

# Content-analyzing Political Texts. A Quantitative Approach

**SONIA ALONSO  
ANDREA VOLKENS  
BRAULIO GÓMEZ**

**CIS**

Centro de Investigaciones Sociológicas

**CUADERNOS  
METODOLÓGICOS  
047**



9 788474 765922

# **Content-analyzing Political Texts. A Quantitative Approach**

**SONIA ALONSO  
ANDREA VOLKENS  
BRAULIO GÓMEZ**

## **CIS**

Centro de Investigaciones Sociológicas

**CUADERNOS  
METODOLÓGICOS  
047**

Centro de Investigaciones Sociológicas  
Madrid, 2012

Editorial Board of the «Cuadernos Metodológicos» Collection

DIRECTOR

Félix Requena Santos, *President of the CIS (Centro de Investigaciones Sociológicas)*

BOARD MEMBERS

Francisco Alvirra Martín, *Universidad Complutense de Madrid*

M.<sup>a</sup> Ángeles Cea D'Ancona, *Universidad Complutense de Madrid*

Jesús M. de Miguel Rodríguez, *Universidad de Barcelona*

Modesto Escobar Mercado, *Universidad de Salamanca*

J. Sebastián Fernández Prados, *Universidad de Almería*

Juan Ignacio Martínez Pastor, *Universidad Nacional de Educación a Distancia*

SECRETARY

M.<sup>a</sup> Paz Cristina Rodríguez Vela, *Director of the Publications and Research Development Department*

The publication standards and Instructions for authors can be found at:  
<http://www.cis.es/publicaciones/CM/>

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior written permission of the publisher.

«CUADERNOS METODOLÓGICOS» COLLECTION, ISSUE 47

Catalogue of Publications of the General Administration of the State  
<http://publicacionesoficiales.boe.es>

First edition, July 2012

© CENTRO DE INVESTIGACIONES SOCIOLOGICAS

Montalbán, 8 - 28014 Madrid

© Sonia Alonso

© Andrea Volkens

© Braulio Gómez

ALL RIGHTS RESERVED

Printed and made in Spain

NPD: 004-12-012-2

ISBN: 978-84-7476-592-2

Depósito legal: M-22678-2012

Typesetting and printing:

CASLON, S.L.

Matilde Hernández, 31 - 28019 Madrid



This book is printed on 100% recycled, chlorine-free paper.

## Table of contents

Introduction .....	5
<b>I. THE CONTENT ANALYSIS OF POLITICAL TEXTS .....</b>	<b>9</b>
1.1. The history of content analysis in political science .....	9
1.2. Content analysis: some important definitions .....	11
1.3. Content analysis step by step .....	13
1.3.1. Sampling .....	13
1.3.2. Unitizing .....	15
1.3.3. Quantifying and coding .....	17
1.3.3.1. Manual codification .....	19
The classification scheme .....	19
The coding handbook and coder training .....	21
1.3.3.2. Computerized codification .....	23
Dictionaries .....	24
Computer software .....	25
1.4. Validity and reliability .....	30
1.4.1. Correlative validity (cross-validation) .....	32
1.5. The virtuous circle of content analysis .....	36
<b>2. THE MANIFESTO PROJECT CONTENT ANALYTICAL APPROACH .</b>	<b>39</b>
2.1. The saliency theory of party competition .....	40
2.2. Content analyzing Manifestos .....	41
2.3. Selecting texts (sampling) .....	43
2.4. Selecting coding units (unitizing) .....	44
2.5. Classification schemes (coding and quantification) .....	45
2.5.1. Classification scheme of national-level manifestos .....	45
2.5.1.1. Sub-categories within the Standard Classification Scheme for national-level manifestos .....	47
2.5.2. Classification scheme of multi-level manifestos: territorial authority claims and multi-level policy preferences.	50
2.5.2.1. Territorial authority claims .....	50
2.5.2.2. Multi-level policy preferences .....	52
2.5.3. The classification scheme of regional-level manifestos .	53

