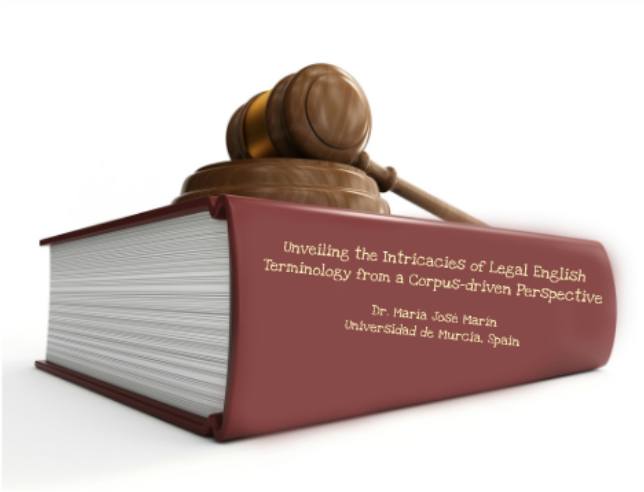
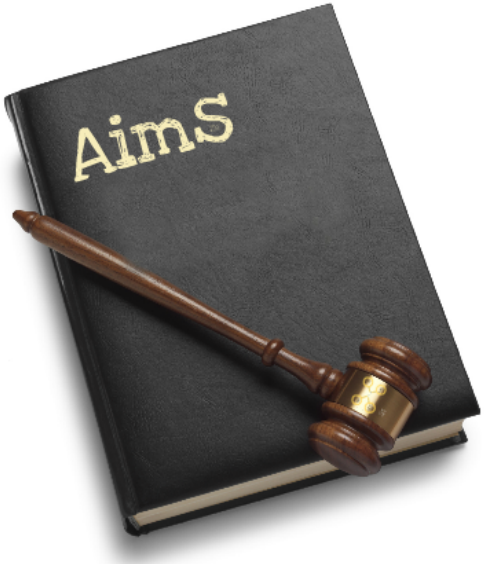
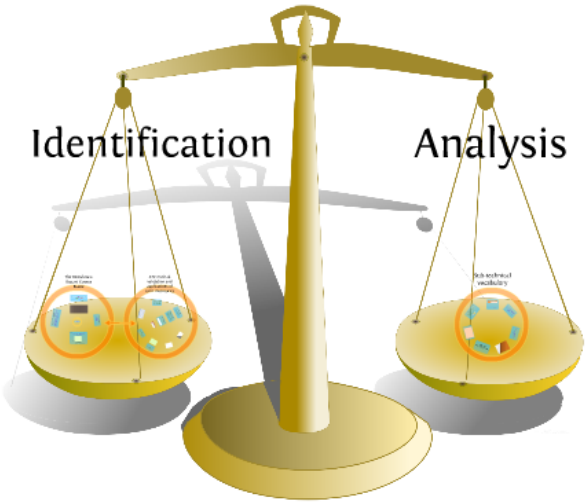
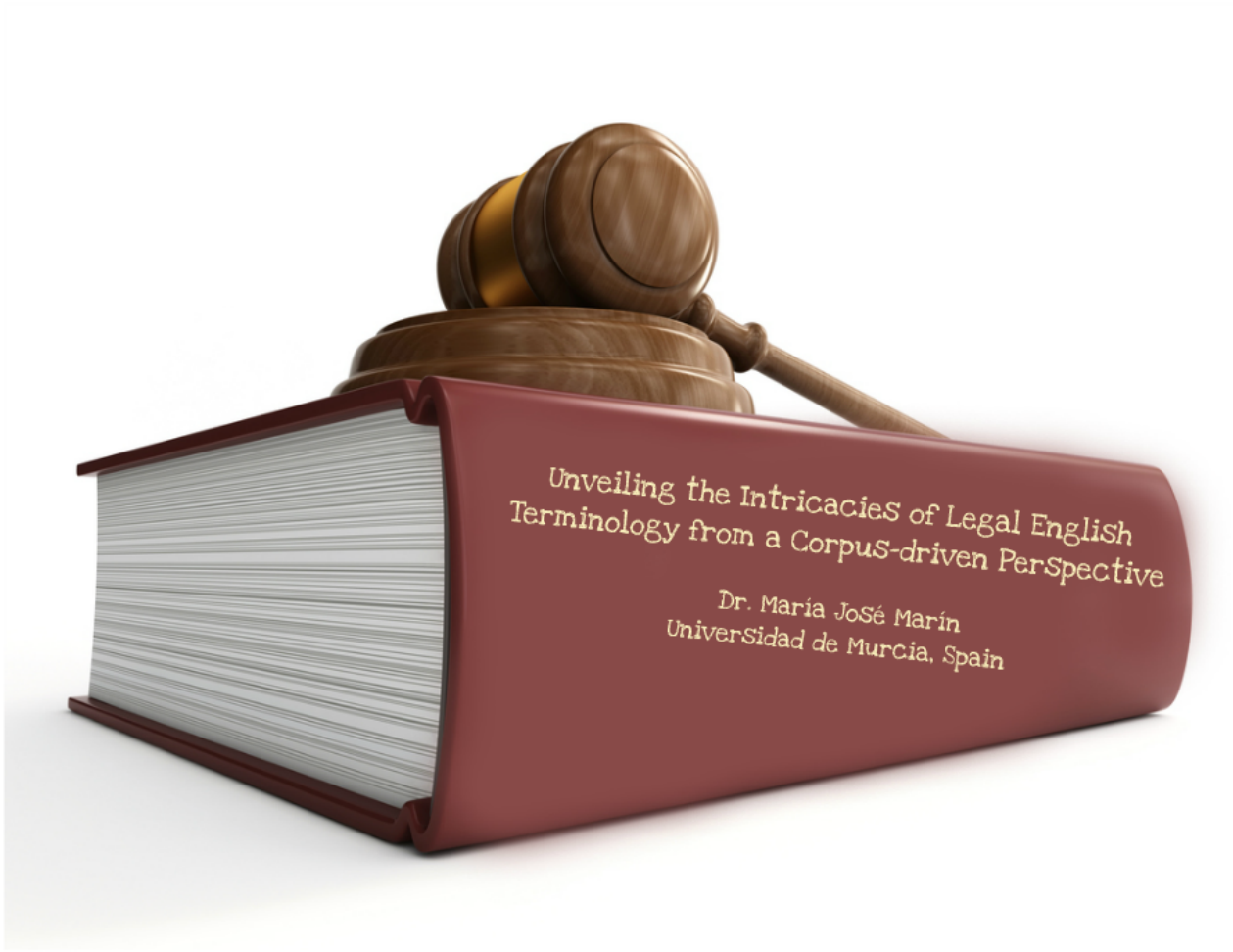
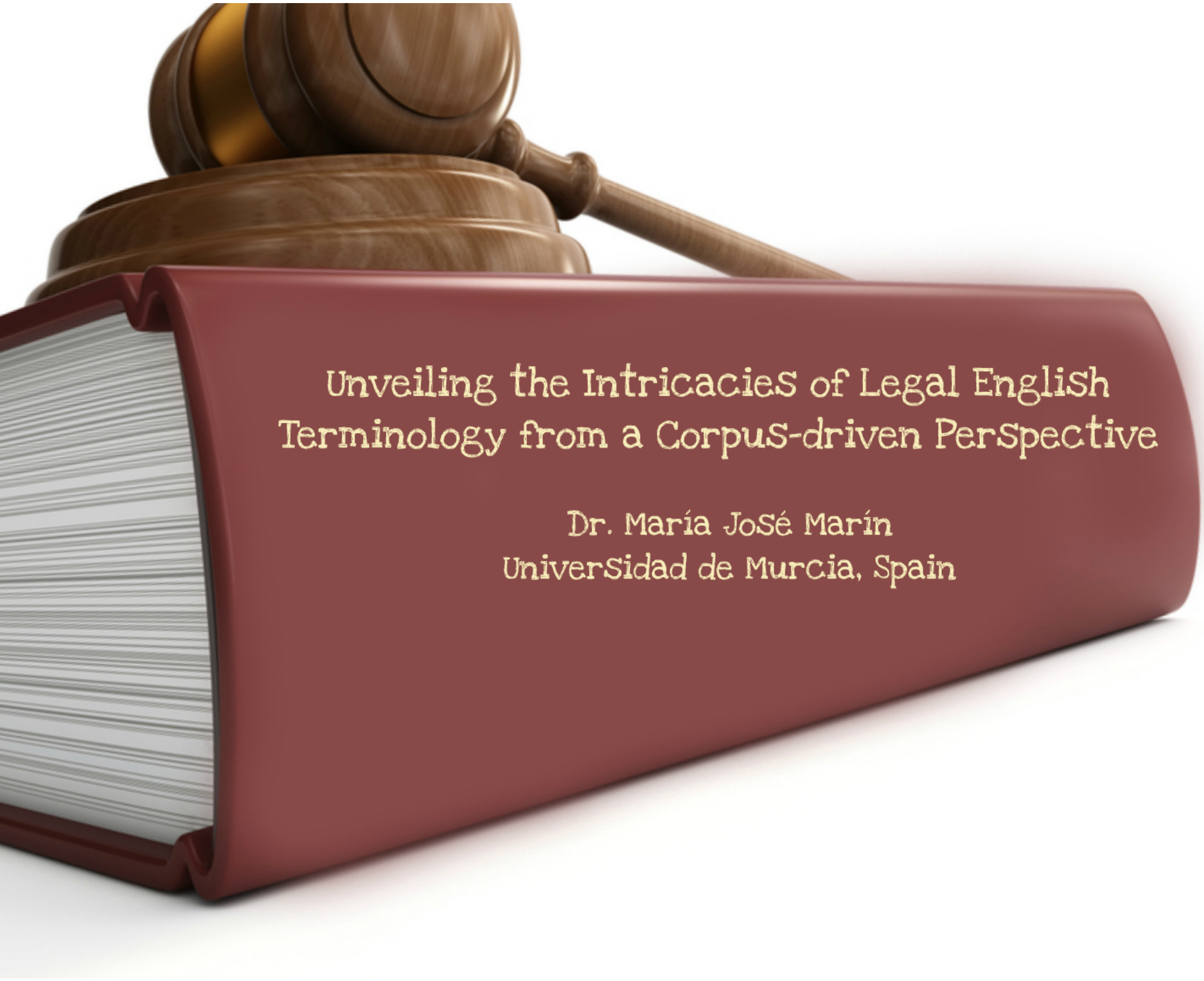

