# 28/06/2024

## Notas a partir da leitura do Codebook de Castro

Fui lendo o codebook e anotando questões importantes para pensar os procedimentos de uma análise automatizada replicando os resultados.

**Sobre uso de codificação manual ao invés de automatizada**

“The coding was carried out by hand. This allowed us to also code agreements or disagreements that were not easily identifiable through sentence construction (e.g. by the use of verbs such a support or oppose), but whose identification needed an understanding of the substantive topics under discussion” (Castro and Kammerer 2017)

### Relatórios e dados a serem incluídos ou excluídos

- Use of Daily ENB reports.Exclude first daily report of the meeting, as it is summary of previous/context; If there are multiple daily reports, final ENB report is also excluded

On procedures for 'interventions' variable: use of semi-automatic procedure, counting mentions to country.

For this part, exclude sections that do not summarize respective day negotiations, commenting on previous rounds or outcomes.

"These include mainly an introductory paragraph, a section summarizing the previous negotiations history (“A brief history of the UNFCCC and the Kyoto Protocol” is a common heading for such a section), a section displaying informal comments by experts and negotiations on the current status of negotiations (“In the corridors”), and a section providing actual analysis of the meeting’s negotiation process and outcomes (“A brief analysis of the XXX meeting”)." (p. 16)

“Countries may at times be mentioned in the ENBs when they are not actually speaking at the negotiations (e.g. to describe the nationality of the chair of a meeting, or the nationality of the speaker of a coalition, or the author of a specific proposal being discussed). To avoid counting these instances, we used the fact that the ENBs generally use full capital letters (e.g. “INDIA”) whenever they refer to a country or coalition when it is speaking at the negotiations, while they use normal capitalization (“India”) whenever a country is mentioned but not as a speaker.” (p. 16)

Para as coalizões, pode aparecer como o nome por extenso ou por sigla.

### Formato dos dados

"Each observation in the dataset consists of a bargaining interaction between a country dyad at a specific negotiation day about a specific negotiation topic. The nature of such interaction may be cooperative (country 1 “speaks on behalf of”, “supports”, “agrees with” or “speaks with” country 2) or conflictive (country 1 “criticizes”, “opposes” or “attempts to delay a proposition by” country 2). [...] Only those interactions that actually took place are included. Thus, country pairs that did not interact on a specific negotiation day are not listed for that day. On the other hand, if a country pair interacted several times ithin a day, then each of these interactions is considered a separate observation" (Codebook, p. 2)

- Variable "relation", which codes type of interaction is very detailed. Draws on specific words and formulations to identify type of relation. These words/formulations may appear in different sentences, which might be a challenging complexity.

Sometimes, 'agreement'/'opposition' might be inferred, not explicitly written in the report (Codebook, p. 5-6). This will not be attempted for now. Hopefully, this will not generate huge distortions. If so, we might rethink.

- Variable "topic" also has very well documented coding rules. I will not attempt to recode it automatically, but might test topic modelling in the future.

## Agregação dos dados

Na base, temos dados entre díadas de países, apresentados de forma desagregada e a cada dia. As interações são classificadas em múltiplas relações, mas Castro as ordena em um gradiente de cooperação-oposição e sugere linhas de corte.

No paper do *Does it matter who participates*, eu tive que agregar um ‘score de cooperação’. Na ocasião, optei por uma soma onde cooperação e oposição se anulam [Checar o procedimento que eu adotei]

Lendo Castro et al. (2014), vemos uma proposta alternativa de agregação, o % de supportive statementes:

Formally, if we define *zijt* as the variable that counts the instances in which country *i* makes a statement in support of country *j*'s position during negotiation day *t*, and *nj* as the total number of statements by country *j* during the negotiations between December 2007 and December 2009, we can express our dependent variables, the percentage of supportive statements, as: [*85 é o total de dias de neg. no artigo*]

Uma imagem contendo Forma

Descrição gerada automaticamente

## Plano inicial para a expansão da base

Já tenho o código da coleta e os dados devidamente coletados. Se quiser expandir depois, é só rodar pro período seguinte. Por hora, vou tentar replicar os dados de Castro a partir do que já tenho. Provavelmente, quando eu expandir pro período que ela não cobre, vou ter que revisar algumas coisas como novas coalizões (nome por extenso + sigla do ENB). Antes disso, porém, vou tentar fazer a coisa funcionar em um nível comparável ao da base já existente.

1. Coleta dos ENBs e organização de planilha com os dados relevantes (Data da negociação, evento de referência, etc.)
2. Limpeza da base como um todo: excluir os relatórios que não me interessam
   1. Segundo os procedimentos de Castro, devem ser mantidos apenas os DAILY REPORTS.
   2. Destes, devem ser excluídos i) o primeiro relatório, normalmente dedicado a um resumo/contexto prévio e ii) o último relatório, normalmente resumindo as negociações e analisando resultados. Inclusão do último só se daria quando só houver um único relatório. Não é difícil criar essa regra, mas vale dar uma analisada manual no que estiver sendo excluído.
3. Para cada relatório: organização dos dados para a análise
   1. Exclusão das seções que não resumem as negociações – introdução, resumo das negociações, etc.  *Aqui não sei como é a facilidade para identificar e limpar. Se não for simples (como no caso de headers ‘history of xxx’), posso checar se uma solução rápida como retirar 1º§ é muito ruim ou dá pro gasto*.
   2. Separação em parágrafos como unidades de análise
4. Nos parágrafos, identificar:
   1. Há países/coalizões sendo mencionados? Se não, excluir. Se sim, identificar os SPEAKERS (em caixa alta).
   2. Há outros países/coalizões sendo mencionados na mesma frase do SPEAKER? Se sim, identificar como os objetos da relação. Se não, checar no resto do parágrafo.
   3. Há uma estrutura clara para a relação? E.g., PAÍS A apoia País B. Se sim, já classificar como interação correspondente. Se não, testar modos mais complexos de interação (e.g., países em frases distintas, etc).

É nesse ponto que a coisa vai começar a ficar mais complexa. Vou construir os códigos até essa etapa e, ao chegar, começo a analisar mais a fundo os caminhos para a classificação. Aqui, vou ter que olhar pras várias categorias e construir certa sequência de análise. É meio trabalhoso e vai ter muito teste na ida e volta, mas acho que é viável pras estruturas mais claras. Nas com frases diferentes/concordância e oposição não explícita, não sei bem a viabilidade, vemos depois.

Quando estiver aceitável, rodar para todo um período e comparar scores com o da base de Castro.