2.6.	The coding handbook .....	56
2.7.	Code sheets .....	59
2.8.	Validity testing .....	60
2.9.	Reliability testing .....	63
2.10.	Coding exercises .....	65
3.	APPLYING THE CONTENT ANALYTICAL DATA OF THE MANIFESTO PROJECT .....	73
3.1.	The Manifiesto Project datasets .....	73
3.2.	The construction of indicators and scales .....	75
3.2.1.	Saliency and position measures .....	76
3.2.2.	How to create your own scale .....	78
3.2.3.	The left-right ideological scale .....	80
3.2.4.	The centre-periphery scale .....	84
3.2.5.	The immigration-anti-immigration scale .....	85
3.3.	Empirical applications of Manifiesto data .....	87
3.3.1.	The crisis of ideology and the median voter .....	88
3.3.2.	Parties, promises and policies .....	91
3.3.3.	Multilevel party competition .....	92
3.3.4.	New issues in electoral competition .....	94
3.3.5.	Second-order elections hypothesis .....	97
3.3.6.	The thesis of programmatic trans-nationalization .....	98
4.	Commented bibliography .....	101
5.	Bibliographic References .....	103
6.	APPENDIXES .....	117
6.1.	Definitions of the Classification Scheme for National and Regional Manifestos .....	117
6.2.	Sub-categories for Central and East European countries and for Mexico .....	129
6.3.	Solutions to coding exercises of regional-level manifestos .....	134
	Index of tables and figures .....	141

## Introduction <sup>1</sup>

This book is the result of a chance encounter at the Social Science Research Center Berlin (WZB). In 2006 the authors discussed their closely related research interests on national and regional parties' policy positions. Out of these discussions emerged the idea of investigating the newly emerging cleavage line between supra-national institutions, on the one hand, and sub-national institutions, on the other, with the nation-state level under increasing pressure from both sides simultaneously. However, we soon realized that for this initial idea to be feasible we needed to develop a methodology for measuring political parties' positions in multi-level electoral contexts. We started this endeavour by designing and writing a handbook on the content analysis of party programs in multi-level settings (Alonso *et al.*, 2012). At this point, the editors of the book series *Cuadernos Metodológicos* offered us the opportunity of broadening the handbook, placing it in the wider context of the methodology of content analysis, on the one hand, and the particular context of party manifestos analysis, on the other. Accordingly, the book now comprises three major parts addressing three inter-related audiences.

Chapter 1 takes a «do-it-yourself» approach to the content analysis of political texts, addressed at political science students and researchers who want to test their own research hypotheses through the use of this methodological tool. The landscape of publications explaining the methodology of content analysis in Spain is a poor one, particularly if we compare it with other places, such as Germany and the Anglo-Saxon countries. Only two Spanish translations of important reference books in English and French exist and they are already quite old. One is Laurence Bardin's *L'analyse de contenu* (1977), translated into Spanish in 1986; the other is Karl Krippendorff's *Content Analysis: An Introduction to Its Methodology*, translated from the first 1980 version into Spanish in 1990. A new updated version of Krippendorff's 1988 book was published in 2004 but has no Spanish translation yet. To our knowledge there

---

<sup>1</sup> The work reported in this book has been funded in part by the Ministry of Economy and Competitiveness (Spain) through the National Plans of Research, Development and Technological Innovation (PN I+D+i) under the project *The Effect of Political Decentralization on Parties' strategies. Extending the Comparative Manifestos Project to the Spanish Regional Elections* (CSO2009-11241).

are only two text-books on content analysis originally written in Spanish; one dates back to 1981 and the other was published in 2001<sup>2</sup>.

While many content analytical textbooks are available in English and German, most of them focus on media research<sup>3</sup>. To be sure, mass media provide important source texts for political science studies and all content-analysis rests on models of communication. But freshmen need to know that there are many different types and techniques of content analysis. The choice of one type over another critically hinges on the theoretical concept to be measured by the content-analysis technique. Because political science concepts differ from those of media research, the type of content analysis to be chosen differs as well. In this book, we concentrate on approaches prevalent in current political science research and we use abundant examples from recent publications.

