Australian Government

Australian Law Reform Commission

Explanatory Note to Complexity and linguistic data

This note covers the following two data sets:

- **As made Acts – Complexity and linguistic data**: This data set covers all Acts made by the Australian Parliament since 1901. The data set was generated by textually analysing each as made Act published on the Federal Register of Legislation, and by combining this data with data from several other ALRC data sets. Full methodology notes are available here.
- **In force Acts – Complexity and linguistic data**: This data set covers all in force principal Acts of the Australian Parliament. The data set was generated by textually analysing each in force Act published on the Federal Register of Legislation, and by combining this data with data from several other ALRC data sets. Full methodology notes are available here.

The first 25 columns of these data sets are derived from the core legislative data sets published on the DataHub. To understand these columns, please consult the Explanatory Note for core legislative data. This Explanatory Note only covers the 60 columns that are unique to the two data sets listed above.

These 60 variables have been created by the ALRC largely through using the R programming language to computationally analyse the text of each Act. As explained in the Table on the next page, several variables were created by matching data among the following data sets published on the DataHub:

- **As made Commonwealth Acts**: This data set covers a subset of as made Commonwealth legislation, covering all as made Commonwealth Acts of Parliament.
- **In force Commonwealth legislative instruments**: This data set covers a subset of in force Commonwealth legislation, covering all in force Commonwealth Acts of Parliament.
- **Amending legislative relationships**: This data set contains all amending relationships in which one piece of legislation amends another piece of legislation. The data was obtained by scraping the 'Amends' webpage for each as made Act or legislative instrument.
- **Modifications legislative relationships**: This data set contains all relationships in which one piece of legislation modifies another piece of legislation. The data was obtained by scraping the 'Modifies' webpage for each as made Act or legislative instrument.

Column name	Description
yearsInforce	The year extracted from the ‘numberYear’ variable subtracted from 2022.
amendmentsPerYear	‘allAmendingNum’ variable divided by ‘yearsInforce’ variable
chapter	The number of Chapters marked-up in the HTML. Some Chapters are marked-up without content, and these are removed, as are ‘placeholder’ Chapters that appear in the HTML. Duplicate Chapters are removed based on their full name. For example, ‘Chapter 7—Financial services and markets’ appears in both Volumes 4 and 5 of the Corporations Act 2001. This is only counted once. Older legislation, international treaties, and provisions that amend another Act may not use markup. Depending on the Act, this may therefore be an undercount
schedules	Number of Schedules marked-up in the HTML. Some Schedules are marked-up without content, and these are removed, as are ‘placeholder’ Schedules that appear in the HTML. Duplicate Schedules are removed based on their full name. Older legislation, international treaties, and provisions that amend another Act may not use markup. Depending on the Act, this may therefore be an undercount.
part	Number of Parts marked-up in the HTML. Some Parts are marked-up without content, and these are removed, as are ‘placeholder’ Parts that appear in the HTML. Duplicate Parts are removed based on their full name. Duplicates are therefore only counted once. Older legislation, international treaties, and provisions that amend another Act may not use markup. Depending on the Act, this may therefore be an undercount
div	Number of Divisions marked-up in the HTML. Some Divisions are marked-up without content, and these are removed, as are ‘placeholder’ Divisions that appear in the HTML. Duplicate Divisions are removed based on their full name. Duplicates are therefore only counted once. Older legislation, international treaties, and provisions that amend another Act may not use markup. Depending on the Act, this may therefore be an undercount.
subdiv	Number of Subdivisions marked-up in the HTML. Some Subdivisions are marked-up without content, and these are removed, as are ‘placeholder’ Subdivisions that appear in the HTML. Duplicate Subdivisions are removed based on their full name. Duplicates are therefore only counted once. Older legislation, international treaties, and provisions that amend another Act may not use markup. Depending on the Act, this may therefore be an undercount.
section	Number of Sections marked-up in the HTML (eg 423A, 732). Older legislation, international treaties, and provisions that amend another Act may not use markup. Depending on the Act, this may therefore be an undercount.
subsection	Number of Subsections marked-up in the HTML (eg (1), (12))
paragraph	Number of Paragraphs marked-up in the HTML (eg (a), (aa)).

