Topic Refinement for Use with NLP

Zachary Elkins, Roy Gardner, and Ashley Moran

Rationale for Topic Refinement

Our methods using the CCP ontology and NLP tools to analyze Chilean consultation responses enabled detailed analysis of voluminous data. They also revealed challenges in redeploying an existing ontology created for human consumption instead for use with NLP tools. Using the CCP ontology with NLP tools revealed much about how the topic formulation and structure can shape their ability to accurately match with semantically similar text.

CCP topics that were well-specified and highly differentiated from each other matched very well to appropriate text segments. These included, for example, CCP topics on constitutional adherence, children's rights, environmental protection, freedom of expression, international law, municipal structures, right to health care, right to life, right to shelter, separation of religion and the state, and structure of the judiciary. These are all highly specified topics that do not have similar topics in the CCP ontology, and NLP tools were able to match them with great precision to appropriate text segments. However, topics that were similar to other CCP topics and topics that had under-specified or ambiguous descriptions were more difficult to use with NLP tools.

Similar topic sets often *over*-matched to text on related topics. For example, CCP topics differentiating many types of equality with respect to gender, race, national origin, and other features all matched to text on a single type of equality on, say, gender. CCP topics on a range of trial procedures—such as the right to a fair trial, right to a public trial, right to a speedy trial, and right to a jury—all matched to constitutional text related to any one of these rights. Likewise, NLP tools had difficulty differentiating between similar topics sets on constitutional amendment, criminal procedures, political parties, voting, well-being, institutions like the executive and legislature, and institutional features like selection, removal, and terms of office.

Under-specified topics often *under*-matched to text. For example, topics with undefined key terms such as 'indigenous people' did not match to text on those topics. And, topics with descriptions that included unrelated or conflicting concepts diluted the topic meaning and *mis*-matched topics to unrelated text. For example, topic descriptions that included text on what a topic does *not* include or conditions that may or may not apply—meant to be helpful to human analysts—fed incorrect information into the NLP tools that caused them to mis-match topics to unrelated constitutional text.

Initial Revisions

Initial CCP topic refinements, with NLP applications in mind, sought to retain what worked in the ontology and change what didn't. We revised all topics in the CCP topic set, with the following principles in mind:

- Make the fewest changes possible.
- Use the phrasing of the current topic label and description to the extent possible.
- Cut extraneous text that is not central to defining the topic and that seems to dilute the topic meaning and match topics to unrelated text. Specifically, this cut explanatory text that introduced unrelated or confounding topics, as well as general words common to many topics, such as provision, condition, requirement, national, state, citizen, right, duty, establishment, jurisdiction, government, and constitution.
- Expand topic descriptions to add (i) differences between similar topics, such as adding more robust definitions that distinguish between specific types of equality, (ii) specific

cases of general concepts, such as adding “proportional” and “majoritarian” to general election system topics, and “unicameral” and “bicameral” to the legislative structure topic, and (iii) definition information from the CCP codebook.

- Add key CCP topics that are in the CCP codebook but were not in the pared-down topic set used on Constitute. Specifically, this added topics for democracy, rule of law, minimum voting age, and social security.
- Move essential topic keywords into the topic description. This supports our plan to remove the keywords field from encoded text to avoid repetition seen across the description and keyword fields, ensure a uniform approach across topics (since some topics didn’t have keywords), and remove keywords that dilute the topic meaning.
- Standardize grammatical formats in case this may affect topic matching. Specifically this (i) wrote topic labels in the most concise form—e.g., labeling a topic “voting restrictions” rather than “restrictions on voting,” (ii) wrote topic descriptions in active voice where possible—e.g., noting a topic “prohibits certain political parties…” rather than “certain political parties… are prohibited,” and (iii) wrote descriptions in the form of either a noun followed by prepositional phrase or (where needed) a verb followed by noun and prepositional phrase.

These refinements to cut extraneous text and expand topic descriptions generated more robust, distinct definitions. For example, 17 “equality topics” in the original CCP ontology had largely the same topic label and description, but for a few words naming the specific type of equality protected (in bold italics in Table 1). Our revised CCP ontology thus elaborated and differentiated the specific types of equality and removed extraneous text, as Table 1 shows.