Content analysis is back at the top of the agenda in political science, in great part due to the advance in computer software and hardware and to the availability of machine-readable texts. This methodology has increasing scientific returns the larger the number of political texts that can be dealt with simultaneously. Human coding may be less attractive in this respect than computer coding but it is equally necessary. Without the human codification approach, computer-based content analyses would be unable to test the validity of their concepts and results and without validity there is no possibility for scientific advancement. Computer-based content analysis will still need to compare its results with those of human coding in order to ascertain the correlative validity of the computerized findings and to calibrate its results. Human-coding will continue to be a necessary first step for the development of classification schemes based on particularly complex concepts. In this respect human coding will always serve as the pre-test of a classification scheme on a sample of texts before moving on to large coding production with computer software. Human-coding will continue to be an inevitable first step in the construction of dictionaries to be used by software packages.

Although major breakthroughs have taken place in computerized approaches to content analysis, there are still limitations to what you can expect from computer-coding. Ahead of us is still a long way before we can teach computer software packages how to capture «meaning» the way humans do. Un-

---

<sup>2</sup> Jesús Martín-Barbero's *Introducción al análisis de contenido* (1981) and Javier Gil and Víctor Hugo Perera's *Análisis informatizado de datos cualitativos* (2001), respectively. Apart from these two books, Pablo Navarro and Carolina Díaz have a book chapter entitled «Análisis de contenido» in the edited book by Juan Manuel Delgado and Juan Gutiérrez *Métodos y técnicas cualitativas de investigación en ciencias sociales* (1994). Jaime Andréu Abela has a working paper on content analysis techniques entitled *Las técnicas de Análisis de Contenido: Una revisión actualizada* (2002), as part of the collection Documentos de Trabajo of the Fundación Centro de Estudios Andaluces. Textbooks in Spanish on discourse analysis are relatively more abundant but discourse analysis is just one type of content analysis, one that will not be covered by the present book.

<sup>3</sup> There are some books in Spanish on content analysis applied to the mass media, but they are all quite old (Janowitz, 1965; Kientz, 1974; Clemente, 1991; Serrano Pastor, 1993; Royo Vela, 1997).

doubtedly, this will be one of the most active areas of development of this methodology in the future. This book may well be a first step for those who want to venture themselves in this future agenda because they will need a good understanding of human coding approaches before they move on to computers.

Chapter 2 of the book addresses those students and researchers who want to apply a particular approach of content analysis, namely the analysis of electoral programs for measuring political parties' policy preferences and ideological positions. In this respect, the book will dwell mainly on the classification scheme of the Manifesto Research Group/Comparative Manifestos Project (from now on Manifesto Project) which was developed at the beginning of the 1980s in order to estimate the policy preferences of national parties by content analysing their national election programs. Since then, the Manifesto Project classification scheme has been used to content-analyze the national election programs of parties from more than 50 countries since the Second World War as well as the speeches of party leaders and government declarations. A late side development of the Manifesto Project has been the adaptation of the classification scheme to multi-level politics, in order to measure not just policy preferences but also territorial authority claims at regional and supra-national electoral levels. Until now, the Manifesto Project has clearly demonstrated that its classification scheme travels well across countries from different world regions, across types of political actors, and across research questions.

The Manifesto Project has become the most used empirical reference in the study of political parties. In this sense, it is a success story. Over the last 30 years, the simple idea of measuring policy preferences of parties by content analyzing election programs has grown into a whole research field (Robertson, 1976; Budge, Roberston and Hearl, 1986). There is no other dataset that allows us to measure the evolution of political parties' ideological positions and electoral competition since the Second World War until the present. There are more than a hundred academic publications that have made use of the Manifesto's work and data in one way or another during the last thirty years. Moreover, fast-growing computerized techniques of content analysis for measuring party policy positions need the Manifesto data in order to test the validity of their findings.

