Software Tools for the Visualization of Definition Networks in Legal Contracts[*]

Michael Curtotti
Research School of Computer Science
Australian National University
Canberra, Australia
michael.curtotti@anu.edu.au

Eric McCreath
Research School of Computer Science
Australian National University
Canberra, Australia
eric.mccreath@anu.edu.au

Srinivas Sridharan
University of California
San Diego
California, United States
srsridharan@ucsd.edu

ABSTRACT

This paper describes the development of prototype software-based tools for visualizing definitions within legal contracts. The tools demonstrate visualization techniques for enhancing the readability and comprehension of definitions and their associated characteristics. This contributes to more accurate and efficient drafting or reading of contracts through the exploration of the meaning and use of definitions including via word clouds, multilayer navigation, adjacency matrix and graph tree representations.

Categories and Subject Descriptors

H.5.2 [**User Interfaces**]: Natural language; H.5.4 [**Hypertext and Hypermedia**]: Navigation; I.7.2 [**Document Preparation**]: Format and Notation; I.7.5 [**Document Capture**]: Document Analysis

General Terms

Human Factors

Keywords

definitions, legal contracts, word clouds, network visualization, contract visualization, text visualization, graph metrics

1. INTRODUCTION

This paper addresses the visualization of definition use within contracts. It is part of ongoing research on the development of software-based tools for reading and writing legal rules in contracts and legislation and aims to improve accessibility of legal documents and increase the efficiency and accuracy of legal rule creation [4, 5, 6].

This paper reports the development of prototype software tools demonstrating novel applications of visualizations for the representation and analysis of definition networks within contracts. The software tool enables a user to input text via a web interface and presents the user with a number of alternative visualizations of definitions in a contract: single layer pop-up hyper-linking of defined terms as they are used and representation of frequency and other information; application of 'word cloud' techniques to enable the rapid and global visualization of the 'usage' of a defined term and 'obfuscation' of a defined terms (metrics reflecting both the semantic content and graph theoretic role of the term); multi-layer hierarchical navigation tools enabling in-situ navigation of 'definition networks' from the rule where a definition is used; visual presentations of definitions as a link and node graph; and matrix representation of definition usage within a contract.[1] In Section 6 below we describe these visualizations further.

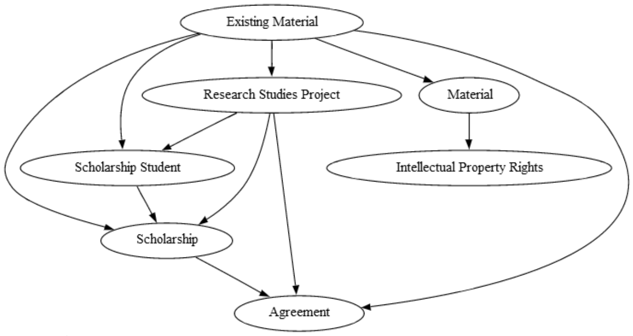


Figure 1: A node-link graph diagram showing the relationships between defined terms which have been extracted from a natural language contract.

Contracts are semi-structured documents, and are usually explicitly organized in a tree-like structure consisting (primarily) of rules and sub-rules. In the Australian case these structures are typically referred to as 'clauses' and 'sub-clauses' with each clause ideally addressing a discrete topic. Definitions typically occur as a small glossary or dictionary embedded within a single 'definition' clause.

Definitions form a substantial part of typical contract texts and are used by drafters to control meaning and presentation of text. In this paper we use 'defined term' to refer to the definition label and 'defining text' to refer to the natural language which expresses the meaning of the defined term. 'Definition' refers to the entire structure. While formally definitions are intended to simplify drafting, they can also be used in larger contracts as a tool to modify meaning in the favour of the drafter's client in ways that become increasingly difficult to analyze for the other party as the complexity of definitional relationships increases. Such definitions can also result in

[*] A colour print version of the paper is available at: http://cs.anu.edu.au/people/Michael.Curtotti/papers/icail2013.pdf

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the Owner/Author. Copyright is held by the owner/author(s).