Column name	Description
paragraphsub	Number of Subparagraphs marked-up in the HTML (eg (i), (iv)).
paragraphsub_sub	Number of Sub-subparagraphs marked-up in the HTML (eg (A), (B))
notes	Number of Notes marked-up in the HTML. Older legislation, international treaties, and provisions that amend another Act may not use markup. Depending on the Act, this may therefore be an undercount.
above_sec	Sum of chapter + schedules + part + div + subdiv
sec_below	Sum of section + subsection + paragraph + paragraphsub + paragraphsub
amend_sch	Number of Amending Schedules marked-up in the HTML (ie Schedules of a piece of legislation that contain amendments to another piece of legislation). This variable can help identify legislation that contains amendments to other pieces of legislation
amend_part	Number of Amending Parts marked-up in the HTML (ie Parts of a piece of legislation that contain amendments to another piece of legislation). This variable can help identify legislation that contains amendments to other pieces of legislation.
items	Number of Items marked-up in the HTML (ie 'ItemHead'). Items generally contain amendments to another piece of legislation. This variable can help identify legislation that contains amendments to other pieces of legislation.
tokenCount	Counts the number of 'tokens' using R's Quanteda package, splitting the document using 'fasterword'. This analysis also does not include front matter, endnotes, and tables of contents. Does not split hyphens. This methodology means it is the same as or very close to a word count produced by Microsoft Word.
substantiveTokenCount	This is based on counting tokens after the tokens have been 'cleaned' to remove non-substantive words. Cleaning involved removing: • tokens with numbers (e.g. 798ABA); • tokens in brackets that are between 1 and 3 characters long (e.g. provision numbers such as (12) or (a)); • all punctuation; and • all stopwords. ***Stopwords*** are words that are removed for certain types of text analysis, such as determining the number of unique words in a text (i.e. its vocabulary). The ALRC uses R's Quanteda package, which includes 172 stopwords. Stopwords include 'the', 'its', and 'this', for example. This analysis also does not include front matter, endnotes, and tables of contents.

Column name	Description
Definitions_number_defs	Number of defined terms marked-up in the HTML. Because defined terms are accompanied by a definition, this is also a count of definitions.
Definitions_unique_defs	Number of defined terms with duplicated defined terms removed (eg if 'property' is defined three times in an Act, it is only counted once here).
Definitions_used	The number of times a potentially defined term is used in the legislation. Determined using a list of all terms marked up in the HTML as defined terms in the piece of legislation. The use of a term is not counted where it appears in the use of another defined term (to avoid duplication). For example, 'financial product advice', a defined term, is counted and the use of 'financial product', another defined term, is not counted when it appears in that defined term. Terms are 'potentially' defined because not all definitions apply for all provisions in a piece of legislation, so a term may be used in an undefined sense even if defined for other provisions.
Definitions_word_count	The number of words that are potentially defined in an Act, determined using the same approach for the 'Definitions_used' variable but counting words comprising the terms rather than uses of the terms. For example, while 'financial product advice' will only count as one use of a defined term it will count for three words that are potentially defined. Words are 'potentially' defined because not all definitions apply for all provisions in a piece of legislation, so a word may be used in an undefined sense even if defined for other provisions.
Definitions_relative_total_words	The percentage of words in the Act that are potentially defined. Produced by dividing the 'Definitions_word_count' variable by the 'tokenCount' variable.
Tagged_concepts_number_defs	Number of bold and italicised terms marked-up in the HTML (*category B investment*).
Tagged_concepts_unique_defs	Number of bold and italicised terms with duplicated defined terms removed (eg if 'property' is defined three times, it is only counted once here).
Tagged_concepts_used	The number of times a term is used in the legislation that is potentially affected by a bold and italicised term. Determined using a list of all terms marked up in the HTML as bold and italicised terms in the piece of legislation. The use of a term is not counted where it appears in the use of another bold and italicised term (to avoid duplication). For example, 'financial product advice', a bold and italicised term, is counted and the use of 'financial product', another bold and italicised term, is not counted when it appears in that bold and italicised term. Terms are 'potentially' used because not all bold and italicised terms apply for all provisions in a piece of legislation, so a term may be used in an undefined or untagged sense even if defined or tagged for other provisions.
Tagged_concepts_word_count	The number of words that are potentially affected by a bold and italicised term in the legislation, determined using the same approach for the 'Tagged_concepts_used' variable but

Column name	Description
	counting words comprising the terms rather than uses of the terms. For example, while 'financial product advice' will only count as one use of a bold and italicised term it will count for three words that are potentially affected by a bold and italicised term. Words are 'potentially' bold and italicised because not all bold and italicised terms apply for all provisions in a piece of legislation, so a word may be used in an undefined or untagged sense even if defined or tagged for other provisions.
Tagged_concepts_relative_total_words	The percentage of words in the Act that are potentially affected by a bold and italicised term. Produced by dividing the 'Tagged_concepts_used' variable by the 'tokenCount' variable.
inforceModificationsNum	The number of in force modifications (also known as notional amendments) to the Act. Based on the 'ModifiedBy' page. This page is accessed through the 'Series' page of the legislation. The ALRC collects this data in the in the **Modifications legislative relationships** data set published on the DataHub. Data matching with the **In force Commonwealth legislative instruments** data set ensures that only in force modifications are counted.
inforceModificationsPages	The number of pages of in force modifications (also known as notional amendments) to the Act. Based on the 'ModifiedBy' page. This page is accessed through the 'Series' page of the legislation. The ALRC collects this data in the in the **Modifications legislative relationships** data set published on the DataHub. Data matching with the **In force Commonwealth legislative instruments** data set ensures that only pages from in force modifications are counted.
allModificationsNum	The number of as made modifications (also known as notional amendments) to the Act, including those not in force. Based on the 'ModifiedBy' page. This page is accessed through the 'Series' page of the legislation. The ALRC collects this data in the in the **Modifications legislative relationships** data set published on the DataHub.
allModificationsPages	The number of pages of as made modifications (also known as notional amendments) to the Act, including those not in force. Based on the 'ModifiedBy' page. This page is accessed through the 'Series' page of the legislation. The ALRC collects this data in the in the **Modifications legislative relationships** data set published on the DataHub. Data matching with the **As made Commonwealth legislative instruments** data set allows pages to be counted.
allAmendingNum	The number of amendments to the legislation based on the 'Act Amendments' tab that appears on the 'Principal + Amendments' page for each Act on the Federal Register of Legislation. This page is accessed through the 'Series' page of the legislation. The ALRC collects this data in the **Amending legislative relationships** data set published on the DataHub.