Having been managed by the Comparative Manifestos Project (CMP) from 1989 to 2009, the Manifesto data collection is now being updated and extended in a new phase of the project — the Manifesto Research on Political Representation (MARPOR) — with particular concern for analyzing processes of party-citizen representation in Latin America. It is thus particularly appropriate at this point to publish in Spanish a guide to this content-analysis methodology. The second chapter should facilitate the extension of the Manifesto Project data collection to Latin American countries and serve as a preliminary guide to the procedures to be applied in selecting, preparing and analyzing them.

For the existing Manifiesto data collection, this chapter is a valuable documentation for political scientists operating in Spanish, be it in Spain or in Latin America <sup>4</sup>.

Chapter 3 of the book caters to users of the Manifiesto Project data collection. Manifiesto data can be applied to test a broad range of hypotheses in which parties play a role because the concept of policy preferences travels far and wide among different political science theories. Therefore, the third chapter presents a summary of the main research questions and hypotheses that have been tested in political science using the Manifiesto data while at the same time offering many tips about how to use the data for your own particular research.

While each chapter addresses a slightly different audience, they are closely inter-connected and, therefore, we think that there is good reason to combine the three parts under one book. Firstly, we hope that the success story represented by the Manifiesto Project example will help students of content analysis to see the final pay-offs of applying this slow, time-consuming, and repetitive methodology. Secondly, the Manifiesto Project is one of the very few truly cross-cultural approaches of content analysis. Opposed to most content-analytical approaches developed for one particular language and country or one politico-cultural area only, the Manifiesto Project approach proved its applicability to different languages, to parties from established industrialized countries as well as from newly developing democracies, and to cultures as different as those of the USA, Sweden, Israel, Turkey, and Korea. Thirdly, the approach was developed to travel across times. Today, it covers data from six decades between 1945 and 2012. This scope demonstrates one of the most important advantages of content analysis compared to other research methods, namely, that it can be applied retrospectively to texts written in former times. Last but not least, the presentation in detail of a concrete content analysis methodology like that of the Manifiesto Project serves to highlight the importance of the concept-validity-reliability triangle at the core of every content analysis technique.

We hope that this book will serve, in one way or another, as an appetizer to many students and researchers and encourage numerous innovative approaches of applying the method to new research questions.

---

<sup>4</sup> In English, it has been covered in the Mapping Policy Preferences series by Oxford University Press (Budge *et al.*, 2001; Klingemann *et al.*, 2006).

## The Content Analysis of Political Texts

Political texts are the most widely available source of evidence about politics. Practically everything that gets done in politics is recorded in protocols or minutes, issued as directives or laws, reported as proceedings of committees or legislative bodies, or even broadcasted. Political causes are advocated in recorded speeches and interviews, pamphlets, posters, and, of course, party platforms and manifestos.

### 1.1. The history of content analysis in political science

Content analysis has a long history. Its initial steps as a methodological approach took place in the realm of media research and political communication (Berelson, 1952). The idea was to read mass communication texts, particularly newspapers, in a systematic way in order to derive empirical data for testing scientific hypotheses. Content analysis soon spread beyond these journalistic roots into various fields of social research, including political science, where the focus on newspapers gave way to the analysis of presidential campaigns, presidential inaugural speeches and political propaganda already in the 1940s. According to Holsti, «[d]uring the 1940s, political research using propaganda materials accounted for nearly 25 per cent of all empirical content analysis investigations» (Holsti, 1969: 21). At this point in time, the influence of content analysis in political science studies was so dominating that Kaplan defined content analysis as «the statistical semantics of political discourse» (Kaplan, 1943: 230). Lasswell's *The Language of Politics* (1949) represents, in this respect, a peak of content analysis with a focus on political documents. By the late 1960s, 75 percent of all content analysis research took part in three disciplines: socio-anthropology, general communications and political science (Holsti, 1969).