ICAIL '13 Jun 10-14 2013, Rome, Italy
ACM 978-1-4503-2080-1/13/06

[1] http://buttle.anu.edu.au/contracts/

meaning being 'hidden', as meaning may not be apparent from the surface text of a legal rule. Because of their complexity, such structures can also result in errors such as inconsistency of meaning in a hierarchy of definitions.

This paper is organized as follows. Section 2 reviews relevant literature. Section 3 describes the extraction of definitions from a contract document. Section 4 discusses the representation of definitions and their relationships as networks. Section 5 briefly outlines the data and tools used in undertaking this work and briefly canvasses analysis of that data. Section 6 presents prototype visualizations exploiting the network characteristics associated with definitions. We present our conclusions in Section 7.

2. RELATED WORK

There are four areas of related work we wish to describe: work relating to the study of contracts at the broadest level; natural language processing for the extraction of definitions; information and graph visualization and studies in relation to Word Clouds.

2.1 Studying Contracts

Contracts are studied from a wide range of perspectives and disciplines. The principles for interpreting contracts as sources of legal rules is an extensively studied domain. Contracts have also been widely studied from the point of view of economic and social theory [19]. Work more directly relevant to the high level aim of creating software-based tools to enhance the reading and writing of contracts is also potentially broad. Curtotti et al. [4, 5] review work including in the field of machine learning, the logical representation of legal rules, e-contracts and studies of corpora of contracts.

2.2 Natural language processing for the extraction of definitions

Work on the application of natural language processing to definitions in general text is extensive, however a considerable part of this work is dedicated to extraction of definitions from unstructured general prose. It thus addresses a more complex and difficult problem than that of extraction of definitions from semi-structured texts, such as contracts. Degorski et al. [9] apply enhancements to machine learning for definition extraction from unstructured text. Others employ rule based approaches for the extraction of dictionaries from text [17]. Winkels et al. [20] and Maat et al. [7] report work in the parallel legislative domain including definition extraction, in the context of Dutch legislation. Definition extraction remains an active area of research with a view to improving precision and recall of such definition extraction [18].

2.3 Information and Graph Visualization

Information visualization is centred on the users of data and is concerned with the representation of complex data in ways that facilitate its comprehension. Information visualization employs graphical presentations of data to exploit the visual capacities of users in identifying patterns and relations in data. It is used in text mining and may provide advantages such as the ability to display a large amount of data at once, enhance identification of relationships and clustering in data, provide interactivity to users or allow users to move from micro to macro quickly [10, pp 190 et seq]. Some forms of such data visualization are commonly known (e.g. histograms and line graphs). Others are more recent technologies developed for the visualization of large data sets. Concept set graphs are a commonly used tool in text mining showing hierarchical relationships between concepts. Graphs may also show the network of associations between concepts or entities found in texts, and the weight of those associations. Circle graphs can be used to show the strength of multiple associations between terms. A plethora of more complex visualizations have also been employed including self organising maps, hyperbolic trees and fisheye diagrams[10, pp 194 et seq]. Among the variables that can be adjusted to enhance graph visualization are layout (including tree layout, 3D representation, spring layout, space division and matrix layout), clustering, sampling or filtering for large graphs, zooming and panning, animation, focus plus context [3]. A site that provides both a software tool for a range of common visualizations and demonstrations of their application is the Java Infovis toolkit site.[2] Visualizations are commonly employed in the field of network analysis (including for example analysis of social networks). Most commonly as the node and link diagram used in graph theory, but enhanced with information describing the entities and relationships represented by nodes and links. An alternative representation also employed in social network analysis are adjacency matrices which provide a two dimensional array representing nodes and associations (or strength of association) between them [pp 4 et seq and pp 259 et seq][8].

