Eszter Katona [1], *Mihály Fazekas* [2]

Hidden barriers to open competition: Using text mining to uncover corrupt restrictions to competition in public procurement

Working Paper series: GTI-WP/2024:01

April 2024, Budapest, Hungary

[1] ELTE & Government Transparency Institute

[2] Central European University & Government Transparency Institute

Abstract

Public procurement accounts for one third of government spending across the world, while it is also particularly vulnerable to corruption. Large amounts of open administrative data enabled a rich literature on measuring corruption. However, scholarship largely focuses on structured information on government tenders, neglecting text fields which are particularly suitable for hiding wrongdoing. To address this gap, this article identifies strategies for limiting competition by tailoring tendering terms to a favoured bidder. We argue that subtle, text-based strategies are employed by corrupt actors when more visible strategies for favouritism, such as non-competitive tendering procedures, are undesirable or impractical. Using data on all published government tenders in Hungary between 2011-2020 of 119,000 contracts, we deploy a host of traditional regression and advanced machine learning models such as Random Forests. We find that specific phrases in bidding conditions, product descriptions and assessment criteria lead to single bidding in otherwise competitive markets. Including texts improves model accuracy from 77% (structured variables only) to 82% (structured and all text data together). We unpack our complex machine learning models by pinpointing terms conducive to deliberate market access restrictions such as overly specific bidding eligibility criteria. We demonstrate that text mining has the capacity to advance our understanding of corrupt behaviours and to better target anti-corruption policies.

Keywords: public procurement, text mining, corruption, measurement, Hungary

Acknowledgements

The authors are grateful for numerous colleagues commenting on earlier versions of this draft, in particular to participants of conferences: 11th Annual Conference on New Directions in Analyzing Text as Data, ECPR Joint Sessions (Measuring Corruption: State of the Art, Challenges, and Advancements); ICSA 2022 (Challenges and recent advancements in corruption risk assessment); and Szociológiai tudás és közjó. MSZT 2022. évi Vándorgyűlés. In addition, we are indebted to Pál Susánszky, Balázs Váradi, and Renáta Németh who shared detailed comments on earlier drafts.

Introduction

Public procurement represents 15% of global GDP and about 1/3rd of total government spending, encompassing everything from school meals to nuclear power plants. As such spending is largely discretionary and highly complex, it is particularly vulnerable to corruption. Correspondingly, allegations of corruption and politicians favouring connected companies are rife in both high and low integrity countries. Thanks to a wide coalition of government, civil society, and business reformers (Adam et al, 2020), public procurement has evolved into one of the most data rich government functions with hundreds of millions of contracts published on various government publication websites and open data repositories.

This combination of large amounts of public resources allocated, high risk of corruption, and unprecedented open data spawned a large literature in the last 10-15 years, proposing novel measurements for corruption and corruption risks (*Lyra et al, 2022*). Aside for qualitative studies, the literature has almost exclusively focused on structured information on government tenders such as the procedure type followed (e.g. open tenders versus direct awards), fulfilling publication requirements (e.g. publishing the call for tenders) or supplier characteristics (e.g. tax haven registered company) (*Fazekas et al, 2018*). It has largely neglected textual information which is particularly suitable for hiding favouritism because the high degree of technical, financial, and legal complexity makes the insertion of seemingly benign, but competition-restricting conditions easy. Hence, in spite of intense scholarly and policy interest, we still know too little about subtle forms of corruption, their magnitude, and what drives them.

To address this gap, this article makes use of hitherto under-utilized textual data in government tenders to study corruption and favouritism in competition for government contracts. Specifically, we aim to

> *identify strategies for limiting competition for government contracts by tailoring bidding conditions to a potentially favoured company.*

By doing so, we expand on existing corruption risk measurement frameworks with the use of textual information.

Our starting point is the understanding of corruption as limited access to public resources, that is unjustified restriction of competition in public tenders for the benefit of a connected bidder(s) (*Fazekas et al, 2016*). A large number of corruption strategies have been identified in the qualitative literature, many of which have also been estimated in large-scale administrative datasets around the world. Against this background, we argue that subtle, text-based corruption strategies are employed by corrupt actors when more visible strategies for corruption, such as non-competitive tendering procedures, are undesirable or impractical. Hence, our conceptual framework expands on the repertoire of identified corruption strategies and gauges trade-offs and substitutions between more and less visible strategies.