Column name	Description
allAmendingPages	The number of pages of amendments to the legislation based on the 'Act Amendments' tab that appears on the 'Principal + Amendments' page for each Act on the Federal Register of Legislation. This page is accessed through the 'Series' page of the legislation. The ALRC collects this data in the **Amending legislative relationships** data set published on the DataHub. Data matching with the **As made Commonwealth Acts** data set allows for the counting of pages.
entropyScore	Calculated using the below equation from Patrick McLaughlin et al, 'Is Dodd-Frank the Biggest Law Ever?' (2021) 7(1) Journal of Financial Regulation 149, 170. Does not include endnotes, table of contents, and excluded structural elements. Does not split hyphens. Also excludes stopwords, numbers, and alphanumeric words (eg 601AKC). $$H(D) = -\sum_{w \in W_D} p_w \log_2 (p|w),$$ 'where D is a document, H(D) is the Shannon entropy of document D, WD is the set of unique words occurring in document D, and pw is the probability of encountering one of these words at a random point in the text—that is, the frequency of that word as a percentage of the total word count.'
unique_tokens	Counts the number of unique 'tokens' using R's Quanteda package. Does not include front matter, endnotes, and tables of contents. Splits hyphens. Excludes stopwords, numbers, and alphanumeric words (eg 601AKC), as per cleaning for 'substantiveTokenCount'.
mean_word_length	Arithmetic mean length of tokens that appear in the 'substantiveTokenCount' analysis.
act_refs_data	A text search for the number of references to 'Act' that appear in the text of the legislation. The code then deletes results that are immediately preceded by any of the following terms: '[Tt]his', 'An', ')', '[Tt]hat', '[Tt]he', or '[Aa]pplication'. The ALRC identified these terms did not indicate an external cross-reference.
section_crossref_data	A text search for references to dozens of search terms that indicate a cross-reference, such as 'section' followed by a number, 'last paragraph' (which is common in older Acts), and 'next

Column name	Description
	sub.{0,2}?section'. Results preceded by '[Tt]his' were removed (as in 'this section'), as this term suggested the result was not a cross-reference to another provision of the Act.
Conditional_statements_word_count	Counts the number of the following that appear in the text of the legislation: 'if', 'except', 'but', 'provided', 'when', 'where', 'whenever', 'unless', and 'notwithstanding'.
Obligations_word_count	Counts the number of the following that appear in the text of the legislation: 'must', 'shall', 'may not', 'prohibited', 'required', 'may only', and 'cannot be'.
Must_not_word_count	Counts the number of references to 'must not' in the legislation.
Offences_word_count	Counts the number of references to '[Oo]ffence' in the legislation.
Offences_language_word_count	Counts the number of the following that appear in the text of the legislation: 'commits an offence' and 'guilty of an offence'.
Reasonableness_word_count	Counts the number of references to 'reasonabl.*' in the legislation.
Modifications_word_count	Counts the number of the following that appear in the text of the legislation: 'omit.*', 'insert', and 'substitute'
Discretions_word_count	Counts the number of the following that appear in the text of the legislation: 'Minister.{0,40}?may' (thereby capturing, for example, Minister for Defence may), 'Governor-General may', and '[Rr]egulations may'.
Legislative_instruments_word_count	Counts the number of references to '[Ll]egislative instrument' in the legislation.
Regulations_word_count	Counts the number of references to '[Rr]egulations' in the legislation.
Civil_liability_language_word_count	Counts the number of references to '[Cc]ivil penalt.*' in the legislation.
Civil_liability_word_count	Counts the number of references to '[Cc]ivil penalty:' in the legislation.
Good_faith_word_count	Counts the number of references to 'good faith' in the legislation.
Unfair_word_count	Counts the number of references to 'unfair.*' in the legislation.
Fair_word_count	Counts the number of references to 'fair.*' in the legislation.
Unjust_word_count	Counts the number of references to 'unjust.*' in the legislation.
May_word_count	Counts the number of references to 'may' in the legislation.
parliament	The term of Parliament in which the Act received Royal Assent.
parliamentDate	The date on which the term of Parliament began. This is the day after the election that preceded the new Parliament and is not the date on which the new Parliament formally opened, which can be months after the election.