A landmark of this development was the introduction of computers. The first computerized content analysis was reported by Sebeok and Zeps in 1958 (Krippendorff, 2004: 13). Until the late 1960s, the most concerted work in com-

puterized content analysis was done with a computer program called «The General Inquirer», developed by the social psychologist Philip J. Stone (Janda, 1967: 13). The English-language bias of the early General Inquirer was later overcome by new, more general, versions such as TextPack<sup>1</sup> (Mochmann, 1980; Klingemann, 1984).

However, the development of automatic coding by computers experienced a bottleneck during the 1980s, due to the limited availability of machine-readable texts. As Monroe and Schrodt have put it, «[t]he manual entry of texts was at least as costly and time consuming as simply coding them directly from paper or microfilm [...] and consequently human coding remained the norm» (2009: 352). The situation changed with the advance of the internet. The Web «revolutionized the availability of texts» (Monroe and Schrodt, 2009: 352) and, as a result, computer-based content analysis has grown considerably during the last decade.

Today, content analysis —computerized and manual— is an established method in media research. Regular courses are taught in master studies of communication and ever increasing numbers of publications are based on the method (Neuendorf, 2002). In political science, in contrast, the 1940s to 1970s were the halcyon days of content analysis. Thereafter, few content-analytical approaches became known beyond highly restricted audiences in subfields of the profession, with few exceptions. First, policy agenda studies are the bread and butter of political science in the USA. Early studies on agenda-setting in the media were followed by research on the agenda-setting power of the media (Dearing and Rogers, 1996). The latter studies do not only map changes in media attention to issues over time but also measure their impact on the agenda of governments. Changes in the agenda of US governments, its causes and consequences constitute a whole subfield in US political science (Schattschneider, 1960; Bachratz and Baratz, 1962; Cobb and Elder, 1983; Baumgartner and Jones, 1993; Kingdon, 1995; Jones and Baumgartner, 2005), but few attempts have existed to make the underlying content analytical approaches comparable across countries (Baumgartner, Green-Pedersen and Jones, 2006). Second, content analysis has a well developed research tradition in Germany, maybe due to a lingering influence of the Brothers Grimm. In the commented bibliography at the end of the book we recommend two German textbooks for those acquainted with the language. Third, content analysis is a well-established technique for measuring party positions. Publications devoted specifically to the topic of estimating parties' positions are widespread (Budge *et al.*, 2001; *Electoral Studies*, 2006; *German Politics*, 2009; Klingemann *et al.*, 2006; Laver, 2001). The reason for this success story in political science is that party positions play a major role in many normative and analytical theories of parties. While some party theories directly touch on the contents of party programs, there is a host of related research questions to which content analysis

<sup>1</sup> For a textbook in Spanish about how to apply TextPack, see Alaminos y Penalva (1995).

of election programs can provide answers. Fourth, although to a lesser degree, content analysis is a technique applied in empirical research on political discourse.

## 1.2. Content analysis: some important definitions

There are as many definitions of content analysis as there are content-analytical textbooks. Most of these textbooks start out from—or at least refer to—the famous Lasswell's formula (1948): «Who says what in what channel to whom with what effect?» While Lasswell was concerned with studying mass media and political propaganda, the formula can be used to address all kinds of communication. It directs our attention to five components of communication—the communicator, the message, the channel, the receiver, and the effect of the communication—and, accordingly, five areas of communication research based on these components. For this reason, content-analysis research starts with a theory about the message to be analyzed: «Why available texts came into being, what they mean and to whom, how they mediate between antecedent and consequent conditions ...» (Krippendorff, 2004: 82). These aspects are called the contexts of the text.