Unveiling the Intricacies of Legal English Terminology from a Corpus-driven Perspective
Dr. María José Marín
Universidad de Murcia. Spain


AimS

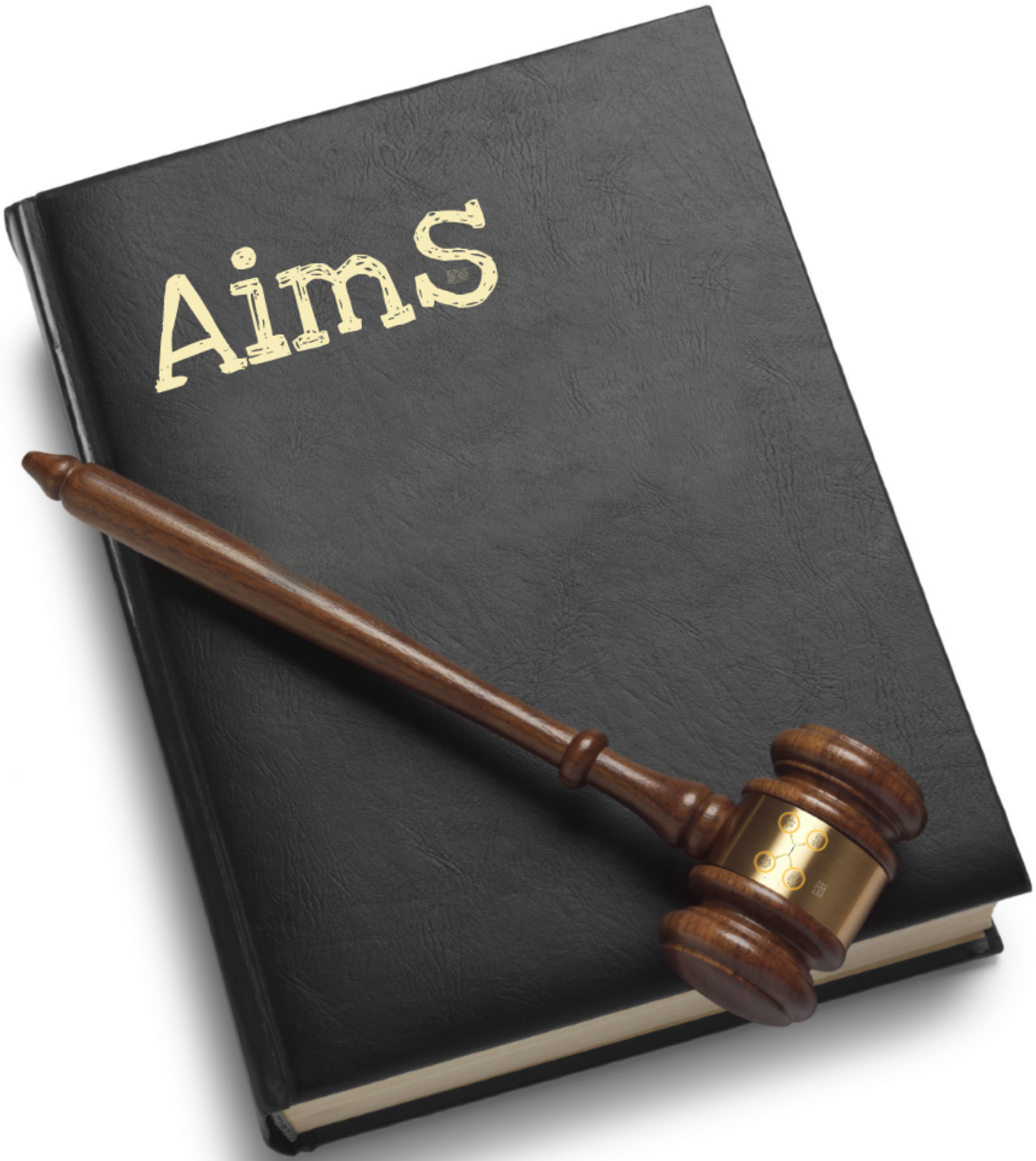

Identification
Analysis


Unveiling the Intricacies of Legal English Terminology from a Corpus-driven Perspective
Dr. María José Marín
Universidad de Murcia, Spain