A large part of existing corruption risk indicators is identified and validated using predictive models with single bidding on competitive markets as dependent variable (*Fazekas and Kocsis, 2020*). This literature predicts single bidding with the use of structured procurement information on the products purchased, the characteristics of the tender and its outcomes. This research takes these models as a starting point and further improves their performance by adding text-based indicators. Given that procurement texts can describe a range of tendering features, we explore which types of textual information are most important for predicting corruption risks, understood as limited competition in public tenders.

We analyze online available, official government data on about 120,000 Hungarian public procurement contracts from between 2011 and 2020. The Hungarian dataset is part of the Government Transparency Institute's Global Contracts Database and can be accessed at opentender.eu/hu/download. We use text mining methods to extract, pre-process and analyze the textual and structured information. First, we replicate past research predicting a single bid submitted on an otherwise competitive tender, making use of a host of structured product and market features such as contract value (control variables), in combination with well-documented corruption risk indicators such as non-open procedure types (baseline models). Then we train Logistic Regression, Random Forest, and Boosting models adding word n-grams and text meta-features to the baseline models.

Our findings point out that the models using textual information outperform the replicated baseline models in predicting single bidding. This confirms our expectation that subtle, text-based corruption strategies can be used in addition to other, more visible strategies for achieving corrupt ends. To explore which texts are most important for measuring corruption risks, we trained different models using texts containing bidding requirements for bidding firms, the award criteria used to score bids, and product descriptions. We found that award criteria are the least impactful for predicting single bidding, while the text in product description has the highest predictive power. Unpacking the highest prediction accuracy model, demonstrates that frequent restrictive technical and financial conditions coupled with highly specific product descriptions greatly increase the probability of single bidding in Hungary.

Our text-as-data approach contributes to both the academic literature and policy applications aimed at understanding, measuring, and identifying corruption. First, unlike most text-as-data approaches which only use textual information for predicting single bidding and related outcomes, we explicitly combine already established models using structured variables with text-based features. This allows us to assess the added value of text mining methods in the literature. Second, we depart from the text-as-data literature in the field by building on all readily available text fields in public procurement announcements. This allows us to better understand different types of competition restrictions hidden in different legal and technical texts, some applying to the product, some to the bidder, some to the bidding process itself. Qualitative evidence pointed out the simultaneous relevance of all 3 text types, yet explicit testing and comparisons in large-scale text analysis has not been done yet. Third, we unpack our complex and flexible machine learning models on top of identifying the best performing model, again, going beyond the state-of-the-art in this field (e.g. Acikalin et al, 2023; Modrušan et al, 2020). Opening up the black box of predictive models allows both to directly relate to qualitative studies predicting certain types of restrictive terms, and to offer risk predictions which are more actionable for practitioners. Finally, our novel results in corruption risk measurement aligned with prior literature can also serve as a valuable input into better targeting anti-corruption policies and for corruption investigations such as the precise identification of investigative leads.

The manuscript is structured as the following: First, we outline the conceptual framework of the analysis defining corruption in public procurement and how corrupt transactions are conducted. This section derives empirical expectations guiding the empirical analysis. Second, we introduce our large-scale administrative data and methods. Third, we review our findings and finally we outline our conclusions and areas for further development.

Conceptual framework

Following prior literature analysing corruption risks in public procurement, we understand corruption predominantly as a limitation to open access to public resources (*North et al, 2009*). This framing departs from often used principal agent theories or accounts equating corruption with bribery. The key expectation of non-corrupt, well-functioning public procurement markets as laid down in laws and also supported by academic theories is open and fair competition. This means that all those companies who reasonably can deliver the requested goods and services should be able to bid and should be assessed fairly (*Yukins, 2007*). These principles are violated when certain bidders are treated unequally, for example by being excluded even though they could reasonably participate in the tender. This violation of principles of good public procurement chime with broader concepts in political science revolving around impartiality in the implementation of public policies (*Rothstein and Teorell, 2008*).

Hence, we define corruption in public procurement as the allocation and performance of public procurement contracts by bending prior explicit rules and principles of open and fair public procurement in order to benefit a closed network while denying access to all others (*Fazekas et al, 2016*). This definition implies a specific measurement approach. Indicators follow from this definition which capture biases in the procurement process that are typically used for favoring a selected bidder. Moreover, those indicators also follow from this definition which point at successful competition restriction, that is tendering process outcomes indicating limited competition and repeated success of the same firm. For example, when a public buyer artificially creates a situation of emergency (e.g. deliberately announcing the tender late given known project deadlines) and uses it to award a non-competitive contract to a connected company, we talk about a corrupt scenario. However, it is important to bear in mind that low competition is not equal to corruption, instead when competition is deliberately limited to favour a particular bidder is when we can talk about corruption.