Political science is mainly concerned with three types of analyses: the internal analysis of a text, the causes of the text, and the effects of the text. The first and most obvious type is the internal analysis of the message. Typical political science questions concerning an internal analysis are: What positions do authors take in the text? How important are the positions to the authors? These descriptive questions take an analytical turn once causes or effects come into play. Theoretically, causes can be related to the communicator, the channel, or the receiver of the message. For instance, we may hypothesize that the position-taking may depend upon the left-wing or right-wing ideology of the politician, on the fact that it was drafted for an internal party meeting or a speech in an election campaign, or on whether she addressed her core voters or the floating voters in the middle of the ideological spectrum. Effects in this example could simply be whether the receiver understood the message or whether the message left an impact on voting behaviour. These examples show that, once causes or effects come into play, the analysis of the content of the text itself is expanded by additional information on the communicator and/or the receiver of the message or even extended by additional data such as voting behaviour.

While such questions on the content of political messages, its causes, and effects are familiar from everyday talk about politics, there is a crucial distinction to the content analytical method. Most content analysts agree with Holsti (1969) when he characterizes content analysis as objective, systematic, and general in its core features. The method is objective due to the rules and procedures which must be formulated for each step of the research process and

recorded in a handbook; it is systematic because the rules must always be applied; and it is general because the analysis must have theoretical relevance. These three characteristics will be further elaborated in the following sections of this chapter:

Krippendorff provides the most technical definition of content analysis: «[A] research technique for making replicable and valid inferences from texts (or other meaningful matter) to the contexts of their use» (2004: 18). We prefer this definition for two reasons. First and foremost, Krippendorff reminds us that content analysis is a technique that needs to be learned before it can be applied. Second, Krippendorff's definition contains two of the most important technical terms of content analysis, namely reliability and validity. In ordinary life, the number of inferences to be drawn from a text is tantamount to the number of readers. Content analysis, by contrast, is a research technique that requires drawing *replicable* and *valid* inferences from texts. Replicable means that all who apply the technique arrive at the same results. Reliability describes the degree to which this is achieved. Validity means the extent to which the results of the content analysis correspond to the theoretical concepts the content analysis set out to measure. Validation is the measurement process by which the results of content analysis are corroborated by facts. Reliability and validity differentiate content analysis from the interpretation of texts in ordinary life and will, therefore, be the focus of a separate section in this chapter:

At this stage of our introduction, we turn to the overall goal of content analysis, the data reduction process. Reducing data refers to the efficient representation of large volumes of data, i.e. quantification. Content analysts disagree about whether the method should always be quantitative or may also be qualitative. In this book we disregard qualitative approaches because in political science the method is used to reduce large amounts of texts to manageable numerical figures. In any case, we agree with Berelson (1952: 114-134) who, while acknowledging the qualitative features of quantitative content analysis inherent in the transfer from texts to numbers, considers that the final goal of the procedure is the quantification of contents. We cannot delve deeper into this methodological debate at this point. Instead, we refer the interested reader to a classic article by Kracauer (1952).

In content analysis, data reduction proceeds in three steps. The first step concerns the classification of text units. Armony and Armony (2005), for instance, analyzed the discourse in an internet forum in order to study the way in which national identity was conceived by Argentines during the 1991 financial and political crisis. In a first step, they marked (among others and with the help of computers) the keyword «Argentina» in all texts entered in a specific forum on national identity. The tagging of the keyword is, in this example, the first step of reducing a weeklong discussion to a manageable amount of numbers, i.e. a (1) whenever the keyword appeared during the discussion.

In the second step, the amount of numbers produced in step 1 was further reduced. There are many possibilities of reduction. The most widely used

method is to report the *frequency* of appearance (absolute or relative). In our example, one could count how many times the word «Argentina» appeared in the contribution of each participant. The use of frequencies, however, is based on two implicit assumptions: first, that the frequency with which the word appears is a valid indicator of importance of this term for the person who uses it, also called its *saliency*. Depending on interpretation, importance and saliency are tantamount to the concern, the focus of attention, or the intensity of an item. Second, it is assumed that each time the keyword «Argentina» appears it should be given equal weight, i.e. each case of occurrence counts the same as any other; even if in some cases an adjective or attribute in the sentence may give special importance to some instances of occurrence. These assumptions are questioned by researchers that defend a more qualitative approach to content analysis.