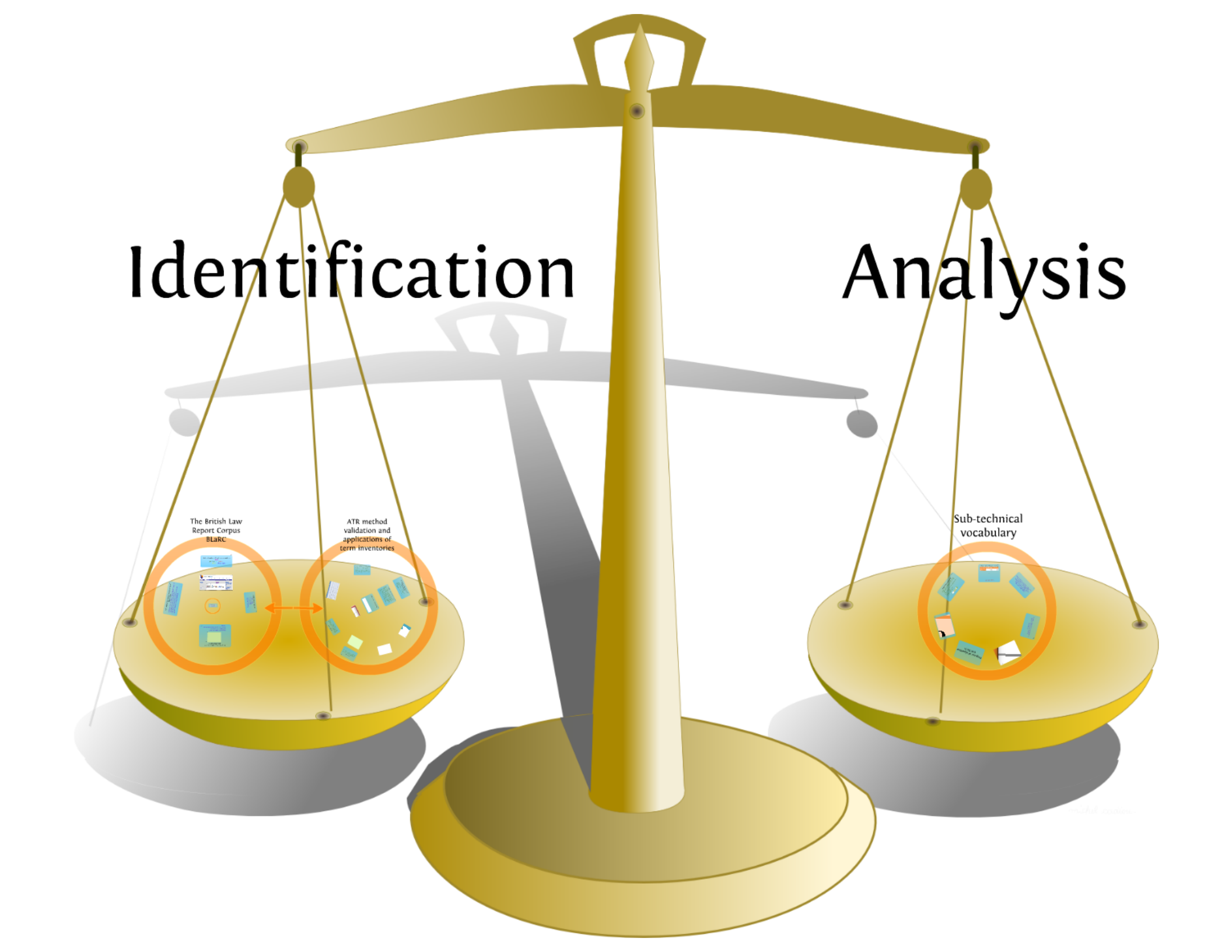


# Unveiling the Intricacies of Legal English Terminology from a Corpus-driven Perspective

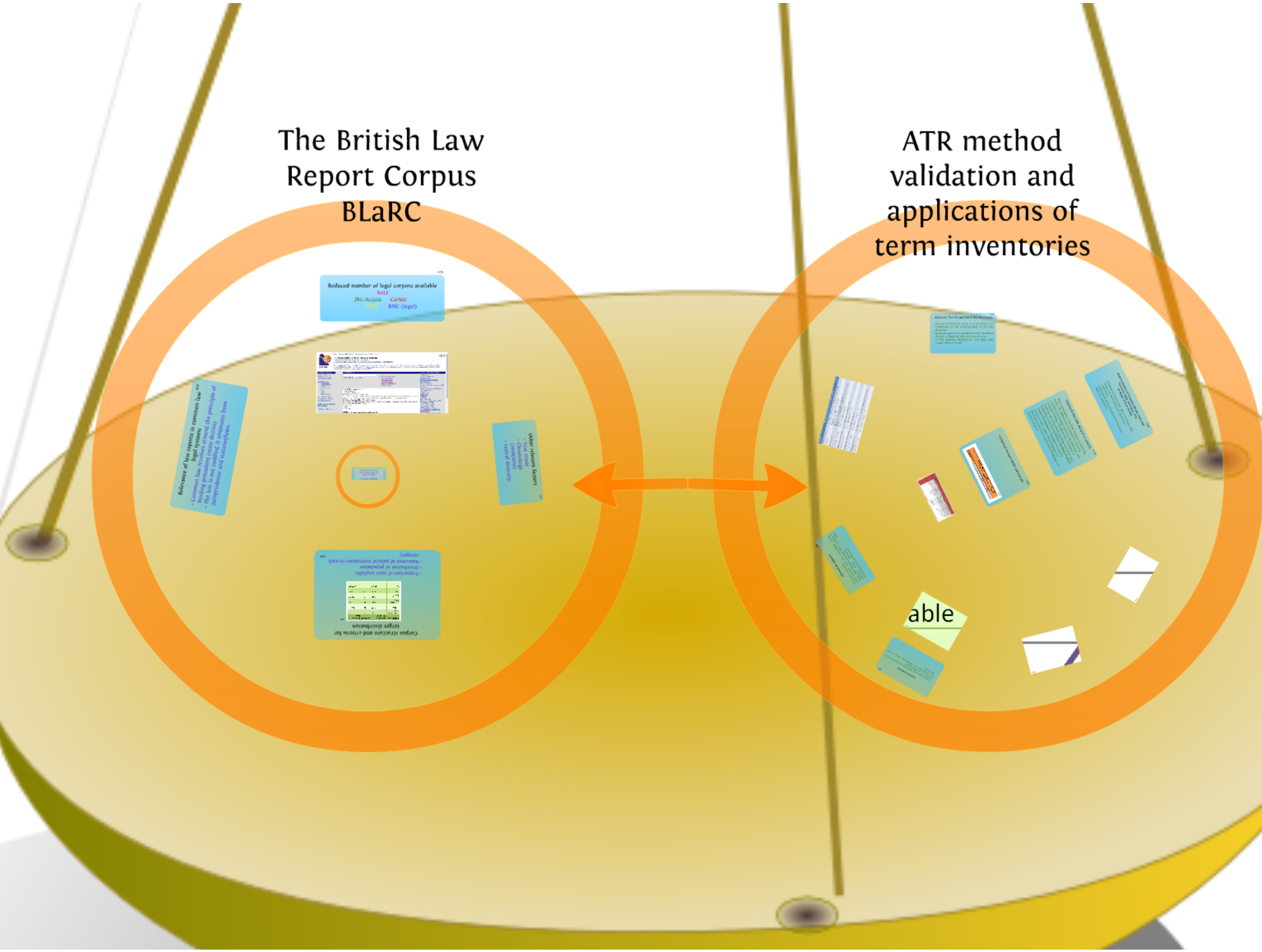
Dr. María José Marín
Universidad de Murcia, Spain


AimS

- Providing a reliable source of information for the study of legal English owing to the scarceness of legal corpora available.
- Validating varied ATR methods to determine their efficiency in legal term extraction.
- Producing inventories of single and multi-word legal terms for varied applications.
- Analysing such a complex phenomenon as sub-technicality in legal English from a quantitative perspective.


Identification
Analysis
The British Law Report Corpus BLaRC
ATR method validation and applications of term inventories
Sub-technical vocabulary


Identification


The British Law Report Corpus BLaRC
ATR method validation and applications of term inventories
Reduced number of legal corpora available
BoLC
JRC-Acquis
Cortec
BNC (legal)
able

# The British Law Report Corpus

## BLaRC

Reduced number of legal corpora available

BoLC

JRC-Acquis Cortec

HOLJ BNC (legal)




# British and Irish Legal Information Institute

Access to Freely Available British and Irish Public Legal Information

DONATE via Virgin Money Giving - DONATE via Charities Aid Foundation - Major Sponsors

Welcome to BAILII, where you can find British and Irish case law & legislation, European Union case law, Law Commission reports, and other law-related British and Irish material. BAILII thanks The Scottish Council of Law Reporting for their assistance in establishing the Historic Scottish Law Reports project. BAILII also thanks Sentral for provision of servers. For more information, see About BAILII.

## BAILII RESOURCES

Recent Decisions Lists
Recent Additions Lists
New Cases of Interest

BAILII Databases
- United Kingdom
- England and Wales
- Scotland
- Northern Ireland
- Jersey
- Ireland
- Europe
- Other Documents

A-Z case name index
A-Z legislation title index
A-Z other titles index

Leading Case law by Subject
Link to BAILII

BAILII Annual Lectures

## SEARCH BAILII

in All Databases Search

- Find by Case Citation
- Find by Case Title
- Case Law Search
- Legislation Search
- Other Materials Search
- Advanced Search

**The BAILII Lecture 2016**
SAVE THIS DATE
The BAILII Lecture 2016
On Wednesday 9 March 2016 at 5:30pm
The Trustees of the British and Irish Legal Information Institute (BAILII) are delighted to announce that the Third Annual BAILII Lecture will be given by Lord Thomas of Cwmgiedd, Lord Chief Justice of England and Wales:
**Developing Commercial Law Through the Courts**
By invitation only. Those wishing to attend will need to notify us by sending a confirmatory email. The lecture will take place in the auditorium of Freshfields:
65 Fleet Street
London
EC4Y 1HT

**BAILII is the most popular free legal website!**

## WORLD LAW RESOURCES

Asia (AsianLII)
Australasia (AustLII)
Canada (CanLII)
Common Law (CommonLII)
Cyprus (CyLaw)
Droit Francophone
Germany (Juristisches Internetprojekt Saarbrücken)
Global Legal Information Network
Hong Kong (HKLII)
Ireland (IrLII)
Italy (ITTIG)
JuriBurkina
JuriNiger
Kenya (KenyaLaw)
University of Montreal (LexUM)
New Zealand (NZLII)
Pacific Islands (PacLII)
Philippines (LawPhil)
Southern Africa (SAFLII)
UK Territories & Dependencies
USA (Cornell)
World Legal Information Institute

# Relevance of law reports in common law legal systems

- Common law revolves around the principle of binding precedent (stare decisis).
- The law is not codified, it emanates from jurisprudence and statutes/laws.

# Corpus structure and criteria for target distribution



| Jurisdictions | Available texts 2008-2010 | % of total | Final word target |
|---|---|---|---|
| Commonwealth Countries | 152 | 0.91% | 55,693 |
| UK Courts | 4,273 | 25.64% | 4,246,965 |
| England and Wales | 8,972 | 54.06% | 3,322,810 |
| Northern Ireland | 2,006 | 12.1% | 736,263 |
| Scotland | 1,209 | 7.29% | 495,466 |
| **Total** | **16,612** | | **8,857,197** |

- Proportion of texts available
- Distribution of population
- Relevance of judicial institutions in each category

| Jurisdictions | Available texts 2008-2010 | % of total | Final word target |
|---|---|---|---|
| Commonwealth Countries | 152 | 0.91% | 55,693 |
| UK Courts | 4,273 | 25.64% | 4,246,965 |
| England and Wales | 8,972 | 54.06% | 3,322,810 |
| Northern Ireland | 2,006 | 12.1% | 736,263 |
| Scotland | 1,209 | 7.29% | 495,466 |
| **Total** | **16,612** | | **8,857,197** |

| | | | |
|---|---|---|---|
| England and Wales | 8,972 | 54.06% | 3,322,810 |
| Northern Ireland | 2,006 | 12.1% | 736,263 |
| Scotland | 1,209 | 7.29% | 495,466 |
| **Total** | **16,612** | | **8,857,197** |

- Proportion of texts available
- Distribution of population
- Relevance of judicial institutions in each category