A number of measurement instruments have been proposed on the basis of this theoretical framework leading to a wide range of promising and a few validated corruption risk indicators, or proxies (*Fazekas et al, 2018; Gnaldi et al, 2021, Villamil et al, 2023*). Almost exclusively, these indicators make use of structured fields such as procedure type used, contract value, or bidder location (**Table 1**). The attractiveness of these indicators is that they rest on readily available or at least readily processible information in public procurement administrative records and datasets (*Fazekas and Saussier, 2018*).

Table 1. Overview of selected corruption risk indicators in public procurement

Source	Indicator(s) used	Country	Year	Sector
Di Tella and Schargrodsky (2003)	Difference in prices of standardised products such as ethyl alcohol	Argentina	1996–2007	Health care procurement
Olken (2007)	Differences between the officially reported and independently audited prices and quantities of road construction	Indonesia	2003–2004	Infrastructure (roads)
Hyytinen, Lundberg, and Toivanen (2008)	Number and type of invited firms; use of restricted procedure	Sweden	1990–2008	Cleaning services
Bandiera, Prat, and Valletti (2009)	Price differentials for standard goods purchased locally or through a national procurement agency	Italy	2000–2005	Standardised goods (e.g. paper)
Klašnja (2015)	Single bidder auctions; non-open procedure types	Romania	2008–2012	General procurement
Chong, Klien, and Saussier (2015)	Negotiated procedure type	European Union	2008–2012	General procurement
Auriol, Flochel, and Straub (2016)	Exceptional procedure type	Paraguay	2004–2007	General procurement
Coviello and Gagliarducci (2017)	Number of bidders; same firm awarded contracts recurrently; level of competition	Italy	2000–2005	General procurement
Ferwerda, Deleanu, and Unger (2017)	Contract level elementary risk indicators such as short advertisement period	EU	2006–2010	General procurement
Fazekas and Kocsis (2020)	Composite risk score including elementary indices such as single bidding, or short advertisement period	EU	2009–14	General procurement
Luciánodra, Milani, Millemaci (2022)	Composite score with focus on contract complexity	Italy	2007-2017	Public works
Decarolis, Giorgiantonio (2022)	Composite risk score including absence of tender call, page and word number of calls, open tender days.	Italy	2009–2015	Public works (sector of goods, services, and works)

Source: adapted and extended from Fazekas et al, 2018

This rich prior literature has identified and validity tested a range of indicators which proxy corrupt strategies in public procurement. The underlying strategies typically make use of features of the tender which are easily visible, verifiable to outsiders. For example, not advertising a call for tenders on a government publication website is by default verifiable for auditors, civil society or interested bidders. Implementing these strategies requires some legal expertise regarding some key features of public procurement rules, e.g. rules defining when direct awards can be made as opposed to running an open tendering procedure. However, they rarely require sophisticated technical and economic skills which are needed for tailoring tendering terms to a pre-selected bidder. These different features, visibility versus technical sophistication, are what set aside hitherto extensively studied compared to understudied corrupt strategies.

A corrupt group, in particular a corrupt procuring entity, will resort to visible, but easy to implement strategies when it is confident that monitoring agents will not uncover or punish it for corruption; or when it is confident that its corrupt acts will look legal. However, whenever it needs to hide its corrupt dealings more carefully or the legal framework makes it hard to conceal corruption, it will resort to more subtle methods of excluding non-connected bidders and favouring those with connections. Among the wide range of corruption techniques identified in the academic and policy literature (e.g. OECD, 2007), tailoring tendering terms favouring a particular bidder are by far the most widely used, at least based on case studies. Hence, tailoring tendering terms to a particular company and its products can be used to constrain competition in public procurement (i.e. limit access to public resources) *in addition to* visible, formal corruption strategies. These arguments lead to the following hypothesis:

> *H1: Constraining conditions and criteria in tender texts represent an avenue to unjustified limitation to competition in addition to formalistic, visible procurement process biases.*

However, not all sections of the tendering documents can be used to limit competition in the same way. There are 3 major areas of the tender documentation which can be used for subtly favouring a certain bidder:

1. **Product description**: This section of the tender documentation precisely defines the products (goods, works, and services) that are purchased (*Gorgun et al, 2020*). Hence, this is where the specific products of the favoured company can be targeted, in essence excluding all non-connected competitors with substitute products.
2. **Eligibility criteria**: This section of the tender documentation defines the preconditions whose fulfilment is necessary for any eligible bidder (*Rabuzin and Modrušan (2019*); *Modrušan et al., 2020*). This is where unwanted competitors can be excluded without being even considered as bidders.
3. **Award criteria**: This section of the tender documentation defines the scoring rule for eligible bids, that is once a company passed the eligibility criteria and its products fit the product description. Typically, this section only defines price as the main criteria for ranking submitted bids. However, when different price-related criteria and quality features are scored, a range of subtle scoring rules can be inserted which favour the connected bidder.

Nevertheless, each of these tender texts may only partially represent additional strategies to formalistic, visible competition restrictions. When the reason for non-competitive procurement procedures is product specificity or uniqueness, restrictions in the product description may correspond to the use of non-open procedure type and the non-publication of call for tenders. However, biases in award criteria are only needed for corrupt goals when there is competition expected between the connected and non-connected bidders, that is the procedure type is open and there was a call for tenders published.

More broadly, we can expect from an efficiently operating corrupt group to use these different tendering sections flexibly, placing the competition constraining conditions in the parts which fit the given context best. For example, if a favoured firm has a unique product, tailoring the product description is the optimal way for corrupting the tender. Hence, there is no need to place further competition constraining conditions in the other sections. However, if the favoured company only sells generic products e.g. office chairs, inserting favoritistic conditions in the product description will be hard. This will make the other 2 sections more attractive avenues for corruption, by for example enabling competition but including unfair or easily gamed award criteria (i.e. conditions which are subjective hence can be used for favouring a connected bidder). Given the different functions of the 3 sections of the tender documentation and their correspondingly different corruption potential, we hypothesize:

> *H2: The 3 different main parts of the tender documentation (eligibility criteria, product description, and award criteria) can be used in additive, partially overlapping corruption strategies.*

Data and indicators

Public procurement data

We use official governmental data on Hungarian public procurement between 2011 and 2020. The database is derived from online published public procurement announcements at the national publication portal www.kozbeszerzes.hu as collected and processed by the Government Transparency Institute. For full information see: https://opentender.eu/hu. The database contains all public procurement tenders and contracts conducted under Hungarian Public Procurement Law. The information is published in standard publications forms such as

1. Calls for tenders,
2. Contract award notices, and
3. Contract modification and correction notices.

As not all these kinds of announcements appear for each procedure (e.g. most non-competitive tendering procedures do not need to publish a call for tenders), we only have the variables deriving from contract award notices consistently across every procurement procedure.

As the source announcements are published as html pages, a web scraper algorithm was used to collect the source information. Then a structured database was created by parsing the html data into a pre-defined structure, containing variables with clear meaning and well-defined categories such as standardized procedure types or contract award announcement dates following the DDMMYYYY format. Furthermore, errors, inconsistencies, and omissions found in the source data were corrected or removed. After the data cleaning processes the final database has sufficient quality for scientific research as prior publications with this data demonstrated it (*e.g. Fazekas et al, 2016*). For a full description of database development, see Fazekas and Tóth (2016).

The novel aspect of this database, which has not yet been analyzed is a corpus of raw texts. Six different text fields are available, which can be grouped into three main types:

1. Product description: tender title, tender description;
2. Eligibility criteria: personal, technical, and economic requirements; and
3. Award criteria: scoring rules for submitted bids.

Crucially, for the interpretation of the results and for understanding the limitations of the data, these text fields do not correspond to the full tender documentation, rather only encompass the summary and key fields in the public announcements. In all cases, there is more detailed and highly specific documentation which is only available for registered bidders, hence could not be collected by opentender.eu. In our analysis we analyse the collected texts following a careful pre-processing step, for more details see below.

While the dataset we use is extensive covering a long time period and very rich in detail, it suffers from a number of quality and scope limitations. First, as the source information follows standard publication formats which frequently change, the database is not always consistent over time (e.g. some variables may be missing for some publication types) and some errors might remain due to inconsistent source information. Second, the dataset only contains contracts above the mandatory publication thresholds of about 50,000 EUR where regulatory requirements, including transparency norms, are more stringent (for more information see: http://europam.eu/?module=country-profile&country=Hungary#info_PP). Moreover, there are other exceptions from publication requirements in the Hungarian Public Procurement Law, for example sectors such as high value defence spending are typically exempted hence not contained in our database. Third, there might be missing texts due to some public buyers not fully following legal requirements or exploiting loopholes for avoiding public scrutiny.

