structural dimension of this understudied phenomenon, by means of network analysis. We performed the analysis on a corpus of Classical Latin texts. We will illustrate the workflow of this project, from the creation of the corpus to the construction of the networks. We will then analyze the networks and we will show the main features of the network built on the whole corpus, and then of the sub-networks built for each genre represented in the corpus. Finally, we will illustrate how the networks obtained can still be enriched by a closer, qualitative analysis of the texts.

2 Modality and co-occurrence

The definition of modality and its (sub)categories has been a discussed subject among scholars [1]. In this work, I will adopt a theoretical framework based on Nuyts (2016) and Dell'Oro (2019): modality can be described as the expression of necessity, possibility and volition. For the sake of this project, we will take into account only markers of necessity and possibility, in order to study their interaction at the sentence level. Modality can be expressed, at the lexical or at the morphological level, by the modal markers. This study will focus only on lexical modal markers. The modal marker carries a basic value of possibility or necessity, and refers to a more or less extensive portion of text, which is called scope. Let us consider Example 1, from Nuyts (2016):

(1) I'll come down for dinner soon, darling, but I need to finish this letter first.

In this sentence, need is a modal marker of necessity. The modal marker refers to the scope to finish this letter first.

[1] It is beyond the scope of this contribution to make a State of the Art about modality. However, important references for the study of modality are Nuyts and Van Der Auwera (2016) for an overview of different theoretical approaches to the study of modality; Bolkestein (1980), Fruyt and Moussy (2002) and Núñez (1991) for modality in Latin; Lyons (1977), Huot (1974), Sueur (1975), Palmer (2001), Chu (2008), Narrog (2009) and Hütsch (2020) for the co-occurrence of modal markers.

Once having established the definition of modality and the notion of modal markers, we need to define what we mean by 'co-occurrence of modal markers'. We consider two markers to be in co-occurrence if they appear in the same sentence, that is, in the same portion of text which is enclosed between two full stops. See Example 2: in this sentence, can and have to are the two co-occurring modal markers. They express respectively possibility and necessity.

(2) You can go to the cinema, but you have to finish your homework first.

Therefore, the two co-occurring modal markers can produce the following combinations:

- Co-occurring markers of the same type: necessity – necessity / possibility – possibility;
- Co-occurring markers of different types: necessity – possibility / possibility – necessity.

.

3 The method: network analysis in linguistics

Language can be represented as a complex network and analyzed as such. A (linguistic) network is represented by the formula N = (V, E), where V represents the vertices (or nodes) of the network and E represents the edges (or links) between them (i Cancho, 2010). In a linguistic network the nodes represent the linguistic units, and the edges some type of relation between them (Passarotti, 2015). One method for building linguistic networks is using syntactically annotated corpora. The structure of such networks, called dependency networks, is made of lemmas (the nodes) and the relations of syntactic dependency between them (the edges) (i Cancho, 2005; Mehler, 2008). Network analysis has mainly been applied to the study of modern languages. Nevertheless, some studies translated into networks the existing treebanks for Latin and Ancient Greek, in order to discuss different annotation schemas and to study linguistic properties such as non-configurationality (Passarotti, 2015; Ponti and Luraghi, 2018). However, as our study focuses on the co-occurrence of modal markers, which not necessarily are in a relation of direct syntactic dependency, we work with co-occurrence networks. In this type of structures, the links are not determined by a dependency relation between the nodes, but by the fact that they co-occur in a certain context. In our case, the nodes of the networks represent the lemmas of the modal markers that co-occur in the corpus. The nodes are assigned an attribute node_type, which has two possible values depending on the basic meaning of the modal marker (necessity or possibility, represented by the logical operators $\Box$ for necessity and $\Diamond$ for possibility (Portner, 2009). See Table 2 for a complete list). The edges that connect the nodes are established basing on the co-occurrence of the modal markers in the context of a sentence. If two modal markers co-occur at least once in the corpus, an edge will be drawn between the respective two nodes. The edges are given a weight value, which is determined by the number of co-occurrences between the two markers. Another attribute of the edges is their direction: the order in which the markers appear in the sentence determines the direction of the edge. This type of information is important in view of a further closer analysis of the contexts of co-occurrence: by looking at the sentences that host a co-occurrence, we want to be able to determine

if the type of syntactic structure and the sequence in which the markers co-occur has an impact on the semantic and syntactic relation between them [2]. See the Example 3 from the Rhetorica ad Herennium 4, 1:

(3) In praecipiendo expresse conscripta ponere oportet exempla, uti in artis formam convenire possint.
'In instructing one must cite examples that are drafted expressly so that they could conform to the pattern of art.'

In this case, a node oportet will be connected by a directed edge to the node possum, because they co-occur in the same portion of text enclosed between two full stops (our sentence). The qualitative analysis will add the information about the type of modality that they express and the type of syntactic structure (possint is in a final subordinate clause introduced by oportet).

The networks were analyzed by means of the software Cytoscape (Shannon et al., 2003).

4 The data

The networks were built on a corpus of Classical Latin texts, representative of four different genres: historiography, letter, oratory and treatise. The reason why we chose to have such variation in the corpus is that we wanted to be able to study the co-occurrence of modal markers first from a general perspective, and then looking at different genres. In this way, we will be able to determine if a certain type of co-occurrence and the type of syntactic structures in which it is established turn out to be more characteristic for some genres with respect to others. This enables us to detect possible language changes with respect to the type of text, the register, its contents and public. These variables change for all the genres represented in the corpus. In order to avoid that the analysis could be biased by the style of a specific author, we selected texts from (at least) two authors for each genre. In the case of the Epistulae ad familiares by Cicero, although the collection bears the name of this author, it also counts the letters received from his correspondents, which ensures the variation by author in the genre of letters. Thanks to this architecture, the genre-based analysis can be considered representative for each genre; at the same time, single analyses on a specific author are also possible. In Table 1 we show the specific texts and authors that are part of the study.

The corpus is also balanced with respect to the size of the texts: each genre is represented by a number of tokens that spans from 118k to 128k, and the same range of variation exists between texts of different authors. This ensured a lower probability of over-representations for some genres or authors on the analysis of the whole corpus. The texts were downloaded from three different open sources: the Perseus Digital Library (http://www.perseus.tufts.edu), the Latin Library (https://www.thelatinlibrary.com/) and the IntraText Digital Library (http://www.intratext.com/LATINA/). They were then annotated with the parser Stanza (Qi et al., 2020). We chose the annotation model IT-TB, trained on the Index Thomisticus treebank (Cecchini et al., 2018; Passarotti, 2011), among the three available for Latin, as it had the best accuracy on the lemmatization task. This type of results was fundamental for the next step of our study, as the nodes are based on the

[2] Building on, e.g., the work by Narrog (2009)

Table 1: Corpus of Classical Latin.

letters	Cicero Epistulae ad familiares	118k tokens
historiography	Caesar Commentarii de bello gallico; Sallust Bellum Catilinae and Bellum Iugurthinum; Bellum Africanum; Bellum Hispaniense	128k tokens
treatise	Varro De re rustica; Vitruvius De architectura; Rhetorica ad Herennium	123k tokens
oratory	Cicero Philippicae; Seneca the Elder Controversiae	118k tokens

lemmas of the modal markers. The texts included in the corpus were also annotated with metadata, in the format author|genre. By using the metadata, we could easily retrieve the genre and author of the texts from which we extracted the co-occurrences.

In order to get the co-occurrences, we used the list of lexical modal markers elaborated in the framework of the WoPoss project (Dell'Oro (2019), https://woposs.unine.ch/). In Table 2 we present an overview of the markers retrieved in the corpus, with their basic modal value (see Section 2).