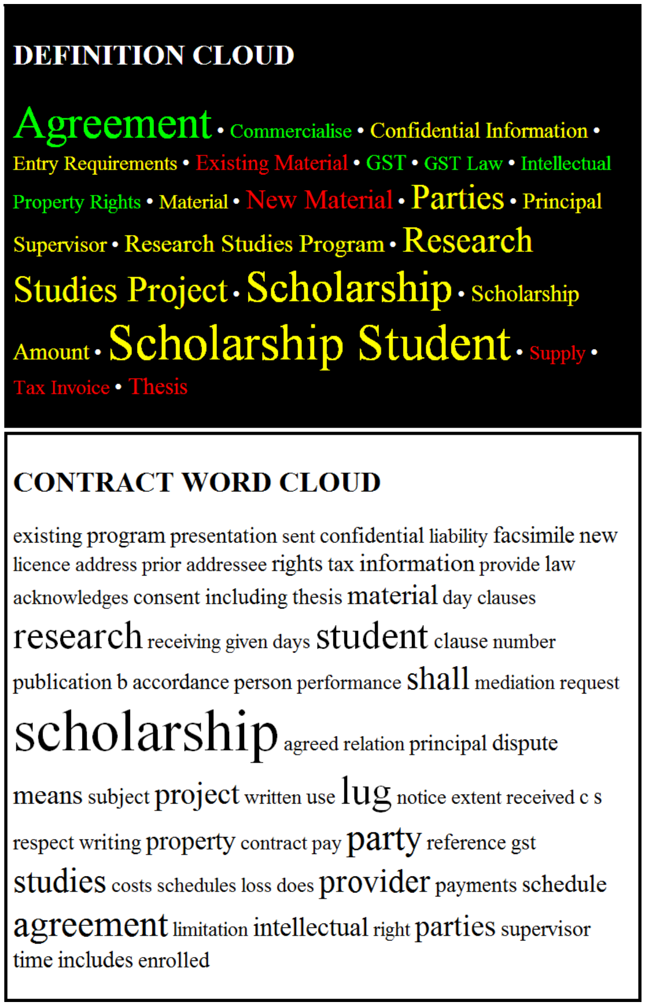

DEFINITION CLOUD
Agreement • Commercialise • Confidential Information • Entry Requirements • Existing Material • GST • GST Law • Intellectual Property Rights • Material • New Material • Parties • Principal Supervisor • Research Studies Program • Research Studies Project • Scholarship • Scholarship Amount • Scholarship Student • Supply • Tax Invoice • Thesis
CONTRACT WORD CLOUD
existing program presentation sent confidential liability facsimile new licence address prior addressee rights tax information provide law acknowledges consent including thesis material day clauses research receiving given days student clause number publication b accordance person performance shall mediation request scholarship agreed relation principal dispute means subject project written use lug notice extent received c s respect writing property contract pay party reference gst studies costs schedules loss does provider payments schedule agreement limitation intellectual right parties supervisor time includes enrolled

Figure 2: Definition cloud and comparative word cloud.

[2] thejit.org

2.4 Word Clouds

Word clouds are a form of information visualization that has become popular in recent years as a way of summarising and visualizing key concepts in a large body of text. Word clouds support functions such as browsing, searching, subject description and formation of an impression concerning the data. A key technique in word clouds is the manipulation of the visual features of text (font, area, width, intensity, colour) and their location within the cloud to suggest importance or other features of the word. Bateman et al. [1] find that font size and weight has a particular effect. Colour can also influence interaction but is ambiguous in its meaning. Position also has an influence. Accordingly they endorse the use of the former while suggesting that colour and position be used with care. Lohmann et al. [16] specifically study the effect of tag position or layout on the effectiveness of certain user tasks such as identification of popular terms, search for particular terms and identification of topics in the word cloud. Based on studies of user interaction with different layouts they do not find a best way to layout a cloud but observe that large tags (font size) are readily identified as 'popular'. They confirm findings by other authors that centering of 'popular' tags within a cloud assists their identification. This effect they find most pronounced with a circular tag layout. They also find the top left quadrant of a word cloud attracts the most attention. Word clouds are not well suited for searching. Halvey et al. [14] also find that font size and position are important, although they note that alphabetical presentation is an aide to finding information.[3]

