YAG research idea form for Interdisciplinary PhD projects

Project title	**Automatic detection of linguistic patterns in legal big data**
Intended supervisors	1. Prof. Michel Vols (Faculty of Law) 2. Dr Martijn Wieling (Faculty of Arts)
Supervisors' disciplines	Vols: Faculty of Law, Public order law Wieling: Faculty of Arts, Computational linguistics
A. Description of research idea (500 words maximum)	Law is everywhere: almost every human activity is regulated. Buying a sandwich, renting an apartment or going to a hospital, all these activities involve legal rules and consequences. For a stable and sustainable society it is essential that its laws are predictable. People need to know what a legal rule means and what likely outcome a potential court case will have. Authorities have tried to improve the law's predictability and transparency by publishing court judgments. For decades, summaries of judgments were published in written journals, which were not easily accessible for the public. Nowadays, courts publish their judgments online. For example, approximately 370,000 individual court judgments can be found on the website of the Dutch judiciary (www.rechtspraak.nl). Similarly, another 52,000 court judgments are published online by the European Court of Human Rights (http://hudoc.echr.coe.int). Each of these judgments contains a detailed and rich description of the facts, procedure, reasoning of the parties and outcome of the case. Of course, public availability of case law will help to improve predictability and transparency, but to analyse hundreds of thousands of legal documents, we need other approaches than the traditional and labour-intensive 'doctrinal analysis' (i.e. close reading of a single or a small number of judgments) conducted by legal researchers. The goal of this PhD project is therefore to combine two distinctly separate disciplines, law and computational linguistics, in developing and evaluating quantitative, computational approaches to improve the predictability and transparency of the law. Techniques from computational linguistics would (for example) enable the automatic extraction, syntactic and semantic analysis of the judgment texts. The extracted features may be used in quantitative analyses identifying common patterns (see Wieling, 2012 for examples, albeit in a different field), which can subsequently be used to predict the outcome of a judgment. Such an approach would clearly be beneficial for the field of law. Surprisingly, a recent study (Vols & Jacobs, 2016) showed that between 2006 and 2016 fewer than 25 publications in Dutch legal journals were published involving statistics to analyse case law. While a quantitative approach to analysing case law is more prevalent in the US, it's primarily focused on specific American legal issues, and frequently contains serious methodological flaws (Epstein & King, 2002; Epstein & Martin, 2014).

	In computational linguistics, specific characteristics of legal texts have been studied (see Francesconi, Montemagni et al. (eds.), 2010), but hardly any studies have attempted to use linguistic characteristics to predict judicial decisions. A very recent exception (also illustrating the timeliness of the project idea) by Aletras et al. (2016) reported an accuracy of 79% in predicting the judgments of the European Court of Human Rights. However, they focused only on a small sample (600 judgments) and used simplistic linguistic features (such as word frequency). The goal of this PhD project is therefore to take a more comprehensive, linguistically-oriented approach incorporating all available data, thereby developing a system which is able to detect common patterns in legal big data and use these to predict the outcome of a judgment.
B. Explanation of interdisciplinary nature of the project (200 words maximum).	From the description of the research idea it should be clear that this project is interdisciplinary. Tools from the field of (computational) linguistics will be used and further developed in order to detect patterns texts originating from the field of law. While the benefit of such an approach for the field of law is immediately obvious (and discussed in the description of the research idea, above), there are also clear benefits for the field of computational linguistics. First of all, by using a large source of legal texts, which has not been previously analysed, this project will help inform research on identifying characteristics of legal texts (Francesconi, Montemagni et al. (eds.), 2010). Furthermore, existing tools from computational linguistics will be adapted in such a way that their performance on legal texts will be improved, making their applicability more general.
C. Expected candidate's profile (200 words maximum).	The ideal PhD candidate should have an interest in the law and adequate technical (i.e. programming) and quantitative skills. Being able to write computer code and running quantitative analyses will be essential for the success of this project. The skill level in programming and/or quantitative analysis may be improved by following courses in the first year of the PhD, but a hard requirement is that the candidate does have some initial skill in these areas and feels confident in improving these skills to the level needed to successfully finish this type of project. Having a formal background in law is not essential, but in that case the candidate will be required to follow one or more introductory law courses during the first year of the PhD. Of course, an interest in legal issues will be necessary to make the project a success. Candidates who are especially invited to apply include students in computer science, computational linguistics, or artificial intelligence with an interest in the computational linguistic analysis of legal texts, or students of law with a (basic) demonstrable skill in programming and quantitative analysis.