5 Network analysis

As announced above, we first performed an analysis of the phenomenon of co-occurrence in Classical Latin, and then we focused on the four different genres represented in the corpus, in order to see how the structure changes with respect to the variable of genre. For the sake of this extended abstract we will focus on some fundamental features and measures of the networks: the number of nodes, their type (□ or ◊) and the specific markers represented by the nodes; the edges between nodes and their weight; the edge count for each node, in order to determine the markers that co-occur the most and the least.

The network built on the co-occurrences for the whole corpus is shown in Figure 1; the networks for the four genres represented in the corpus are shown in 2, 3, 4 and 5. In the layout that we chose for illustrating the networks, squared nodes represent markers of necessity, whereas the diamonds represent markers of possibility. The edges are also customized: their width is proportional to the weight of the edge. The wider the edge, the higher the number of co-occurrences between the two nodes.

It is immediately evident from the structure of these five different networks that there are some differences at the structural level: some networks are bigger, in terms of number of nodes, than others (cf. Figure 1 with the other networks); some networks have a wider diameter (cf. Figure 4 and Figure 1); comparing the networks, the width of some edges shows a higher or lower weight value with respect to different genres.

We can use some network analysis measures in order to give a dimension to the differences that we can spot by simply looking at the images of the networks. In this extended abstract we will show some of these measures, which will be discussed more extensively in the paper[3].

[3] For the sake of this extended abstract we will show the raw frequencies of the co-occurrences and of the co-occurring markers. In fact, this type of value tells us how entrenched a certain type of co-occurrence is in the language represented by the corpus. However, in order to study the mutual

Table 2: Modal markers in the study and basic modal meaning.

Modal marker	Modal base
aeque	□
aequus/iniquus	□
aptus/ineptus	◊
certo/certe	□
certus/incertus	□
debeo	□
decet	□
dubius/dubium	◊
facultas	◊
forsitan	◊
fortasse	◊
forte	◊
ius est	□
licet	◊
licitus/illicitus	◊
meum est	□
necessarius/-um/-o	□
necesse est	□
necessitas	□
necessitudo	□
oportet	□
opus est	□
possibilitas	◊
possibiliter	◊
possum	◊
potestas	◊
probabilitas	◊
probabiliter	◊
queo/nequeo	◊
usus est	□
valet	◊

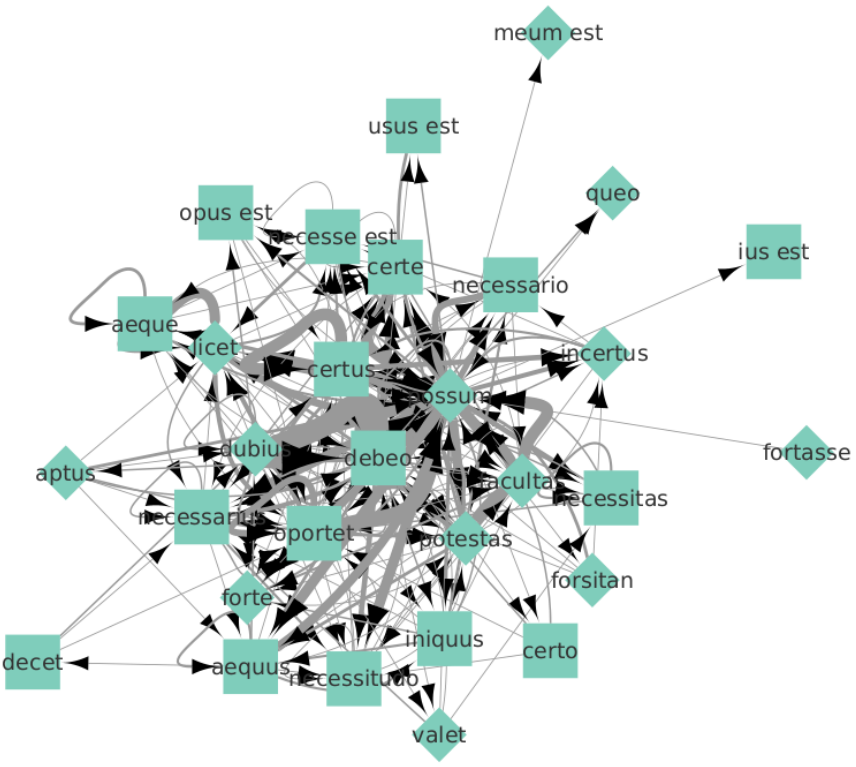

Figure 1: Co-occurrence network in the whole corpus.