3. EXTRACTION OF DEFINITIONS FROM CONTRACTS

To visualize definitions in a contract it is necessary to first extract them and clauses from the contract. We use relatively trivial regular expressions which are applied in three stages: (1) identification and segmentation of the definition clause and other clauses in the text of the contract; (2) segmentation of definitions from each other; and (3) extraction of the defined term and its defining text from a definition. By requiring users to apply simple rules which are widely used in Australian industry practice, such as ensuring definitions ends with a full stop and using standard 'key words' for definition relations (particularly the words 'means' and 'includes') essentially 100% accuracy can be attained on typical contract texts. This result expresses that a realistic and readily attainable solution (minor user editing) can effectively address the accuracy problem which is difficult to fully solve using entirely computational methods. Changes necessary to improve accuracy are easily exposed to the user through a web page and are implementable with a few key strokes. Previous research by the authors using fully automated methods of machine learning, hand crafted rules or hybrid methods reached accuracies of around 80% to 82%, a level of accuracy inadequate to the legal domain [4].

4. REPRESENTATION AS NETWORKS

A graph is an ordered tuple *G = (V, E)* of a set of vertices (or nodes) *'V'* and edges *'E'* between them. An edge links two vertices v_1 and v_2, and may either be directed or undirected [2, p348]. The number of edges associated with a vertex is referred to as its 'degree'. In the case of a directed graph, the in-degree of a vertex is the number of incoming arcs to the vertex. Its out-degree is the number of arcs emerging from the vertex [2, p348 et seq]. We follow de Nooy et al. in defining a 'network' as a graph which has additional information associated with its vertices and edges, i.e. information beyond the simple structural characteristics of nodes and links [8, p7]. Definitions and the clauses in which they occur are thus represented as the vertices (or nodes) of a network. Links between the vertices represent either the occurrence of a defined term in a clause, or the occurrence of a defined term in another definition. The nodes and links form directed graphs which can be analysed, including from a graph theoretic viewpoint to reveal information about a legal document. Figure 1 shows an example of a network between definitions in a contracts.

5. DATA, TOOLS AND ANALYSIS

The work reported in this paper is based on analysis of a set of ten contracts drawn from a corpus of 249 Australian contracts consisting of in the order of 10^6 words. The corpus has been compiled from Australian contracts and contract drafts available on the web. Curtotti et al. [5] report the profiling and analysis of an earlier version of the corpus. The current version of the corpus has been subjected to further data cleaning but is substantively the same as reported above.[4]

The tools used in carrying out the work reported here included project code for analysis and visualization (including online) which is written primarily in python and javascript and libraries for graph analysis or graph visualization (networkX and Graphviz). We also utilize Canviz for web visualization of graphviz output.[5]

On average, definitions represent 17.4% of the core text of contracts in the sub-corpus and these contracts on average used definitions 304 times. These latter results illustrate the significance of definition text as a component of such legal documents.

Definition networks in our sub-corpus had on average a degree of 2.35 with a standard deviation of 2.47. This relationship between average and standard deviation would lead us to anticipate that degree is log-normally distributed [15]. This is in fact a reasonable description of the distribution.

Length is often used as a simple measure of complexity: the longer a definition, the more complex it is likely to be [13]. We found the correlation between out-degree and definition length to be low to moderate with a value of 0.325 over 223 data points (individual definitions). Degree then provides a different indicator of complexity to length.

In some of the visualizations described below we employ a recursive out-degree related measure to represent 'hidden' meaning in a definition. This measure is derived from the overall length of all text recursively referenced through the outward links of a definition. This is effectively a recursive weighted out-degree measure, where length is a measure of weight of the parent and successor nodes.

[3] We note that there is work in the parallel domain of visualization of legislation. Due to limitations of space we do not canvas that work here, but refer the interested reader to *The Visualization of Law*, Curtotti and McCreath [6]. Also work on visualizing contract provisions using non-computerized methods has been undertaken by Haapio and Passera [11, 12].

