Transnational solidarity
at times of crisis
TransSOL

Reliability Report for Work Package 5: Claims Analysis and Social Media Commenting Analysis

Annex to Deliverable 5.1

TransSOL: European paths to transnational solidarity at times of crisis: Conditions, forms, role models and policy responses

WP 5: Media Analysis: Collective Identities and Public Solidarity

Work package leader: Science Po

WP participants: USIEGEN, UNIGE, GCU, UOC, UNIFI, UNIWARSAW, UCPH

Project Information
Project Type: Collaborative Project
Call: H2020 EURO-3-2014: European societies after the crisis
Start date: June 2015
Duration: 36 months
Coordinator: Prof. Dr. Christian Lahusen, University of Siegen
Grant Agreement No: 649435
EU-funded Project Budget: € 2,483,805.00

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 649435.

www.transsol.eu

TransSOL Reliability Report for Work Package 5: Claims Analysis and Social Media Commenting Analysis

This report documents the quality checks conducted for claims coding in TransSOL. Due to its qualitative nature, the coding of social media commenting, being a bit smaller in scale, was trained and monitored at the beginning of the research project and continuously checked by the team leaders. Furthermore, the social media commenting analysis was based on many of the variables already used for claims-making, for which intercoder-reliability had been established before. This section is therefore dedicated to the more comprehensive claims-coding only.

Validity and Reliability in TransSOL Claims Coding

Due to the logistic effort that comes with conducting an analysis of claims in a multi-team setup, we decided to test reliability across teams in English as well as decentralized within teams in the original language. This was on the one hand due to the fact that when using such a more interpretive method of coding, the language context and the broader discourse on the issue under analysis matters a lot for interpretation; and on the other hand, because the actual coding of claims was conducted in the original language while training and instructions, i.e. also the codebook, was in English. There was always at least one person in every team with a close-to native level of English and extensive experience with international cooperation. For the claims-coding, however, coders were hired from outside the project context – mostly Master or PhD students. Not all coders, while being able to follow training discussions, were that advanced in English which is another reason why we decided to split the test in two phases: An *Inter*- and an *Intra*-team reliability test.

Phase 1 included a test of reliability across teams (inter-teams reliability test) as well as a validity test for claims identification based on instructors' coding. We used English language material from an abridged version of the Greek Katherimini online newspaper (http://www.ekathimerini.com), assuming that the English used here would be less complex and easier to code than using, for example, a sample of The Guardian. In that sense, we tried to create equal conditions for all coders, also such with a lower level of English language skills. Coding for this test was conducted as a team effort. Thus, the coding that was submitted to work package leaders for the calculation of reliability scores was the result of decisions made in the team (i.e., n=8 reliability coding samples for 8 teams in total). This was to ensure that the rationale of coding was the same across teams.

Technically, teams were here provided with 10 articles from Katherimini from which they first identified claims for testing the reliability of claims identification. The articles were retrieved using the same key words as for the overall analysis (refuge* OR asyl*). The set of claims identified by teams was then checked by work package leaders and a set of valid claims identified (n=20). This set of valid claims was then re-submitted to teams for the reliability coding of variables. Following the line of reasoning in literature evaluating similar data (e.g. Van der Brug et al. 2015), we measured the agreement on claims identification with percentage agreement: (All coding decisions-decisions deviating from majority)/all coding decisions. Example: 3 coders code 3 claims for reliability; for 1 claim, only two coders agree (= 1 deviating decision); all coding decisions = 3 (coders) * 3 (claims) = 9; 1 deviating coding decision; percentage agreement = (9-1)/9 = 0.89. For further information on the approach, see the documentation of reliability checking of the EUROPUB project (WP2) at https://europub.wzb.eu/codebooks.en.htm