5.1 Nodes

In Table 7 we show the five networks with respect to the number of nodes that they contain and the markers that the nodes represent. Some nodes (aequus, aptus, certe, certo, certus, debeo, dubius, facultas, forte, incertus, iniquus, licet, necessarius, necesse est, necessitas, oportet, possum and potestas) are represented in all four genres. We can conclude that their tendency to appear in co-occurrence with other modal markers does not change with respect to the variable genre. On the other hand, for some other nodes (decet, fortasse, ius est, meum est, opus est, usus est), the co-occurrence with other modal markers is only observed in one or two genres.

We also calculated the Edge Count for the nodes in each genre, in order to see which nodes were the most connected in the network i.e., were involved in the highest number of co-occurrences, and how this measure changed with respect to genre. The results are shown in Table 3. We can see that possum is the most connected node in the four different genres, which results in being the most central node for the four genres. Debeo is also particularly high in the ratings, being the second most connected node for three genres out of four, and still being classified third in the genre of historiography. We also calculated the least connected nodes in the networks, which are shown in Table 4.

relevance of the markers with respect to each other when they appear in co-occurrence, it is important to compare the raw frequency of each marker when it appears in co-occurrence with the absolute frequency of the same marker in the corpus (see Narrog (2009, 165-175) for a more complete discussion). In the paper, we will show both types of value for each measure, and we will reason about the different insights that raw frequencies vs. normalized frequencies can offer on the object of study.

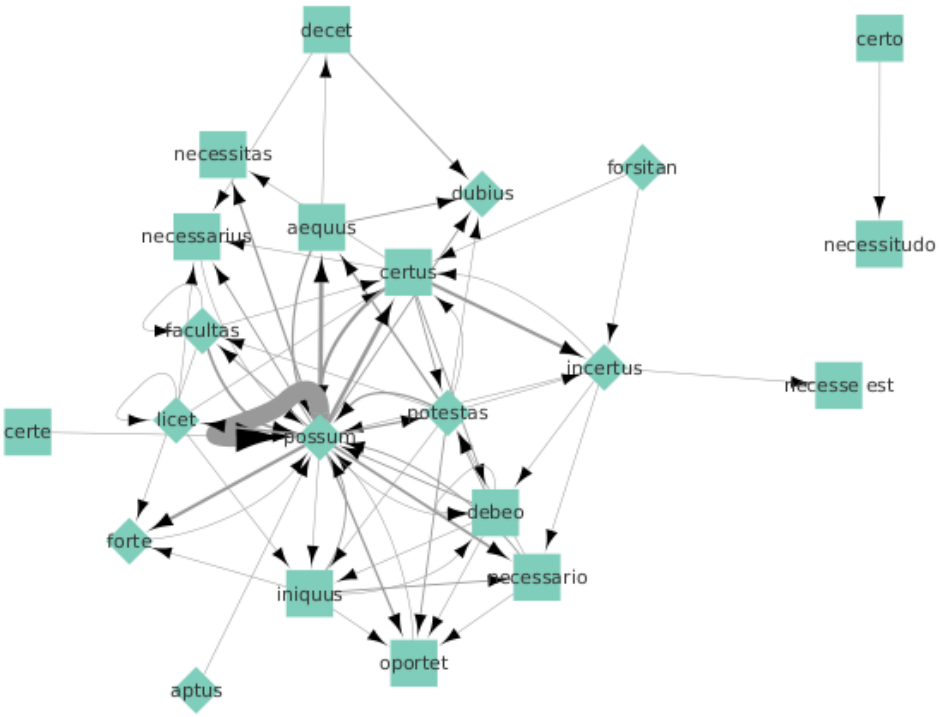

Figure 2: Co-occurrence network for the genre of historiography.