[4] A copy of the corpus is obtainable by contacting the authors via `http://cs.anu.edu.au/people/Michael.Curtotti/`

[5] `http://code.google.com/p/canviz/`

6. VISUALIZATIONS

We report a prototype website demonstrating a web based tool for the extraction and visualization of definition structures from submitted contract texts.[6] A user is able to load the demonstration text or submit a contract conforming to the requirements of the tool and may visualize definition structures within the contract by selecting one of four visualization options.

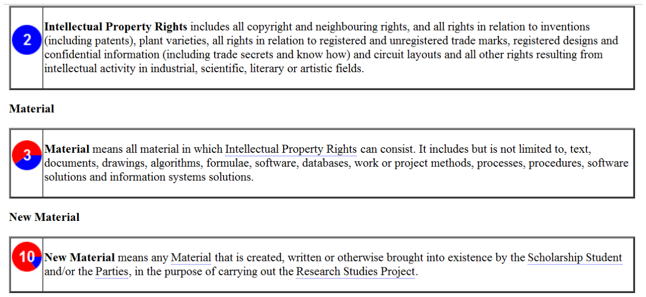
Cloud Visualisations: Figure 2 illustrates a cloud presentation of two measures of definition characteristics: their frequency of use in a contract (usage) (visualized through font size), and how much of the meaning of a defined term is 'hidden' or 'obfuscated' through the referencing of other defined terms by a defining text. Red, yellow, green is a well recognized 'traffic light' representation suggestive of levels of risk and is used in this context to indicate risk prone definition relationships with red suggesting significant hiding of meaning, yellow moderate hiding and green a low proportion of hidden text. In black and white print the colours appear as dark grey, light grey and gray respectively.

The scaling of font size makes it relatively straight forward to determine the probable purpose of the document from which the definitions are drawn as significant terms are emphasised. A more traditional word cloud is provided from the same contract for the purposes of comparison. In this word cloud word font size is a function of word frequency.

Use Case: Such a visualization allows a reader to form an immediate impression of the importance of terms, where complex layered meaning may be hidden and the probable nature of the contract.

In Situ Usage and Obfuscation: Similar information to that conveyed via the Definition Cloud is provided by in situ presentations using a number (to directly represent usage) and a small pie chart icon to represent 'obfuscation'. In this case the ratio of the pie shown in red represents the relative length of hidden text associated with the definition. (See Figure 3)

Use Case: As above.



2 **Intellectual Property Rights** includes all copyright and neighbouring rights, and all rights in relation to inventions (including patents), plant varieties, all rights in relation to registered and unregistered trade marks, registered designs and confidential information (including trade secrets and know how) and circuit layouts and all other rights resulting from intellectual activity in industrial, scientific, literary or artistic fields.

Material

3 **Material** means all material in which Intellectual Property Rights can consist. It includes but is not limited to, text, documents, drawings, algorithms, formulae, software, databases, work or project methods, processes, procedures, software solutions and information systems solutions.

New Material

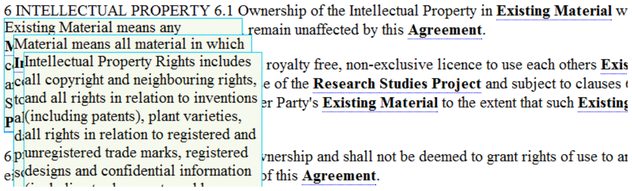
10 **New Material** means any Material that is created, written or otherwise brought into existence by the Scholarship Student and/or the Parties, in the purpose of carrying out the Research Studies Project.

Figure 3: Usage and Obfuscation metrics

Single and Multilayer in-situ Definition Graph Navigation: In most contracts, the only means of navigating to the meaning of a defined term when encountering it in a clause is to scroll to the definition clause (typically at the top of the document), read the definition and scroll back down.

We provide both a visualization that allows single layer access to the meaning of a defined term and navigation through multiple layers of definition referencing. The latter visualization enables a user, from the rule being read, to navigate through the entire tree of definitions referenced by the rule, following the conceptual links between defined terms. By double clicking the definition window the user can cause the pop up to disappear, returning to the original rule.