# Other relevant factors

- Text mode
- Chronology (2008/2010)
- Lexical diversity

CORPUS AVAILABLE AT
FLAX: http://flax.nzdl.org/greenstone3/flax

Tom Cobb's LEXTUTOR

CORPUS AVAILABLE AT FLAX: http://flax.nzdl.org/greenstone3/flax

Tom Cobb's LEXTUTOR

# ATR method validation and applications of term inventories

# Automatic Term Recognition in specialised texts

- Getting to know the terms in a specialised text contributes to the understanding of the text aboutness.
- Automatic access to specialised terms facilitates the task of handling large amounts of data.
- 10 ATR methods validated on 2.6m word pilot corpus subset of BLaRC.

| METHOD | IMPLEMENTATION ON *UKSCC*, THE PILOT CORPUS |
|---|---|
| *TermoStat* (Drouin, 2003) | Online: http://olst.ling.umontreal.ca/~drouinp/termostat_web/index.php |
| Chung (2003) | Manual: Ratio formula applied on excel spreadsheet.<br>*LACELL*: Reference corpus |
| Kit and Liu (2008) | Manual: Rank difference formula applied on spreadsheet.<br>*UKSCC* lemmatised and POS tagged (Tree tagger). BNC lemmatised lists used as reference. |
| *TF-IDF* (Sparck Jones, 1972) | Manual: Formula applied on spreadsheet |
| *RIDF* (Church and Gale, 1995) | Automatic: JATE java tool set (Zhang et al. 2008) |
| *Keywords* (Scott, 2008) | Automatic: *Wordsmith 5.0* (Scott, 2008) |
| *Terminus* (Nazar and Cabré, 2012) | Online: http://terminus.upf.edu/ |
| *C-value* (Frantzi et al., 1999) | Automatic: JATE tools (Zhang et al., 2008) |
| *Termextractor* (Sclano and Velardi, 2007) | Online: http://lcl.uniroma1.it/sso/index.jsp?returnURL=%2Ftermextractor%2F |
| *Textract* (Park et al., 2002) | Automatic: JATE tools (Zhang et al., 2008) |

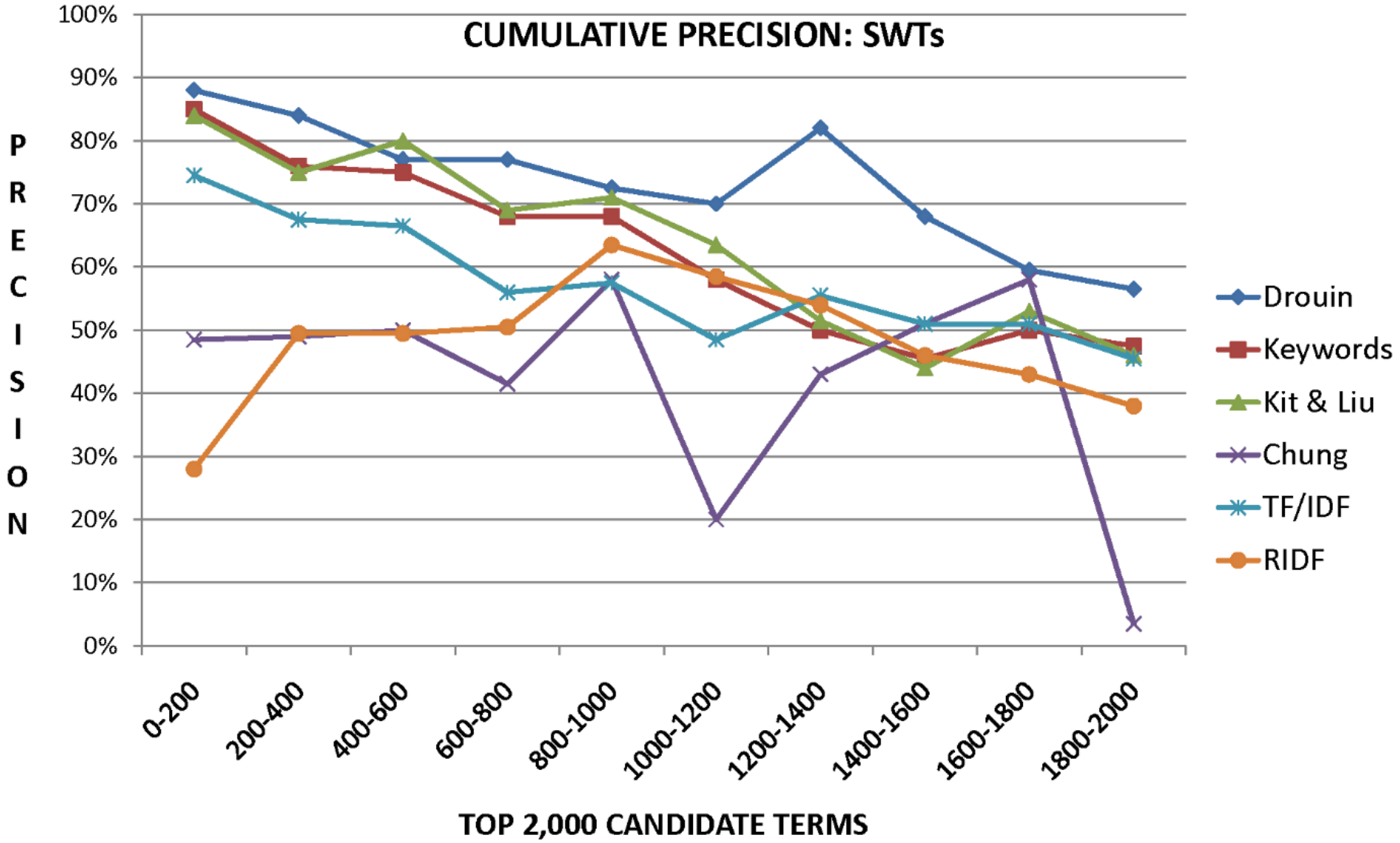
# ATR method validation

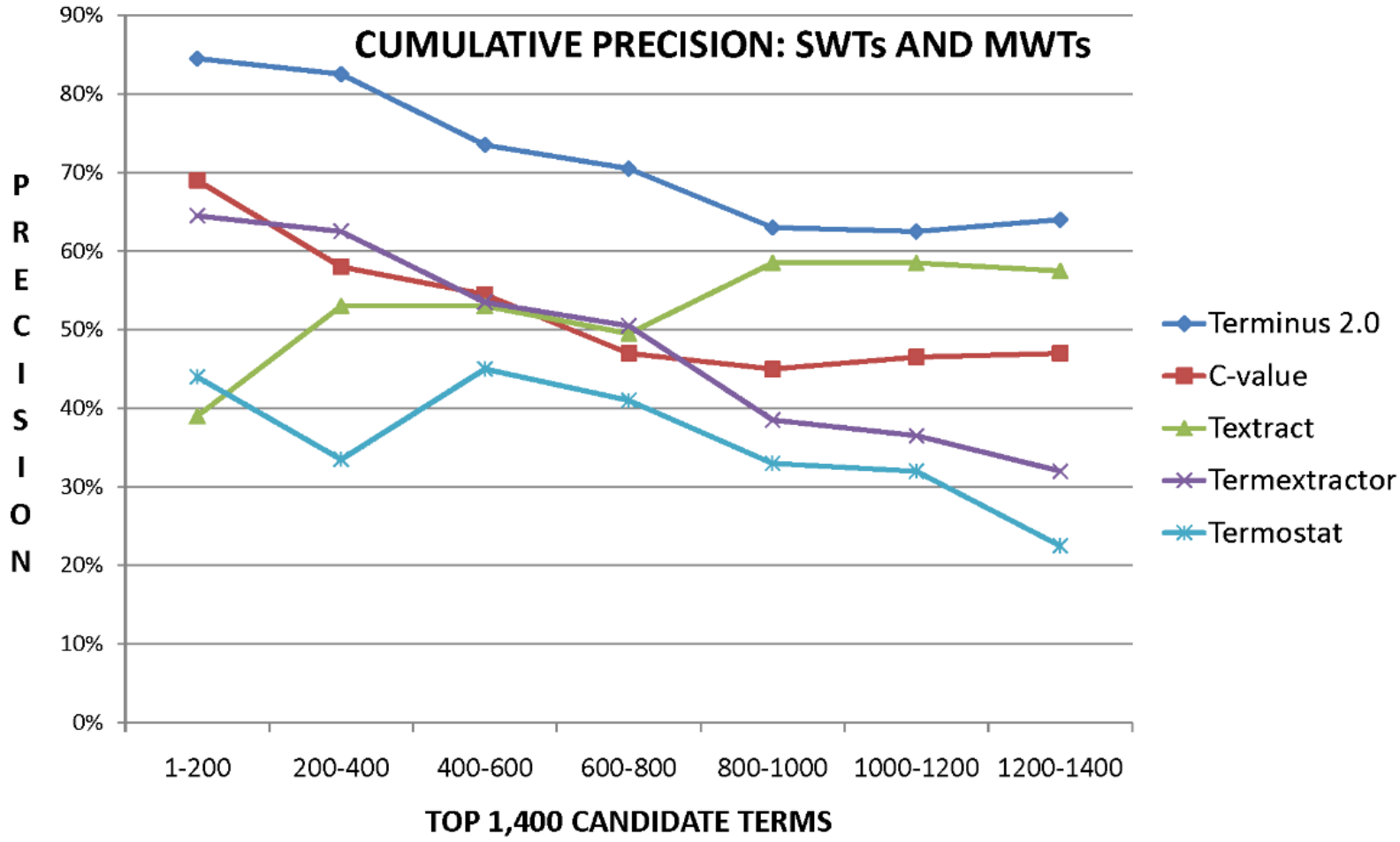
- Automatic validation
- Gold standard used for comparison: 10,088 entry legal glossary obtained form merging 4 different online glossaries + 2 paper dictionaries used as reference.