Table 3: Highest Edge Count for the different genres.

Letters		Historiography		Treatise		Oratory	
possum	40	possum	28	possum	37	possum	28
debeo	30	certus	12	oportet	26	debeo	15
certe	14	potestas	12	debeo	15	licet	13

Table 4: Lowest Edge Count for the different genres.

Letters		Historiography		Treatise		Oratory	
usus est	1	necessitudo	1	ius est	1	valet	1
opus est	1	necesse est	1	iniquus	1	meum est	1
fortasse	1	certo	1	decet	1	incertus	1

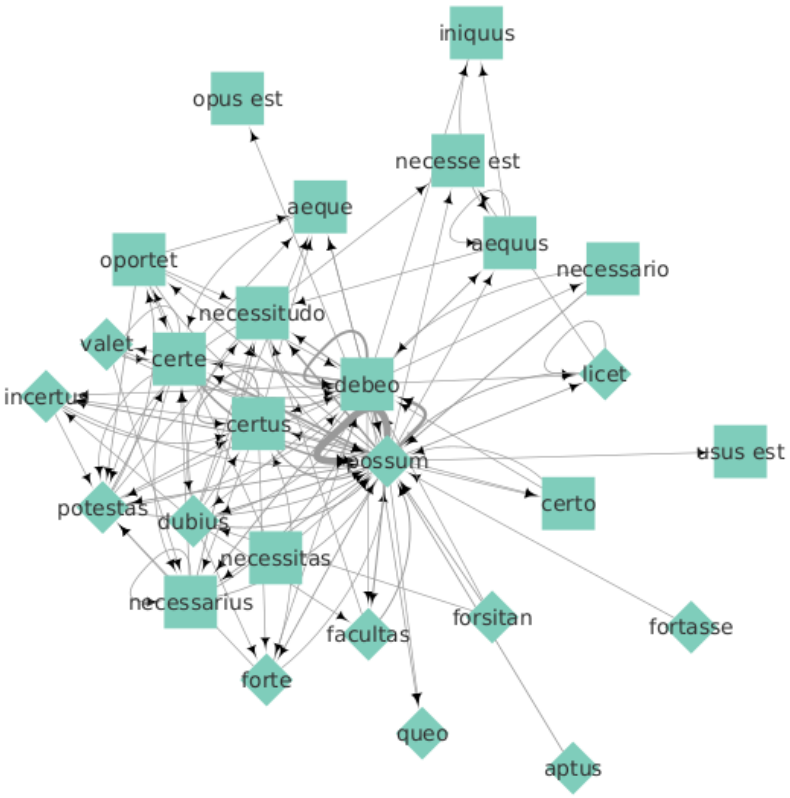

Figure 3: Co-occurrence network for the epistolographic genre.