Table 1. Sample equality topics in original and revised CCP ontologies

	Topic Description	
Topic Label	**Original CCP Ontology**	**Revised CCP Ontology**
Equality regardless of ***origin***	Requires that everyone is treated equally before the law, without regard to their ***country or place of origin***. This may apply to both public and private interactions in some jurisdictions. Mention of anti-discrimination, ethnicity, non-discrimination, race.	Equality regardless of a person's ***country of origin, birthplace, native home, motherland, fatherland, or ethnicity***; protection from discrimination based on ***country of origin***.
Equality regardless of ***race***	Requires that everyone is treated equally before the law, without regard to their ***race***. This may apply to both public and private interactions in some jurisdictions. Mention of anti-discrimination, ethnicity, non-discrimination, race.	Equality regardless of a person's ***race or group defined by shared history or social identity***; protection from ***racism*** or discrimination based on ***race***.

Ontology revisions substantially reduced the incidence of inaccurate matches, and increased by more than 200% the number of text segments matched above our baseline similarity threshold. In particular, revisions successfully differentiated topics within topic sets. This disambiguation is evident in measurements of the semantic distance between topics within some sets in Figure 1.

In Figure 1, the blue curve represents the distribution of distances between topics in the original CCP ontology, while the orange curve that of the revised CCP ontology. In most topic sets (in panels of Figure 1), the orange curve is positioned to the left of the blue, which indicates that its inter-topic similarities are comparatively lower, and that its topics are more highly differentiated. For example, revisions successfully differentiated similar topic sets related to age restrictions, criminal procedure, equality, indigenous rights, and well-being, seen as distinct shifts left in the “age,” “crim_proc,” “equal,” “indigenous,” and “well_being” curves in Figure 1. Revisions also differentiated a subset of each of the executive, legislature, removal, selection, and term topic sets, seen as a flattening or a shift left in part of these curves.

Figure 1. Similarity of topics by category in original and revised CCP ontologies[1]