| | | |
|---|---|---|
| a coelo usque ad centrum | BAC | calendar call |
| a fortiori | Bachelor of laws | calendar year accounting period |
| a mensa et thoro | backdate | called to the Bar |
| a posteriori | back-to-back life sentences | calling-up notice |
| a prendre | BAD | calumny |
| a priori | bad debt | cancellation of removal |
| a priori assumption | bad faith | candidature |
| a verbis legis non est recedendum | badgering the witness | caning |
| a vinculo matrimonii | bail | canon law |
| a.k.a. | bail bond | cap and trade |
| ab initio | bail bondsman | capable |
| AB trust | bailable | capacity |
| ABA | bailee | capacity to contract |
| abandon | bailiff | capital |
| abandoned property | BAILII | capital account |
| abandonment | bailment | capital asset |
| abandonment of residence | bailor | capital case |

# Evaluation procedure

- Precision levels were determined by comparison of the CT lists with the gold standard.
- The comparison was carried out using an excel spreadsheet.


CUMULATIVE PRECISION: SWTs
PRECISION
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
0-200
200-400
400-600
600-800
800-1000
1000-1200
1200-1400
1400-1600
1600-1800
1800-2000
TOP 2,000 CANDIDATE TERMS
Drouin
Keywords
Kit & Liu
Chung
TF/IDF
RIDF


CUMULATIVE PRECISION: SWTs AND MWTs
PRECISION
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
1-200
200-400
400-600
600-800
800-1000
1000-1200
1200-1400
TOP 1,400 CANDIDATE TERMS
Terminus 2.0
C-value
Textract
Termextractor
Termostat

# Implementation of Termostat (Drouin, 2003) and Terminus (Nazar and Cabré, 2012) on BLaRC (8.85 m words)

- Production of term lists. After validation, 2,848 single and multi-word true terms were confirmed.
- Applications of specialised term inventories: didactic exploitation.

# McEnery & Xiao (2011: 364-5) on CL and ESL

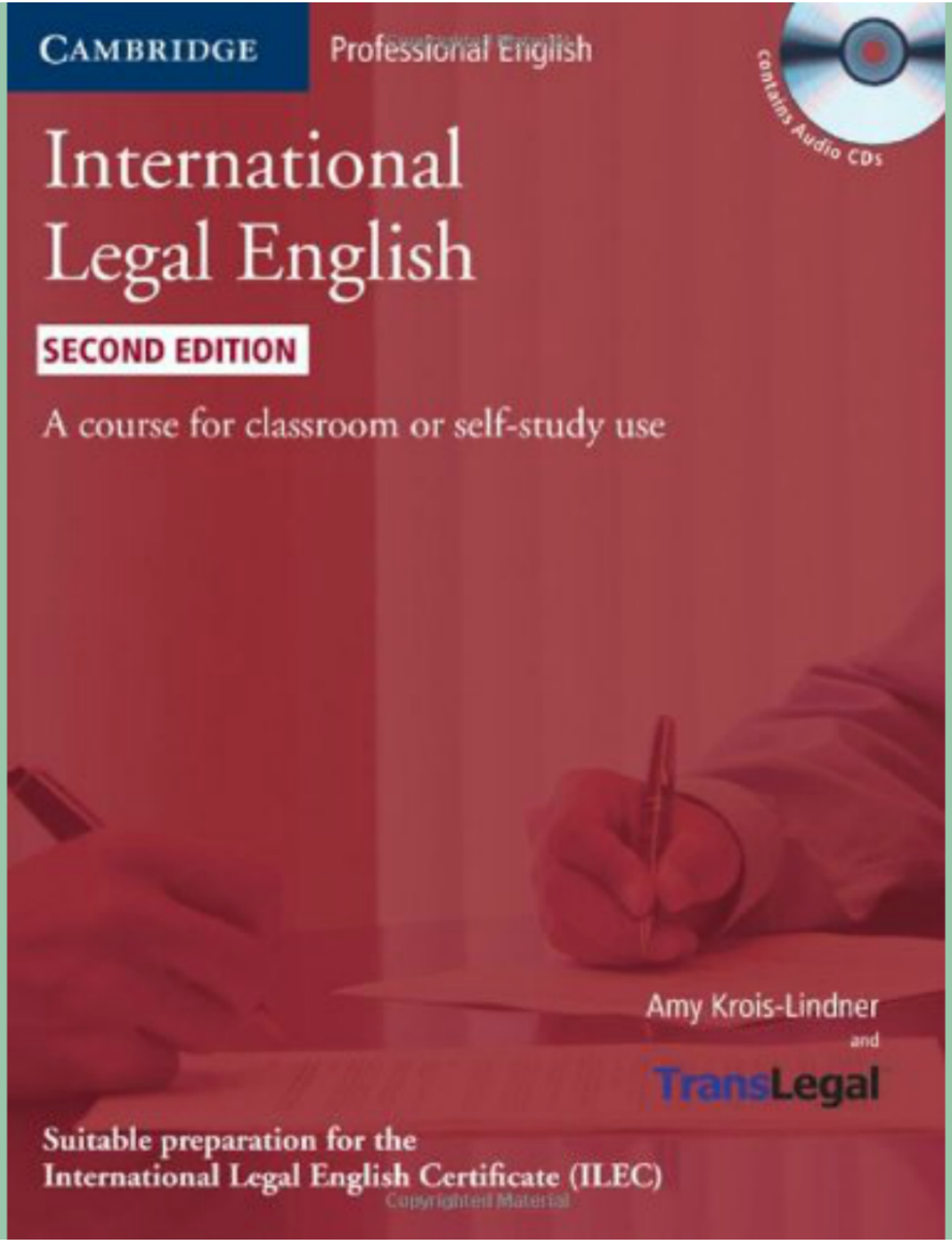
"That convergence has three focuses, as noted by Leech: the **indirect use** of corpora in teaching (reference publishing, materials development, and language testing), the **direct use** of corpora in teaching (teaching about, teaching to exploit, and exploiting to teach) and **further teaching-oriented corpus development**: languages for specific purposes (LSP) corpora, first language (L1) developmental corpora and second language (L2) learner corpora".