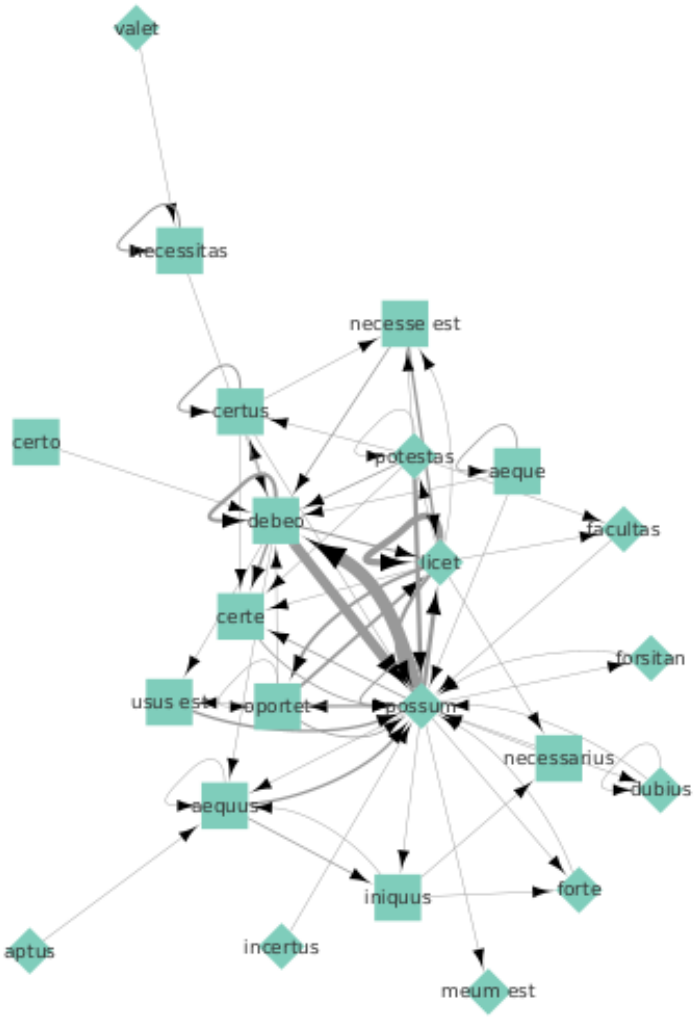

Figure 4: Co-occurrence network for the genre of oratory.

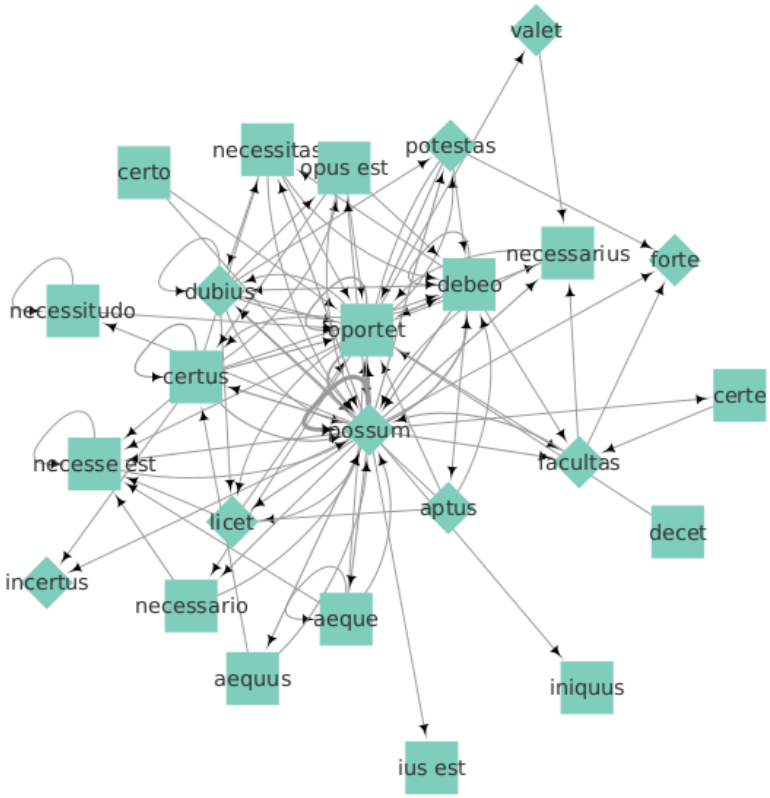

Figure 5: Co-occurrence network for the genre of treatise.

Table 5: Number of edges: whole corpus vs. single genres.

All texts	Letters	Historiography	Treatise	Oratory
247	113	67	95	67

Table 6: Co-occurrences: whole corpus vs. single genres.