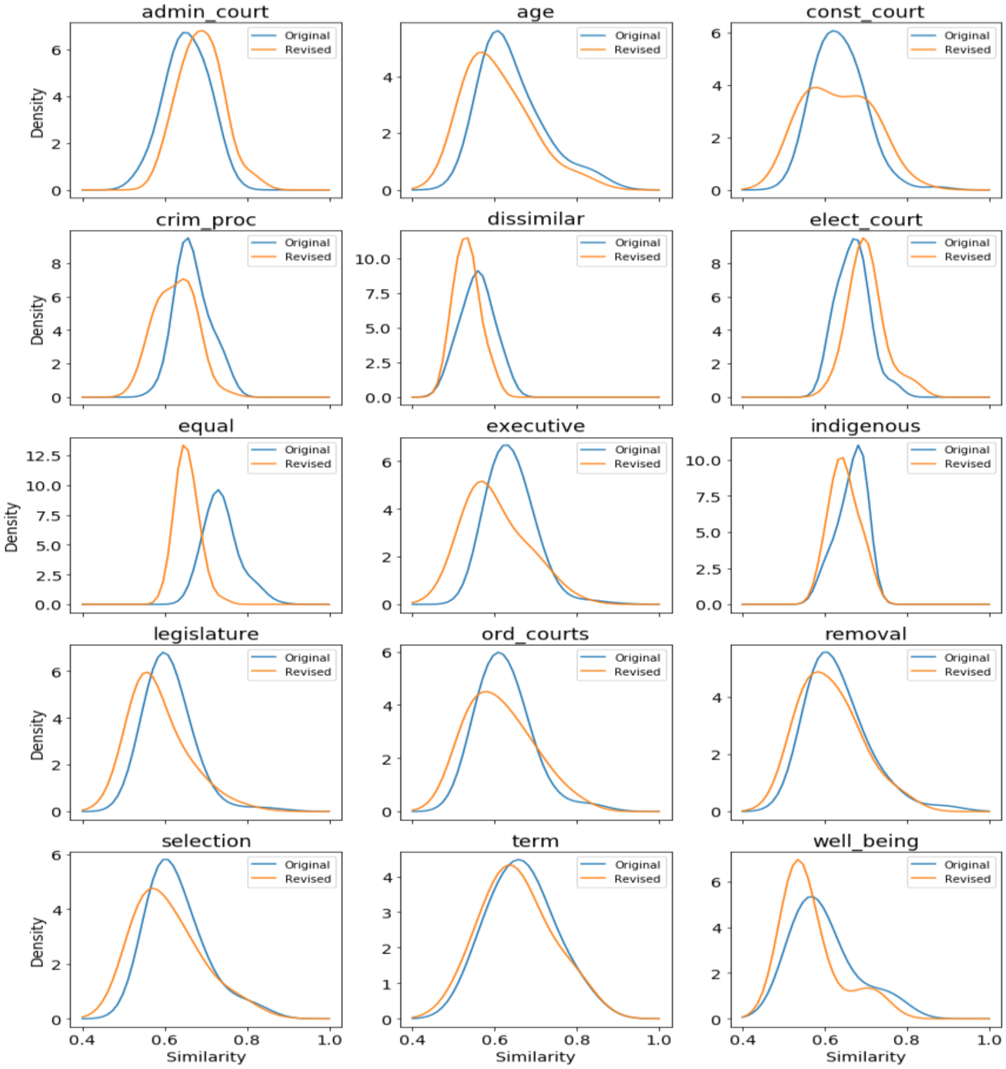
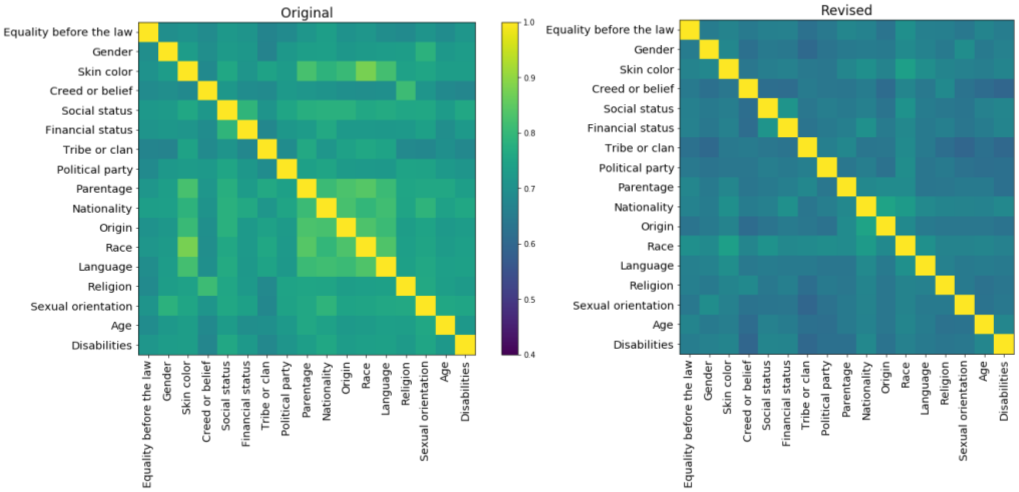


Figure 2 shows the impact of topic refinement on specific topic pairs, using equality topics again as an example. Overall, topic pairs became less similar after refinement, indicated by a spectrum shift from lighter boxes in the original ontology to darker boxes in the revised ontology. Also, the reduction in the number and intensity of bright spots on the map suggests that the topics that were initially *most* similar—those on equality regardless of skin color, parentage, nationality, national origin, race, and language—were successfully differentiated by the revisions.