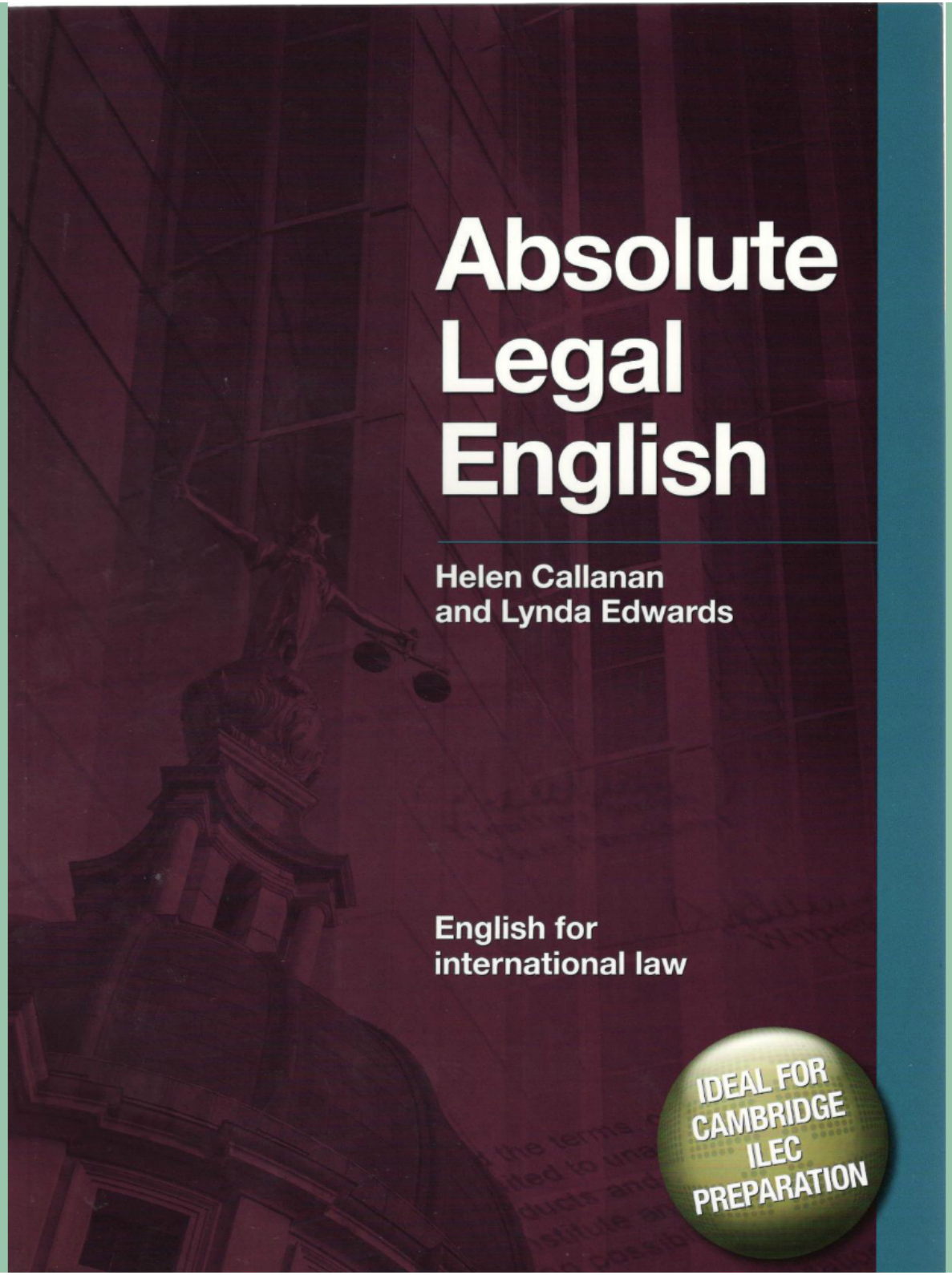
# Comparison between BLaRC and LeGTeXT

CAMBRIDGE

# Professional English in Use

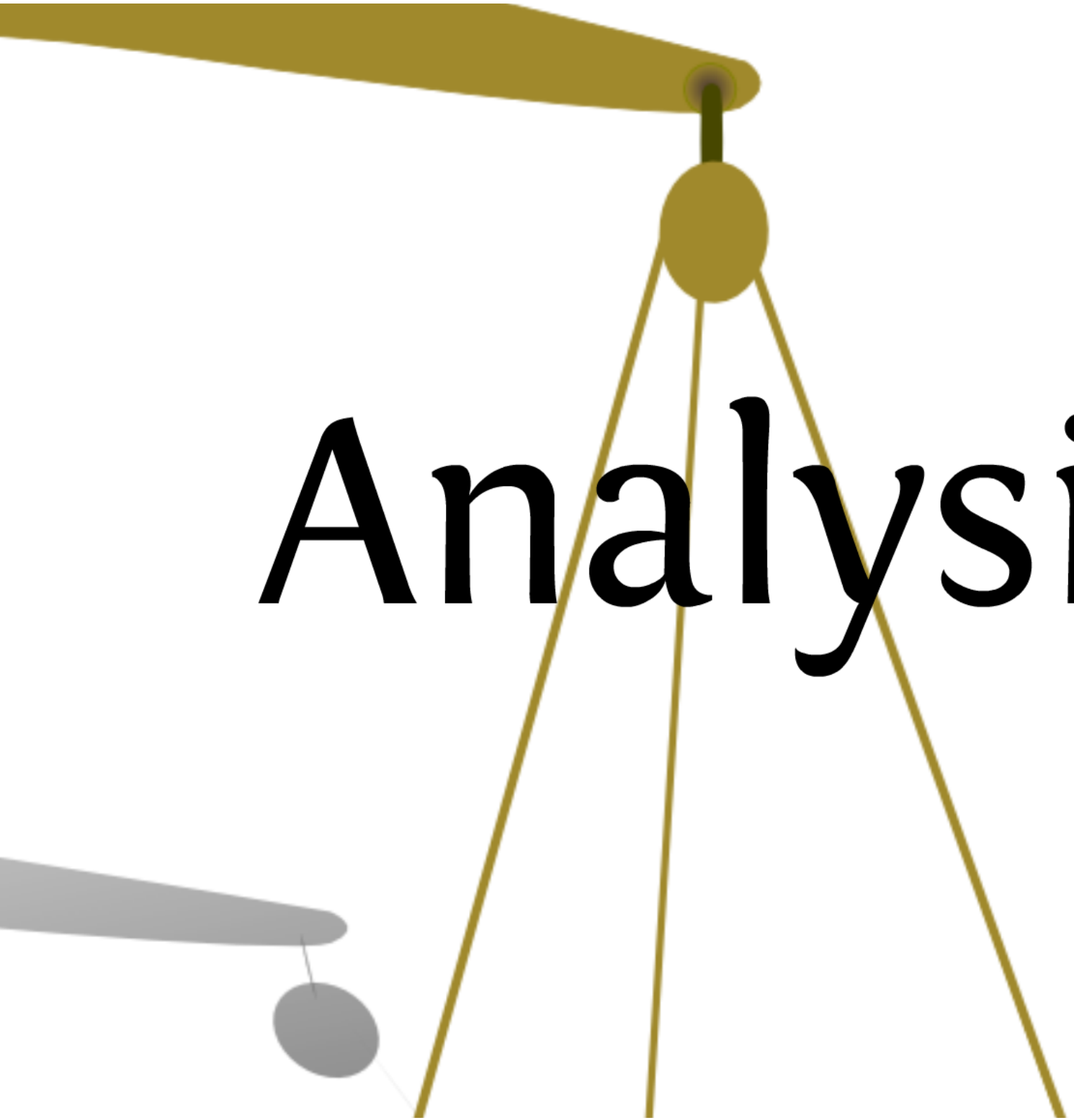

Gillian D Brown
Sally Rice


CAMBRIDGE
Professional English
contains Audio CDs
International Legal English
SECOND EDITION
A course for classroom or self-study use
Amy Krois-Lindner
and
TransLegal
Suitable preparation for the International Legal English Certificate (ILEC)
Copyrighted Material


Absolute
Legal
English
Helen Callanan
and Lynda Edwards
English for
international law
IDEAL FOR
CAMBRIDGE
ILEC
PREPARATION

| Corpus | Term list obtained with TermoStat | Overlap with BLaRC term list | Text coverage provided by BLaRC |
|---|---|---|---|
| **LeGTeXT** | 1570 | 1072/ 67% | 12.37% |

| | LANGUAGE LEVEL | OBJECTIVES | TASK | CORPUS USE |
|---|---|---|---|---|
| **ACTIVITY 1** | **Morphological** | Developing students' **awareness** on **word formation** processes | Forming **word families** of terms like *claim, appeal* or *trial* through **derivational processes** (+ affixes) | Showing **concordances** illustrating the real **contexts** of usage of those terms which can serve to **test** the results of their guesses |
| **ACTIVITY 2** | **Syntactic** | Learning **grammar patterns** (prepositions) | Studying the **collocates** of some legal terms (*claim*, *appeal*, *breach*) to select the **most frequent prepositions** they occur with | The **corpus** should be **processed** to obtain the **main collocates** of the terms selected (pre-training required) |



# Analysis

# Sub-technical vocabulary

# Relevance of sub-technical vocabulary

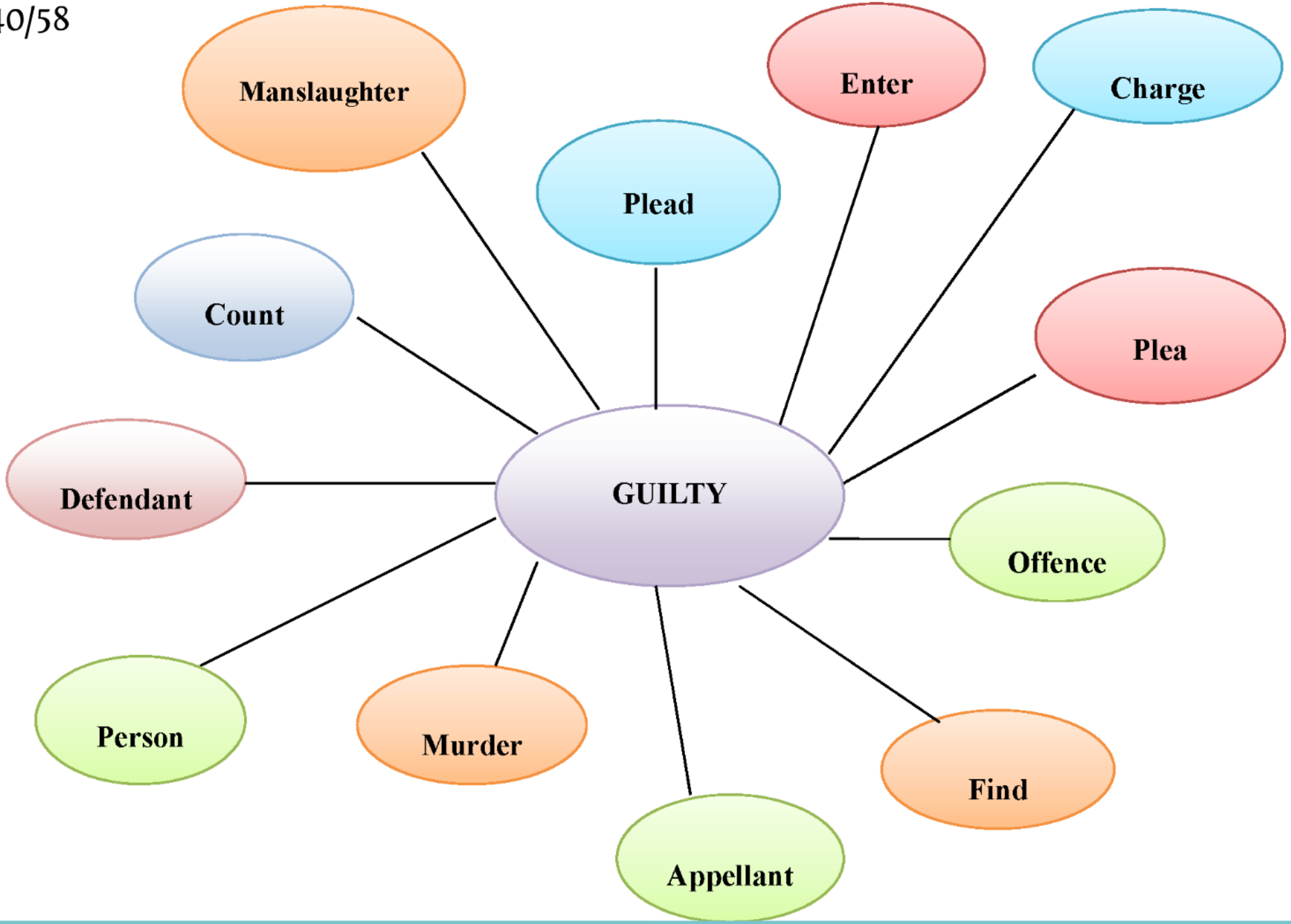
| General Vocabulary Lists | Overlapping Terms *BLaRC/GSL; AWL; BNC* | % Overlap |
|---|---|---|
| West's (1953) *GSL* & | 1,040/2,570 | 40.47% |
| Coxhead's (2000) *AWL* | | |
| *BNC* (2007) | 1,362/3,000 | 45.41% |