For the validity test in *Phase 1*, we contrasted teams' identification of claims with a set of valid claims identified by instructors. Claims identification is the most crucial part of claims coding; on the one hand, it is essential to establish a common understanding of what a claim is to ensure even a basic comparability of the coding. On the other hand, all other coding of the actual claim variables builds on this common understanding; thus, the reliability of the whole coding process squarely depends on the correct identification of the unit of analysis. Measures for validity, especially *precision* and *recall* (e.g., Stryker et al. 2006) have become increasingly established in automated content analysis, where the validity of the computer's coding of a text is a more pressing issue that in manual coding. Nonetheless, these measures can also be applied for a validity test in a claims analysis. Recall provides a measurement of how many relevant items (= claims identified as valid by instructors) were selected (=identified by teams); whereas precision accounts for how many of the selected items (= claims identified by teams) were relevant (= identified as valid by instructors; see Stryker et al. 2006: 414/415). Table 1 provides an overview of results which were overall satisfactory, but required additional training for some teams, especially regarding the identification of valid claims (low recall).

Table 1: Precision and Recall as Measurements of the Validity of TransSOL Claims Coding

Instructors vs.	Precision	Recall
DK	0.87	0.95
DE	0.90	0.90
UK	0.83	0.71
FR	0.88	0.67
GR	0.81	0.81
IT	0.76	0.76
PL	1.00	0.57
CH	0.83	0.95
Average Instructors vs. Teams	0.86	0.79

To illustrate, the Polish team identified only valid claims, and therefore has a perfect precision value of 1. However, it did not identify all valid claims and therefore has a relatively low recall score. The Danish team identified some additional claims that were not labelled as valid by instructors (lower precision) but did in in fact only miss one valid claim (higher recall). Precision scores are more than satisfactory, teams with a recall of lower than .80 were given additional training for claims identification.

For the reliability test, scores are lower for latent variables that required more interpretation by coders (position and value) (e.g., Neuendorf, 2002). Table 2 presents the overall satisfactory results of the reliability test for all teams.

Table 2: Reliability Scores Across Teams (Inter-teams Reliability Test)

Variable	*Rel. Measure*	*Rel. Score*
Claims Identification	% Agreement	82%
Posit	K-alpha (% Agr.)[1]	0,75 (90%)
Actor	% Agr.	96%
Scope	% Agr.	96%
Nationality	% Agr.	86%
Form of Action	% Agr.	87%
Issue Migration Management	% Agr.	96%
Issue Integration	% Agr.	99%
Issue Background and Fate	% Agr.	90%
Issue assoc. with crisis	% Agr.	96%
Issue Public and Civic Activities	% Agr.	99%
Issue Other	% Agr.	100%
Value	% Agr.	73%

Note: 2 coders in the following teams: CH, DE, FR, GR, IT, UK; 3 coders for DK, PL; Krippendorff's alpha, while being the most established measure of reliability, is not well-suited to be performed on rare phenomena, especially in dummy variables due to already low variance (De Swert, 2012). *We therefore percentage agreement measures here and only use Krippendorff's Alpha for the only metric variable in the sample (POSIT).*

For the value variable, we checked reliability on dummy variables for each value of the variable. Results indicate that the interest-based value was difficult to distinguish from the other two values, i.e. rights- and identity-based justifications. This seems mostly due to the fact that in claims by political actors, one might always assume political calculations, even behind morally framed claims (see Table 3).

[1] For percentage agreement: deviations from the majority decision. We allowed an error margin of 1 (i.e., counting the difference between neutral and positive/negative and half a mistake only.). Reference for calculating percentage agreement was the number of claims coded multiplied by the number of teams.

Table 31: Reliability of Identification of Value and Value Categories

	% Agreement
Value 1: Interest-based	0.78
Value 2: Rights-based	0.91
Value 3: Interest-based	0.96
Presence of Value (yes/no)	0.79

The greater problem, however, was to determine whether there was a codable value or not. Here, some teams over-interpreted claims and coded values to a much greater degree than other teams. Therefore, we took a closer look at such claims for which all teams coded a value (N=6). While the number of claims admittedly is rather low, the scores are higher (see Table 4). Thus, we trained coders again on especially the identification but also the coding of value.