Once the data have been summarized through a method such as the frequency, content analysis proceeds to the third step. It then tests the research hypotheses (i.e. patterns and relationships within the findings) through several techniques such as cross-tabulations, associations, and correlations. In our example, multivariate analyses such as factor analysis, multidimensional scaling, contingency and cluster analysis, semantic nodes, and profiles (Krippendorff, 2004: 191-211) are possible once several keywords have been marked. Going back to our example, the third step could then result in a typology of national identities by identifying underlying dimensions of national identity and national identity type scores for each participant.

### 1.3. Content-analysis step by step

Content-analyzing a political text involves decisions about sampling, unitizing, and coding. At the most general level, the latter refers to the choice between human and computerized content analysis to place the units of sampled texts into groups with shared meanings<sup>2</sup>. This section deals with sampling and unitizing because these steps apply to both human and computerized techniques. Thereafter, two separate sections introduce human and computerized coding.

#### 1.3.1. SAMPLING

Sampling is the process by which the researcher limits the number of observations to be analyzed to a «subset of units that is statistically or conceptually

---

<sup>2</sup> A seldom used type of content analysis, the network approach, places relations between objects or persons into groups such as «coalesce» or «does not coalesce», see Carley 1997 for an introduction to this specific method and Kleinnijenhuis, Ridder and Rietberg 1997 as an example for application.

representative of the set of all possible units, the population or universe of interest» (Krippendorff, 2004: 83). Unfortunately, there is no simple rule that tells us how many units to collect except that it depends on your research question. The less case- and time-specific your hypothesis, the more you are in need of selecting subsets of texts from a broad universe of texts.

For example, consider the question of the policy payoffs obtained by the coalition partners of a certain two-party coalition government in country X. If this coalition existed only twice, you can select the whole universe of two coalition agreements and two times two election programs (by each party in the coalition) to be content analyzed and compared. If, however, you are interested in testing hypotheses about the policy payoffs of coalition partners in general, the rules of statistical significance come into play: you need at least thirty units in order to approximate a normal distribution.

Therefore, you need to do more than randomly select 30 cases of coalition agreements from the universe of all coalition agreements across countries and times. Debus (2008), for instance, selected 35 coalition agreements according to two criteria: he took only coalition agreements from countries where party competition had previously been shown to be based on economic and social policy, and he selected countries to represent three different types of party competition, namely uni-polar, bipolar, and multi-polar competition. Thus, he mixed a most-similar system design regarding the dimensions of competition with a most dissimilar system design (Anckar, 2008) regarding the types of competition. In technical terms: he held the dimensions of party competition constant in order to investigate the impact of different types of party competition on policy payoffs of coalition partners.

The choice of coalition agreements as the texts to be content-analyzed is a straightforward one when studying the policy payoffs of coalition partners. However, decisions on sampling can be more difficult when there is no text specifically devoted to the concept of interest for the researcher. What would you do if you would like to analyze the impact of globalization on party competition in national election campaigns? The research project «National Political Change in a Globalized World» (Kriesi *et al.*, 2008) selected as target documents for content analysis two newspapers, one quality and one tabloid, during the two months preceding the election day, for four election campaigns in six different countries. The objective was to track down a new cleavage line that was hypothesized to be on the rise as a result of globalization. In all six countries, three campaigns from the 1990s to the early 2000s and one from the middle of the 1970s were included. The campaign of the 1970s represented an election before globalization set in, allowing for a comparison between the pre-globalization period and the times of globalization. All articles dealing with politics in the two newspapers during the two months before the elections were included in the sample.

The mix of quality papers and tabloids is a regular feature in media studies in order to give a comprehensive account of media content. You may, however,

question the decision to constrain the analysis to two months before each election. Why not one, three or four months? Concerning the duration of election campaigns, there is variation between as well as within countries. If you are in need of determining such variations for making a case selection, more often than not you have to check your source texts to make sure that you do not err on two sides: if a shorter period is chosen you may miss large parts of the campaigns, but if a longer period is chosen you may include cases before the campaign took off.