| **Category 1**: | **Category 2**: | **Category 3**: |
|---|---|---|
| Words denoting a legal concept which are frequently used both in the general and specialised fields not changing their meaning in the legal context. | Words often employed both in the general and specialised fields which change their meaning in the legal context sharing some semantic features with their original meaning | Words occurring more frequently in the specialised field than in the general one which change their meaning in the legal environment acquiring a new meaning. Their new meaning is quite distant or completely unrelated to their general sense. |
| *judge, court, tribunal, law, prosecution, jury, legislation, robbery, theft, guilty,solicitor* | *charge, offence, sentence, claim, decision, grounds, complaint, dismiss, evidence, relief, record, trial, battery* | *appeal, conviction, party, warrant, terms, act* |

In spite of its relevance, no method has been designed to try and measure this phenomenon

# Williams' lexical networks (2001)

- Extension of word associations by considering several collocational levels (limited to 2 in this case).
- More contextual information associated to the network nucleus.
- Collocational span: 5R, 5L
- Frequency threshold: >30


Manslaughter
Enter
Charge
Plead
Count
Plea
Defendant
GUILTY
Offence
Person
Murder
Find
Appellant


Make
Section
Rule
Paragraph
Order
Article
Decision
Notice
OFFENCE
Regulation
Act
Appeal
Court
Claim
Agreement
Clause

| Word | Specialised collocates/ co-collocates (normalised) | General usage collocates/co-collocates (normalised) | Normalised Frequency BLaRC | Normalised Frequency LACELL |
|---|---|---|---|---|
| PURSUANT | 404.40 | 0 | 10.34 | 0 |
| ESTOPPEL | 114.57 | 0 | 8.65 | 0 |
| LIABILITY | 421.69 | 0 | 8.20 | 0 |
| BATTERY | 27.57 | 0.73 | 7.89 | 2.27 |
| CONVICTION | 281.35 | 1.33 | 10.41 | 3.23 |
| SENTENCE | 491.25 | 1.53 | 9.50 | 2.98 |
| DISMISS | 338.64 | 3.20 | 10.06 | 3.81 |
| SOLICITOR | 159.77 | 0.33 | 8.23 | 2.39 |
| RELIEF | 184.18 | 6.08 | 9.88 | 4.45 |
| TRIAL | 666.66 | 2.33 | 9.22 | 3.84 |
| LEGISLATION | 246.44 | 39.7 | 9.23 | 4.2 |
| WARRANT | 30.39 | 1.60 | 7.91 | 3.01 |
| PARTY | 708.36 | 274.13 | 9.22 | 4.73 |
| CHARGE | 167.68 | 64.77 | 9.08 | 4.89 |
| COMPLAINT | 180.22 | 18.18 | 8.79 | 4.70 |
| OFFENCE | 522.93 | 28 | 8.91 | 5.03 |
| GUILTY | 66.55 | 11.96 | 6.87 | 4.25 |
| EAT | 0 | 2.20 | 0 | 3.27 |
| BLUE | 0 | 13.43 | 0 | 3.52 |
| MORNING | 0 | 268.36 | 0 | 4.94 |

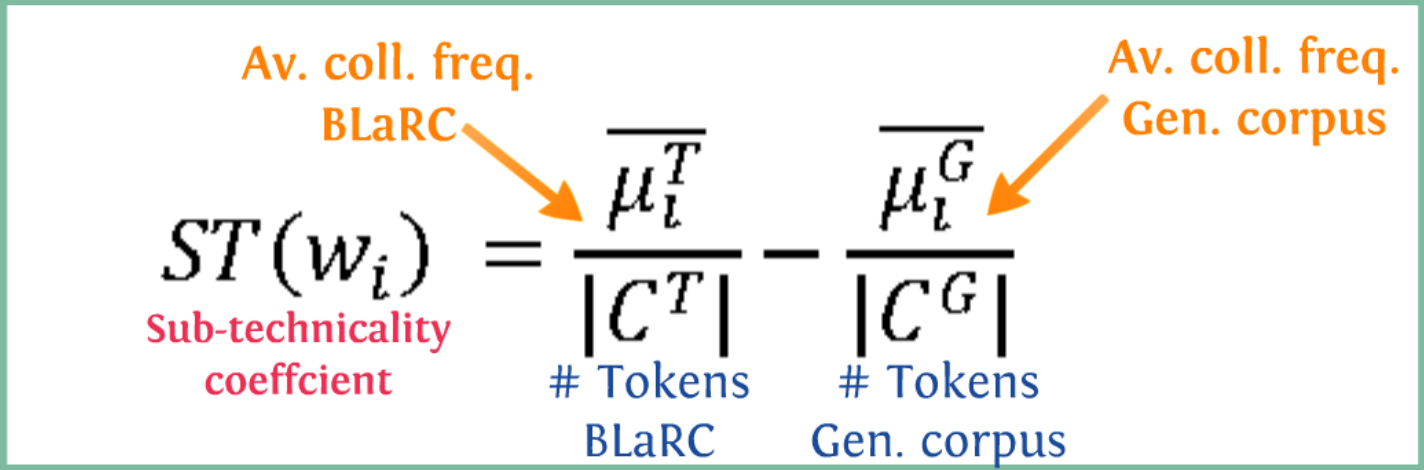
# Proposal of algorithm SUB-TECH

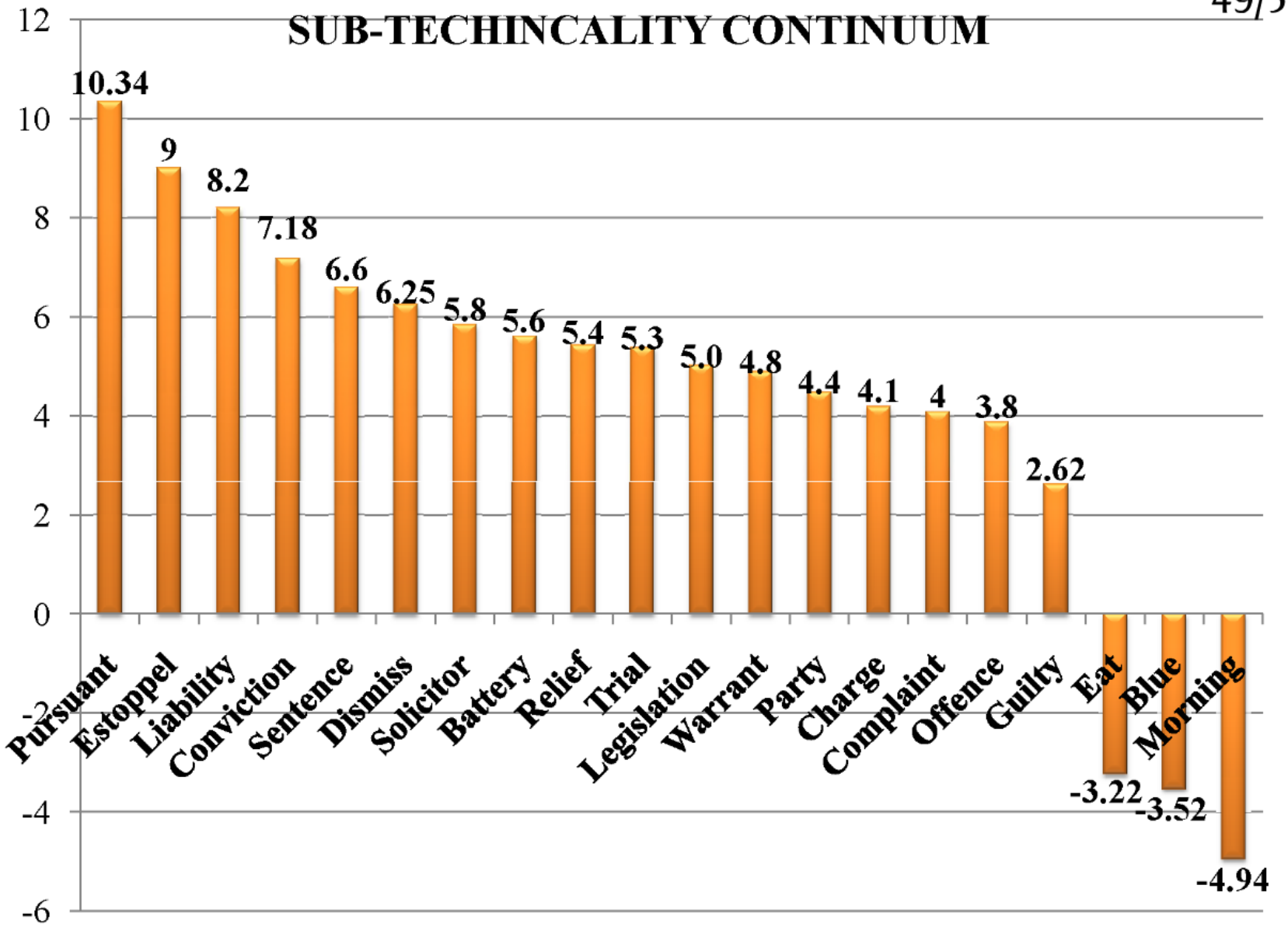
Step 1:
Identification and extraction of the specialised vocabulary from BLaRC applying both Drouin's (2003) TermoStat and Nazar and Cabre's (2012) Terminus 2.0.