Table 4: Reliability of Value Variable tested on Claims for which all Teams identified Value as present

	% Agreement
Value	0.83

Phase 2 was conducted after the across-teams reliability check. It consisted of an intra-team reliability test since all teams employed more coders at the same time. The intra-team reliability tests were conducted in the respective country language on a sample of the newspapers used for coding. For an overview of these newspapers, please consult the Report for Work Package 5 as published on our website transsol.eu.

The work package was the last to be conducted in TransSOL, which in some cases was a logistic challenge: Some teams had to retrain coders and recode data, prolonging the overall process of data collection considerably. Therefore, in some cases, the intra-test could not be completed or documented which is why not all teams and also only the most prominent values in variables are documented in this annex (see Table 5).

Teams here proceeded in the same way as for Phase 1 reliability testing: team leaders first let coders identify claims from a sample of 10 articles and then coded a set of valid claims (n=20). However, this phase of coding was designed to ensure the reliability of coding *between* individual coders, which is why team discussions only took place after the identification and coding were concluded – thus, each coder coded the reliability sample independently from his/her colleagues in the team. For some countries included in this test, this entailed another re-check and correction of the data which then was conducted by team leaders, thus by expert coders.

Table 5: Reliability Scores for Individual Teams (Intra-teams Reliability Test)

	Rel. Measure	*DE*	*DK*	*GR*	*IT*
Claims Identification	% Agr.	81%	89%	96%	97%
Posit	K-alpha	0,78	0,49	0,32	0,83
	(% Agr.)	(96%)	(87%)	(85%)	(99%)
State Actors/Societal Actors	% Agr.	100%	100%	100%	100%
Claimant Scope = National	% Agr.	93%	95%	98%	100%
Nationality of Claimant = Domestic	% Agr.	95%	100%	100%	100%
Issue Migration Management	% Agr.	88%	90%	98%	92%
Issue Integration	% Agr.	88%	95%	95%	98%
Issue Background/Fate and Other Issues	% Agr.	85%	95%	93%	100%

Note: 2 coders in the following teams: DE, IT; 3 coders for DK and GR. For state actors, we aggregated all political actors (actor categories 1, 2,, and 9 in the WP5 codebook for the claims analysis) and subsumed the rest under 'societal actors'. For issue integration, issue categories 2 and 4 were aggregated; the issue category background/fate and other issues subsumes categories 3, 5, and 6 which did, overall, not come up very often.

Coders in the Greek and Italian team benefitted from having been part of a similar project before Moreover, also one coder of the German team had coded claims before, however with a slightly different definition of a claim. This specific coder was trained again and the coding re-checked by the team leader. The main difficulty for claims' identification was to distinguish claims from a mere description but also to distinguish one claim from another in the same text.

We trained coders again to improve the quality of data and re-checked already coded claims. This careful control was very important for the posit variable: the problem here, as it often occurs for the coding of tone, was to assess tone towards refugees as the object of the claim: In some cases, the tone would appear to be negative since the claimant criticized the government's decision regarding refugees, which in turn, however, would mean an expression of support – thus a positive evaluation – for refugees. In addition, it was in some cases difficult to decide if a claim was evaluative or neutral – in many cases, mistakes were made where one coder coded neutral whereas the other coder coded negative or positive.

The issue variable, especially in the German sample, was checked again by the team leader to improve the quality of the data. Due to the fact that a large part of the coding was conducted by one coder who left the team early, a re-checking of data by an expert was the only possibility to correct data in hindsight.

References:

De Swert, K. (2012). *Calculating inter-coder reliability in media content analysis using Krippendorff's Alpha*. http://www.polcomm.org/wp-content/uploads/ICR01022012.pdf

Neuendorf, K. (2002). *The Content Analysis Guidebook*. SAGE Publications.

Stryker, J. E., Wray, R. J., Hornik, R. C., & Yanovitzky, I. (2006). Validation of database search terms for content analysis: The case of cancer news coverage. *Journalism and Mass Communication Quarterly*, *83*(2), 413–430. https://doi.org/10.1177/107769900608300212