Structured variables: Red flags and control variables

Two groups of variables are derived from structured data: corruption risk indicators (aka red flags) and control variables. Corruption risk indicators approximate corruption occurring in individual public procurement tenders and contracts. Importantly, they should not be interpreted as indications of whether corruption has actually occurred.

Following Fazekas et al (2016), we use the single bidding indicator, that is one bid submitted on an otherwise competitive market, as the outcome variable for our models. Single bidding on competitive markets is the simplest indicator of competition restriction, hence a core risk factor of corruption. It has been shown across a wide set of countries, including Hungary, that single bidding is associated with overpricing at the bidding stage, other risk factors such as tax haven registration of the winning supplier (*Fazekas and Kocsis, 2020*), and also perceptions and self reported experiences with corruption (*Charron et al, 2017*).

Moreover, we consider further risk factors which indicate behaviors or situations often leading to deliberately restricted competition in public procurement (*Fazekas and Kocsis, 2020*). These further red flags point at process biases which can be used and in fact often are used to exclude unwanted, unconnected bidders. Of the longer list of such indicators (*Fazekas et al, 2016*), we only adopt those indicators for the subsequent analysis which are applicable to Hungary, while also potentially can be calculated for a wider set of countries. In addition, we also excluded red flags which overlap with text-based indicators, for which we develop a wide set of new indices. Our red flag list, used as predictors of single bidding hence is:

- **Non-open procedure type:** using procedure types which exclude bidders by definition such as direct awards or negotiated procedure without prior publication represent a straightforward way for favouring a connected bidder. Moreover, those procedure types which retain some, but only partially requirements of open competition such as invitation tenders, can also be abused for favoritistic and corrupt ends.

- **No call for tenders published**: When the call for tenders are not published in the official journal, it is much harder for interested bidders to learn about tendering opportunities. Hence, informing the connected bidder about a tender while avoiding a public announcement can restrict competition and disadvantage non-connected bidders.

- **Suspiciously short submission period length**: When an open tender has to be run following transparency requirements, defining an unusually short submission or advertisement period (i.e. the number of days between publishing the call for tenders and the bid submission deadline) can put non-connected bidders at a disadvantage. This happens when bidders who only learn about the tender from the public announcement have too little time to put together high quality, competitive bids, compared to a connected bidder who received the information about the tender earlier.

- **Suspiciously short decision period length**: When the decision period is very short (i.e. the number of days between bid submission deadline and contract award decision date) it can indicate that the buyer did not consider bidders carefully, rather made snap decisions in favour of the connected bidder.

In addition to red flags, our models include a range of control variables which take account of different degrees of complexity, product market specificities, and somewhat different regulations (e.g. state-owned enterprises face different procedure type thresholds than central government ministries). These

variables are contract value, economic sector (2-digit Common Procurement Vocabulary (CPV) codes), year (based on contract award announcement), buyer location (region of the buyer), buyer type (e.g. central government entity), and buyer main activity (e.g. education or healthcare).

Text-based indicators and text processing

We added text-based indicators to the structured variables described above, inspired by the small but fast growing, diverse literature (*Winters, 2014; Fazekas and Kocsis, 2020; Gorgun et al, 2022*). This included creating i) n-grams, that is combinations of 3, 4, and 5 words as predictors; and ii) calculating meta-characteristics of the text fields such as length or uniqueness. We calculated these variables for each of the 3 different text field types introduced above: product description, eligibility criteria, and award criteria.

In order to calculate these variables, the first step is to clean raw text data. We conducted careful preprocessing to standardize texts and to remove noise as much as possible. We carried out preprocessing in multiple steps, trying out different methods and degrees of preprocessing as they can greatly influence modelling outcomes and accuracy (*Denny and Spirling 2018*). Python was used to perform preprocessing. We used the Spacy and NLTK packages, which are available in Hungarian.