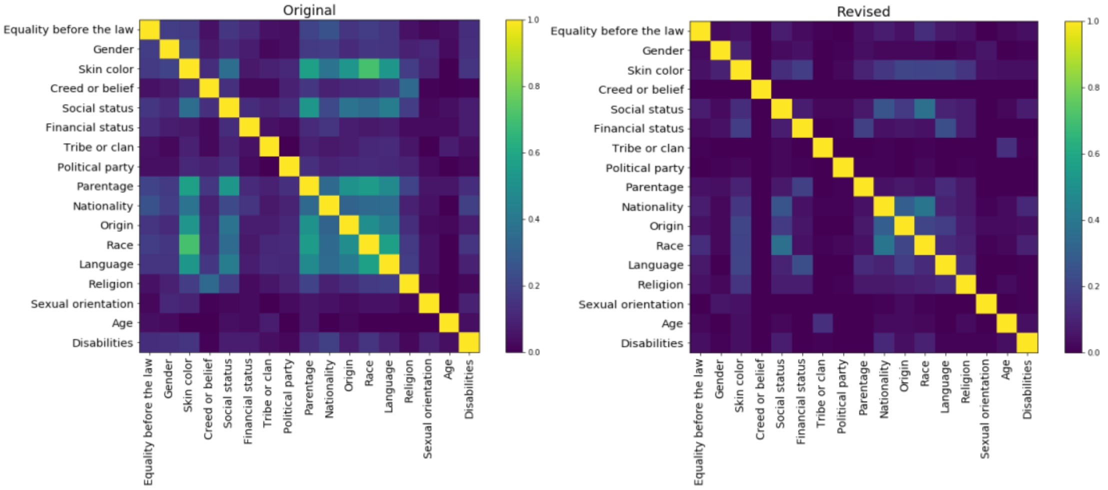
[1] Semantic similarity scores range from zero to one, where a score of 0.0 means the text segments share no meaning, and a score of 1.0 means they are identical.

Figure 2. Similarity of equality topic pairs in original and revised CCP ontologies



The increase in topic differentiation also affects the patterns of topic matches to corpus text. Figure 3 shows the overlap in corpus sections captured by equality topics—i.e., the portion of corpus sections matched to *both* topics in a topic pair. This neighborhood effect is another indication that topics, as specified, may be too similar to one another. Overall, the overlap decreased after topic refinement, seen as a shift from lighter boxes in the original ontology to darker boxes in the revised ontology. The implication is that the refined topics are more discriminating in matching to corpus text.

Figure 3. Overlap of corpus sections found by equality topic pairs in original and revised CCP ontologies



The effect of topic differentiation is also evident upon review of individual topic matches to corpus text. For example, topic revisions substantially reduced incorrect matching on equality topics. Table 2 provides examples of topic matches to text segments in Chilean public consultation responses. Use of the revised ontology seems to have successfully avoided matches of specific equality topics to general equality text. Note that correct matches are shown in green and incorrect in red.

Table 2. Sample text matches to equality topics in original and revised CCP ontologies

	Topic Match and Score	
Consultation Response Text	**Original CCP Ontology**	**Revised CCP Ontology**
Equality. Guarantee equal rights. Guarantee equality before the law. Eliminate social gaps. Start by defining dignity and basic concepts that must be guaranteed, linked to the human rights of every citizen. [Text ID reg/10180]	Equality before the law: 0.73 Equality regardless of social status: 0.71 Equality regardless of political party: 0.71 Equality regardless of creed: 0.71	Equality before the law: 0.73
Equality before the law. [Text ID reg/10300]	Equality regardless of gender: 0.71 Equality regardless of creed: 0.71 Equality regardless of race: 0.7 Equality regardless of religion: 0.7 Equality regardless of social status: 0.7	Equality before the law: 0.71
Equality before the law. Proportionality regardless of social class guaranteed by the constitution with due process. [Text ID reg/13968]	Equality regardless of social status: 0.74 Equality regardless of financial status: 0.7 Equality before the law: 0.7 Equality regardless of religion: 0.7	Equality before the law: 0.71 Equality regardless of social status: 0.7

Text source: Chilean public consultation responses

Importantly, the ontology revisions also differentiated *dis*similar topics. We checked this to ensure that topic refinements did not negatively affect topics that were already well-differentiated in the original ontology and performing optimally in the matching (e.g., compulsory voting, environmental protection, freedom of expression, rights to health and housing, and separation of church and state). The distance analysis confirmed that topic refinements did not make dissimilar topics more similar and, in fact, further differentiated them from each other, as seen in the “dissimilar” graph in Figure 1.