All texts		Letters		Historiography		Treatise		Oratory	
possum-possum	247	possum-possum	64	possum-possum	30	possum-possum	52	possum-possum	91
debeo-debeo	29	debeo-possum	26	aequus-possum	7	oportet-possum	46	debeo-possum	21
certus-certus	16	possum-debeo	20	certus-possum	6	possum-oportet	15	possum-debeo	16

5.2 Edges

In Table 5 we show the number of edges for the whole corpus, compared to the number of edges for the single genre. We immediately notice that letters and treatises are more characterized by the phenomenon of the co-occurrence with respect to the other two genres. In order to see the details of the types of co-occurrence, we looked at the single edges between the nodes, and at their weight. The results are shown in Table 6. The measures for the whole corpus are already informative: the co-occurrence possum with possum, which translates into a self-loop in the networks, is by far the most represented type of co-occurrence in the corpus. We also find it as the most represented co-occurrence in the four genres, sometimes with great quantitative differences with respect to the other two most frequent co-occurrences (it is the case of historiography).

5.3 Other measures: network connectivity

An interesting feature of these networks is their high connectivity. The network connectivity measure is inversely proportional to the number of connected components in the network. A contributing factor being their small size, the number of connected components for these networks is always low. The network built based on the whole corpus data, the one for letters, treatises and oratory only count 1 connected component. The network for historiography counts 2 connected components, which can be easily spotted in Figure 2. The pair certo-necessitudo, connected by a directed edge, represents an isolated component with respect to the rest of the network. This is the only set of data in which the two markers only interact with each other and with no other marker in the dataset. This is especially interesting, considering that the two markers are present in almost the whole corpus: certo appears in all the genres in the corpus, and necessitudo in three of them (see Table 7).

Each node and type of edge will be dedicated a more thorough analysis in the paper. Moreover, other measures that will be discussed in the paper are degree centrality, betweenness centrality, clustering coefficient, closeness centrality. We will illustrate how these measures are able to describe the behavior of the markers in the corpus with respect to the phenomenon of co-occurrence.

6 Qualitative analysis

In this section, we will briefly discuss the type of information that is added during the qualitative analysis, and how it can be used in order to enrich the networks. The sentences that host the co-occurrence are further annotated by:

Table 7: Modal markers in co-occurrence: whole corpus vs. different genres.

All texts	Letters	Historiography	Treatise	Oratory
aeque	aeque		aeque	aeque
aequus	aequus	aequus	aequus	aequus
aptus	aptus	aptus	aptus	aptus
certe	certe	certe	certe	certe
certo	certo	certo	certo	certo
certus	certus	certus	certus	certus
debeo	debeo	debeo	debeo	debeo
decet		decet	decet	
dubius	dubius	dubius	dubius	dubius
facultas	facultas	facultas	facultas	facultas
forsitan	forsitan	forsitan		forsitan
fortasse	fortasse			
forte	forte	forte	forte	forte
incertus	incertus	incertus	incertus	incertus
iniquus	iniquus	iniquus	iniquus	iniquus
ius est			ius est	
licet	licet	licet	licet	licet
meum est				meum est
necessario	necessario	necessario	necessario	
necessarius	necessarius	necessarius	necessarius	necessarius
necesse est	necesse est	necesse est	necesse est	necesse est
necessitas	necessitas	necessitas	necessitas	necessitas
necessitudo	necessitudo	necessitudo	necessitudo	
oportet	oportet	oportet	oportet	oportet
opus est	opus est		opus est	
possum	possum	possum	possum	possum
potestas	potestas	potestas	potestas	potestas
queo	queo			
usus est	usus est			usus est
valet	valet		valet	valet
30	27	22	25	23

- Type of modality expressed by the two co-occurring markers;
- Type of syntactic structure in which the two markers co-occur;
- Connectors (e.g., sed 'but', et 'and' etc.)

These three types of information can be gathered and used as attributes. In the specific, the types of modality would be treated as attributes for the nodes in the network, whereas the type of syntactic structure and the connectors would become edge attributes. This would enhance the potential of applying the network analysis method to the study of modality, and would clustering on multiple levels: lexical, semantic and syntactic.