Use Case: Such a facility is likely to aid both comprehension and increase efficiency of contract reading. It reduces time necessary to access related meanings and anchors the reading experience in the rule itself.



6 INTELLECTUAL PROPERTY 6.1 Ownership of the Intellectual Property in **Existing Material** w
Existing Material means any
remain unaffected by this **Agreement**.
Material means all material in which
Intellectual Property Rights includes
all copyright and neighbouring rights,
and all rights in relation to inventions
(including patents), plant varieties,
all rights in relation to registered and
unregistered trade marks, registered
designs and confidential information
royalty free, non-exclusive licence to use each others **Exis**
e of the **Research Studies Project** and subject to clauses
er Party's **Existing Material** to the extent that such **Existing**
vnership and shall not be deemed to grant rights of use to a
of this **Agreement**.

Figure 4: Tool for multi-layer navigation of defined terms from rule where the defined term is used. Here we navigate the terms 'Existing Material' – 'Material' – 'Intellectual Property Rights' from the clause of the contract dealing with those rights.

Matrix Representation of Bimodal Definition Use Graph: Figure 5 provides a representation of the relationship between clauses and definitions as a weighted bimodal adjacency matrix. Such representations are used in social network analysis,[8] but their application to legal documents is novel.

Each square in the matrix represents a definition-clause relationship and the darkness of the square indicates the relative frequency with which a defined term is used in a particular clause. A column provides a visual summary of the importance of definitions used within a particular clause, while a row summarises the use of a particular definition across the agreement.

Use Case: The bimodal representation provides a potential tool for visualizing the semantic structure of a contract in summary form.

Definition Graphs: Figure 1 is an example of standard node link graph diagram representing a definition network. It shows the relationships between a definition and the defined terms it uses. The visualization provides an immediate sense of the relationship between defined terms. It intuitively represents the complexity of definition use, providing an opportunity to a drafter to consider revision to reduce complexity, or to a reader to explore concepts utilised by a rule. A reader is similarly alerted to semantic relationships. Simple inspection reveals any cycles that may be present in the definition graph. Cycles may represent logical errors or conceptual complexity in the ideas represented by the definition. Graphs of this kind can equally be generated with a rule as the root node of the representation. Although the adjacency matrix visualization provides an indication of 'weight', it only indicates relationships of a clause or rule with definitions to a depth of 1 (i.e. those directly used in the clause text). A directed node link diagram enables the relevant definition network to be explored in full.

Use Case: Provides a graphical representation of the semantic struc-

[6] http://buttle.anu.edu.au/contracts/

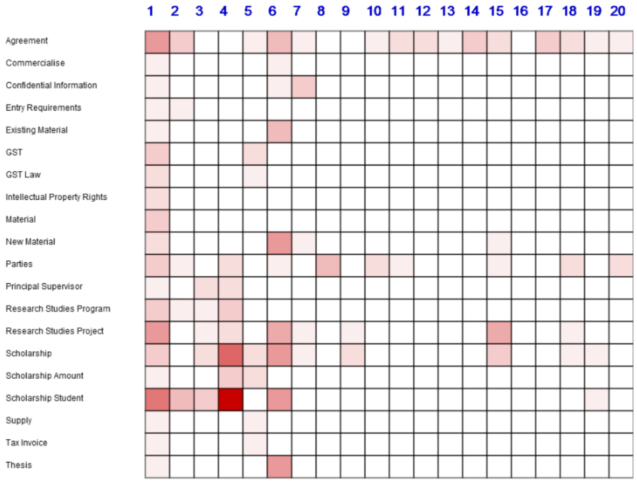


Figure 5: A matrix representation of the relationships between definitions and clauses. Extracted using Pajek.

ture of key terms in a contract, assisting readers in understanding semantic relationships and drafters in removing potential errors or simplifying how terms are defined.