The kind of texts picked as the object of analysis is tantamount. The importance of this decision is shown by the results of a recent analysis on the relevance of populism in the speeches of 41 chief executives in different countries (Hawkins, 2009). In his analysis, Hawkins distinguished four types of speeches: speeches held during election campaigns, ribbon-cutting inaugural speeches of infrastructure projects, speeches held in international meetings, and famous speeches to which special importance was attributed for some reason. The result of this cross-country comparison was that presidents with a populist inclination used populist appeals in campaign speeches twice as much than in ribbon-cutting, international, or famous speeches. Obviously, you would have underestimated the extent of their populism if campaign speeches had not been included in the analysis; however, you would have overestimated the degree of populism if your analysis had included campaign speeches only. This result points to the impact of different receivers on the content of political texts. In other words, the context of the text affects the content of the text. The interpretative corollary is that presidents use populism for mobilizing voters; populism may be a linkage strategy of elites with the masses.

### 1.3.2. UNITIZING

Unitizing refers to the designation of the units to be coded or, in other words, the segment of the texts selected from the universe that is placed in a given category. The units mostly used in political science are the *single word* (or symbol), the *sentence* (or parts of it), and the *item* (the entire speech, article, film, book, etc.). Again, there is no simple rule that can tell you when to choose which coding unit. The only advice we can give without knowing anything about your research question is to try out different units once texts have been sampled.

Single words, for example «environment», are mostly used in computerized content analysis. For many research hypotheses, all the researcher needs to know is the number of occurrences of single words. Consider a simple hypothesis of pressures from emerging new parties on the established parties. You may hypothesize that once Green parties enter the parliament they force all other parties to address the environment. Suppose you have all machine-

readable versions of all the speeches in two parliamentary periods, before and after the entry of the Green party. You could then use a computer program to identify all occurrences of the terms «protection» and «environment» and count them in order to test your hypothesis. This you could do because your hypothesis is just concerned with topics (or thematic concerns).

However, the choice of single words as your units of analysis for this particular example is less straightforward than it may seem. There are many ways to address a topic. Therefore, you need to identify all variations of a term such as «environment». In order to do so, the term is reduced to the stem, in our case «environ-». All occurrences of the stem are tagged. Moreover, you need to identify all the different ways of expressing concern for a topic. For example, instead of talking about «environmental protection», the government may be «concerned with nature». On the other hand, it may be misleading to include all occurrences of the term «environment» as it is also used in phrases such as «an environment of globalization» or «the work environment». Therefore, words as coding units are sometimes interpreted in contextual units (mostly sentences) to make sure that they are captured correctly. Despite these caveats, taking words as coding units is a very efficient way to address many research questions, in particular if confronted with large volumes of texts.

Instead of topics, political scientists are often interested in positions, i.e. topics with directions attached, and in frames, i.e. arguments that justify positions. Continuing with the example above, the researcher may want to know whether the government took action for more environmental protection. Obviously, no single word captures such action-oriented position<sup>3</sup>. Therefore, political scientists often recur to sentences as the natural units of arguments. Since long sentences may contain more than one argument, long sentences are sometimes divided into quasi-sentences.

A unitizing method based on quasi-sentences as units is the *core-sentence approach*, widely used by political scientists during recent years (Kriesiet al., 2008; Dolezal, Helbling and Hutter, 2010). We use a variation of an example provided by Helbling, Höglinger and Wüst (2010: 508) to demonstrate the advantages of this unitizing method:

The Democratic Party supports the new constitution of the European Union because it makes decision-making efficient, but opposes the accession of Turkey to the European Union because we fear rising unemployment rates.

This sentence contains two different positions and, therefore, two core sentences (Table 1.1).

For both quasi-sentences, the actor, the topic, and the direction of the preference are identified. The direction is established as pro (+1) or contra (-1). This method is to be preferred for texts that address more than one actor, as

<sup>3</sup> Koopmans and Ersanilli (2011) call this position-taking claims.