Yet, while initial topic refinements successfully differentiated many topic sets, they did not fully differentiate topics within some categories, such as institutions. Remaining challenges stem in part from the hierarchical structure of the CCP ontology. Topics nested under only one category in the CCP ontology can be effectively differentiated in the current structure. For example, for equality topics, the features added by equality subtopics (gender, race, etc.) appear as stand-alone topics by and large only under the equality topic. The distinctiveness of the second concept added by the subtopic thus allows these similar topics to be differentiated. On the other hand, topics nested under several categories in the CCP ontology pose challenges for differentiation in the current structure.

For example, for institutional categories (head of state, head of government, constitutional court, administrative court, lower chamber, etc.), the features added by subtopics (term, selection, removal, powers, etc.) are stand-alone topics under all of those institutional categories. Thus any success in making the conceptualization of the subtopic more robust under one category (say, elaborating the “term” component of the *head of state term* topic) to differentiate it from other topics in that category (*head of state selection, head of state removal,* etc.) also makes it more similar to “term” subtopics in other categories (*head of government term, constitutional court judge term,* etc.). And vice versa: the more we elaborate the “head of state” concept at the category level to distinguish it from other executive categories, the more similar all topics become within the “head of state” category.

Faceted Revisions

To address the matching challenges that remained after initial topic revisions, we experimented with rewriting CCP institutional topics using a “faceted” approach to classification. This sought to allow the algorithm to better distinguish and match these difficult topic sets. The faceted approach to classification aims to define stand-alone, mutually exclusive topics for each feature of a corpus. This is distinct from a hierarchical approach to classification, which creates subtypes under a superordinate category, thus creating a topic for each combination of superordinate and subtype features (such as *head of state term*, falling under the “head of state” category). A faceted approach, on the other hand, retains broad categories that are each associated with a single topic (such as *head of state* and, separately, *term*).

Our second set of CCP topic refinements thus retained changes made in the initial CCP topic revisions above, while separating any ‘problematic’ hierarchical topics into faceted topics. For institutional topics, for example, we first removed all topics that combined superordinate and subtype features (such as *head of state term, head of state selection, head of government term, head of government selection,* etc.). To replace these, we then created a new topic for each institution (*head of state, head of government, cabinet, administrative court,* etc.) and a new topic for each institutional feature (*term, selection, removal, age, immunity*, etc.). This aimed to create more differentiated topics and reduce opportunities for matching errors.

Table 3 provides examples of institutional topics matched to constitutional text. It shows the topics matched to two constitutional texts, using the original, revised, and faceted CCP ontologies. Use of the revised (but still hierarchical) ontology added additional accurate matches (shown in green), but did not remove all extraneous matches (shown in red). Use of the faceted ontology successfully avoided extraneous matches for these topics.

Table 3. Sample text matches to institutional topics in original, revised, and faceted CCP ontologies

	Topic Match and Score		
Constitution Text	**Original CCP Ontology**	**Revised CCP Ontology**	**Faceted CCP Ontology**
To be President of the Republic or Vice President, it is required: to have natural Paraguayan nationality [Text ID Paraguay_2011/983-5]	Head of state eligibility: 0.63 Head of state term: 0.63 Head of state succession: 0.63 Requirements of citizenship: 0.63	Head of state eligibility: 0.68 Head of gov’t eligibility: 0.66 Head of state age: 0.65	Head of state: 0.65 Head of government: 0.63 Nationality: 0.69 Country of origin: 0.65
The members of the Constitutional Court take the oath of office during a solemn ceremony presided over by the President of the Republic before Parliament, the Supreme Court, the Council of the State and the convened Court of Government Accountability. [Text ID Gabon_2011/505]	Const’l interpretation: 0.67 Const’l court establishment: 0.66 Const’l court opinions: 0.65	Const’l court: 0.68 Const’l court removal: 0.66 Const’l court selection: 0.66	Const’l court: 0.68 Const’l oath: 0.65

Text source: Current constitutions

Our initial revisions provide robust evidence that the topic refinement and faceted approaches presented here can increase the resolution and differentiation of topics.