7. CONCLUSIONS AND FUTURE WORK

In this paper we present work related to the visualization of definition networks. We describe definition usage in contracts and present a number of prototype visualizations of definitions (including visualization of network attributes and selected metrics). Methods widely employed outside the legal field (such as word clouds) show promise for application within the legal field in connection with facilitating comprehension of definition use in contracts. Navigational enhancements such as multi-layer pop up for definition navigation show the potential to facilitate access to the meaning of definitions within the context of rules in which they are employed increasing comprehension and efficiency in contract reading. Tools such as node-link diagrams facilitate an exploration of semantic trees embedded in definition networks. Presentation of metrics associated with the definition network help readers assess the significance and risk of defined term usage.

8. REFERENCES

[1] S. Bateman, C. Gutwin, and M. Nacenta. Seeing things in the clouds: the effect of visual features on tag cloud selections. In *Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*, pages 193–202. ACM, 2008.

[2] I. Bronstein, K. Semendyayev, G. Musiol, and H. Muehlig. *Handbook of Mathematics Fifth Edition*. Springer, 2007.

[3] W. Cui. A survey on graph visualization. *PhD Qualifying Exam (PQE) Report, Computer Science Department, Hong Kong University of Science and Technology, Kowloon, Hong Kong*, 2007.

[4] M. Curtotti and E. McCreath. Corpus Based Classification of Text in Australian Contracts. In *Proceedings of the Australasian Language Technology Association Workshop 2010*, 2010.

[5] M. Curtotti and E. McCreath. A corpus of australian contract language. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and the Law 2011*, 2011.

[6] M. Curtotti and E. McCreath. Enhancing the visualization of law. In *Law via the Internet Twentieth Anniversary Conference, Cornell University*, 2012.

[7] E. de Maat and R. Winkels. Automatic classification of sentences in dutch laws. In *Proceedings of the 2008 conference on Legal Knowledge and Information Systems: JURIX 2008: The Twenty-First Annual Conference*, pages 207–216. IOS Press, 2008.

[8] W. de Nooy, A. Mrvar, and V. Batagelj. *Exploratory Social Network Analysis with Pajek*. Cambridge University Press, 2005.

[9] Ł. Degórski, M. Marcinczuk, and A. Przepiórkowski. Definition extraction using a sequential combination of baseline grammars and machine learning classifiers. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation, LREC 2008*, 2008.

[10] R. Feldman and J. Sanger. *The Text Mining Handbook Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press, 2007.

[11] H. Haapio. Contract clarity through visualization–preliminary observations and experiments. In *Information Visualisation (IV), 2011 15th International Conference on*, pages 337–342. IEEE, 2011.

[12] H. Haapio and S. Passera. Reducing contract complexity through visualization-a multi-level challenge. In *Information Visualisation (IV), 2012 16th International Conference on*, pages 370–375. IEEE, 2012.

[13] J. Hagedoorn and G. Hesen. Contractual complexity and the cognitive load of r&d alliance contracts. *Journal of empirical legal studies*, 6(4):818–847, 2009.

[14] M. Halvey and M. Keane. An assessment of tag presentation techniques. In *Proceedings of the 16th international conference on World Wide Web*, pages 1313–1314. ACM, 2007.

[15] E. Limpert, W. A. Stahel, and M. ABTT. Log-normal Distributions across the Sciences: Keys and Clues. *BioScience*, 51(5), 2001.

[16] S. Lohmann, J. Ziegler, and L. Tetzlaff. Comparison of tag cloud layouts: Task-related performance and visual exploration. *Human-Computer Interaction–INTERACT 2009*, pages 392–404, 2009.

[17] S. Muresan and J. Klavans. A method for automatically building and evaluating dictionary resources. In *Proceedings of the Language Resources and Evaluation Conference*, volume 1, page 30, 2002.

[18] R. Navigli and P. Velardi. Learning word-class lattices for definition and hypernym extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1318–1327. Association for Computational Linguistics, 2010.

[19] J. Paterson, A. Robertson, and P. Heffy. *Principles of Contract Law*. Lawbook Co., 2nd edition, 2005.

[20] R. Winkels and R. Hoekstra. Automatic extraction of legal concepts and definitions. In *Legal Knowledge and Information Systems: JURIX 2012: the Twenty-Fifth Annual Conference*, volume 250, page 157. IOS Press, 2013.