Step 2:
Manual extraction of terms shared both by the general and specialised fields.

Step 3:
Application of Williams' (2001) lexical network model to the list of words selected both in the specialised and general corpora with the aim of comparing results.

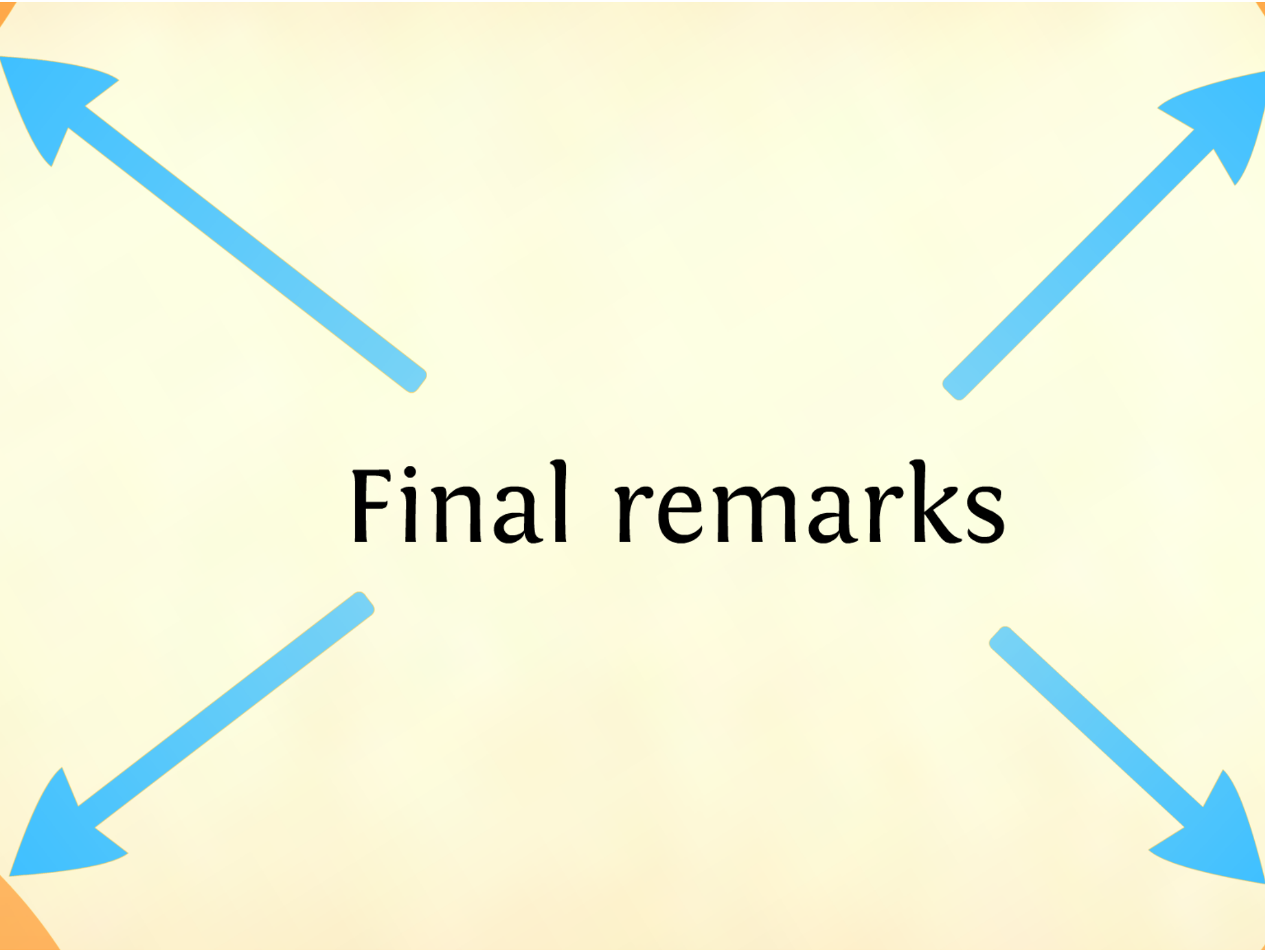
Step 4: Implementation of the formula presented below for the ranking of sub-technical terms along a continuum of specialisation.


Av. coll. freq.
BLaRC
Av. coll. freq.
Gen. corpus
$ST(w_i) = \frac{\overline{\mu_i^T}}{|C^T|} - \frac{\overline{\mu_i^G}}{|C^G|}$
Sub-technicality coeffcient
# Tokens BLaRC
# Tokens Gen. corpus

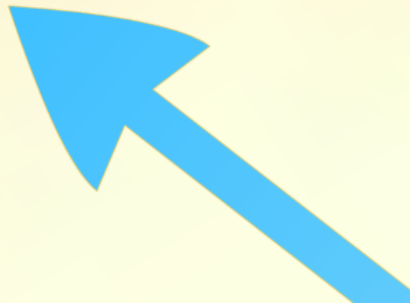

SUB-TECHINCALITY CONTINUUM
12
10
8
6
4
2
0
-2
-4
-6
10.34
9
8.2
7.18
6.6
6.25
5.8
5.6
5.4
5.3
5.0
4.8
4.4
4.1
4
3.8
2.62
-3.22
-3.52
-4.94
Pursuant
Estoppel
Liability
Conviction
Sentence
Dismiss
Solicitor
Battery
Relief
Trial
Legislation
Warrant
Party
Charge
Complaint
Offence
Guilty
Eat
Blue
Morning

Nonetheless, quantitative methods fail to provide a full picture of linguistic phenomena

Further research needs to be conducted to complement the data obtained by applying Sub-Tech for a fuller and also qualitative characterisation of sub-technical legal terms


Final remarks

Providing a reliable source of information for the study of legal English: BLaRC



Selecting the two most efficient ATR methods in legal term extraction

Obtaining a list of specialised single and multi-word legal terms that could be employed as support for the teaching of legal vocabulary

# Proposing a method of analysis of sub-technical vocabulary applying quantitative criteria

## MAJOR REFERENCES


- Cabré Castellví, M. T. 2013. "Panorama des approches et tendances de la terminologie aujourd'hui". In Quirion, J., Depecker, L., Rousseau, L.J. (eds.) *Dans tous les sens du terme*. Ottawa: Presses de l'Université d'Ottawa: 133-152.
- Drouin, P. 2003. "Term extraction using non-technical corpora as a point of leverage." *Terminology*, 9 (1): 99-117.
- McEnery, T, Wilson, A. 2001. *Corpus Linguistics*. Edinburgh: Edinburgh University Press.
- McEnery, T., Xiao, R. 2011. "What corpora can offer in language teaching and learning". In Hinkel, E. (ed.), *Handbook of Research in Second Language Teaching and Learning*. London: Routledge
- Marín, M.J. 2014. "Evaluation of five single-word term recognition methods on a legal corpus". *Corpora*, 9 (1). Endinburgh: Endinburgh University Press.
- Marín, M.J. (in the press, 2016). "Measuring the Degree of Specialisation of Sub-Technical Legal Terms through Corpus Comparison: a Domain-Independent Method".*Terminology*, 22 (1). John Benjamins.
- Solan, L., Tiersma, P. 2012. *Oxford Handbook of Language and Law*. OUP.
- Tiersma, Peter. 1999. *Legal Language*. Chicago: The University of Chicago Press.