7 Acknowledgments

This work stems from my PhD thesis "Co-occurrence of modal markers in Latin: a quantitative and qualitative analysis", which is part of a wider project called A world of possibilities. Modal pathways over an extra-long period of time: the diachrony of modality in the Latin language (WoPoss), see https://woposs.unine.ch/. This project is funded by the Swiss National Science Foundation (SNSF n° 176778), and it is hosted at the Institut des sciences du langage (University of Neuchâtel).

References

Bolkestein, A. M. (1980). Problems in the description of modal verbs: an investigation of Latin. Assen: Van Gorcum.

Cecchini, F. M., M. Passarotti, P. Marongiu, and D. Zeman (2018). Challenges in converting the Index Thomisticus treebank into universal dependencies. Proceedings of the Universal Dependencies Workshop 2018 (UDW 2018).

Chu, X. (2008). Les verbes modaux du français. Paris: Ophrys.

Dell'Oro, F. (2019). Woposs guidelines for annotation. Zenodo, 1–28.

Fruyt, M. and C. e. Moussy (2002). Les modalités en latin: Colloque du Centre Alfred Ernout, Université de Paris IV, 3, 4 et 5 juin 1998. Paris: Presses de l'université de Paris-Sorbonne.

Huot, H. (1974). Le verbe devoir: étude synchronique et diachronique. Paris: Klincksieck.

Hütsch, A. (2020). L'usage des verbes modaux en français et en allemand. Étude contrastive de la combinatoire adverbiale sous l'éclarage quantitatif. Ph. D. thesis, Université de Neuchâtel. Faculté des Lettres et Sciences Humaines - Institut des Sciences du Langage et de la Communication.

i Cancho, R. F. (2005). The structure of syntactic dependency networks: insights from recent advances in network theory. In G. Altmann, V. Levickij, and V. Perebyinis (Eds.), The problems of quantitative linguistics. Chernivtsi: Ruta.

i Cancho, R. F. (2010). Network theory. In P. C. Hogan (Ed.), The Cambridge encyclopedia of the language sciences, pp. 555–557. Cambridge: Cambridge University Press.

Lyons, J. (1977). Semantics, Volume 2. Cambridge: Cambridge University Press.

Mehler, A. (2008). Large text networks as an object of corpus linguistic studies. In A. Lüdeling and M. Kytö (Eds.), Corpus linguistics. An international handbook of the science of language and society. Berlin, New York: Mouton De Gruyter.

Narrog, H. (2009). Modality in Japanese : the layered structure of the clause and hierarchies of functional categories. Amsterdam ; Philadelphia: John Benjamins Publishing Company.

Núñez, S. (1991). Semántica de la modalidad en latín. Granada: Servicio de publicaciones de la Universidad de Granada.

Nuyts, J. (2016). Analyses of the modal meanings. In J. Nuyts and J. van der Auwera (Eds.), The Oxford handbook of modality and mood. Oxford: Oxford University Press.

Nuyts, J. and J. Van Der Auwera (2016). The Oxford handbook of modality and mood. Chichester, UK: Oxford University Press.

Palmer, F. R. (2001). Mood and modality. Cambridge: Cambridge University Press.

Passarotti, M. (2015). What syntax can do for philosophy: a treebank-based network analysis of the verb sum in thomas aquinas. Rivista di filosofia neoscolastica 107(1-2), 309–324.

Passarotti, M. C. (2011). Language resources. the state of the art of latin and the Index Thomisticus treebank project. In Deuxiéme Colloque International ALIENTO, pp. 301–320.

Ponti, E. M. and S. Luraghi (2018). Non-configurationality in diachrony: Correlations in local and global networks of ancient greek and latin. Diachronica 35(3), 367–392.

Portner, P. (2009). Modality. Oxford: Oxford University Press.

Qi, P., Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning (2020). Stanza: A Python natural language processing toolkit for many human languages. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations.

Shannon, P., A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research 13(11), 2498–2504.

Sueur, J.-P. (1975). Etude sémantique et syntaxique des verbes devoir et pouvoir. Ph. D. thesis, Université de Paris X-Nanterre.