Our first step of preprocessing was lemmatization (i.e. finding the dictionary form or root of words). For Hungarian (being an agglutinating language), the otherwise faster stemming does not work. The difference between the two is that the former looks for the actual lemma, while the latter only cuts off the suffixes from the end of the word. We used the HuSpaCy[1] package for this task. Second, lemmatization was followed by stop-word removal. We used the stop-word list available for Hungarian from the NLTK package. We removed a range of frequently occurring stop-words such as the Hungarian equivalents of “a”, “the”, “and”, etc. As part of stop-word removal, we also removed numbers. While numbers may be important - as for example they can indicate laws and sections of laws or specific eligibility parameters - removing numbers improves model interpretability. This is because competition restriction happens through overspecifying legal and technical conditions. However, it is not central to our claims whether restrictions happen through any particular paragraph or eligibility condition, rather their frequency of use.[2] Moreover, we also decided to eliminate words shorter than 2 characters, in addition to the removal of stop words. Looking at the word frequency list and the results of the models, such a strict pre-processing delivered the best balance between interpretability and model accuracy.

After careful text pre-processing, we created predictors from n-grams by converting the processed text data to numerical vectors – in the form of a sparse matrix – suitable for analysis. We used the TF-IDF (Term Frequency-Inverse Document Frequency) vectorizer from the scikit-learn[3] package in Python. TF-IDF reflects the importance of a term in a document relative to a collection of documents.

We also calculated text meta characteristics, albeit for these variables, we did not use the same pre-processing, rather calculated them on the raw, unprocessed texts. These meta variables aim at capturing the macro features of each text either on its own (e.g. law reference) or compared to other texts (e.g. uniqueness score or lexical diversity). Calculating the ratio of numbers and tracking

[1] https://github.com/huspacy/huspacy

[2] Underpinning these claims, our overall model accuracy did not meaningfully improve by retaining numbers through pre-processing. By implication retaining numbers would have increased model complexity at little to no model accuracy gain.

[3] https://scikit-learn.org/stable/

references to laws allows us to retain some of the information removed during pre-processing, i.e. removing numbers. Moreover, including the length of texts (normalized by economic sector average) allows the models to consider the overall frequency of words in a text in addition to the occurrence of specific words and word combinations.

Final dataset

The final, cleaned dataset used in the analysis contains a little over 119,000 contracts from 2011-2020. The number of contracts follows a cyclical distribution across years (**Figure 1**.)

Figure 1. Number of contracts awarded by year, Hungary, 2011-2020

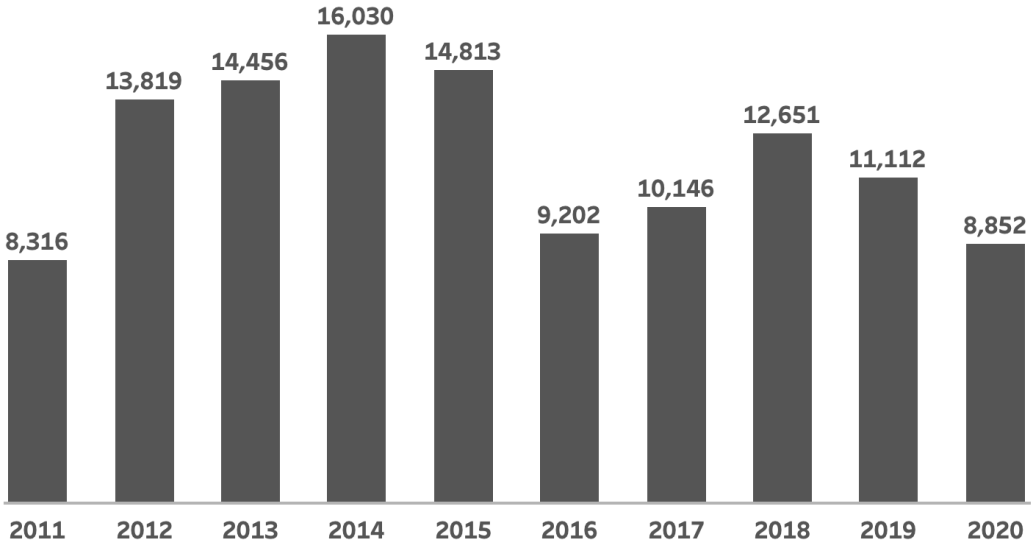

Table 2 lists the variables used in the analysis along with their definitions and descriptive statistics. First, we show the outcome variable used in the analysis, single bidding, which has 24% of awarded contracts receiving only 1 bid and 76% receiving 2 or more bids. Second, we describe the control variables used: contract value, product sector (2-digit CPV code), tender year, buyer location, buyer type (e.g. central government bodies), and buyer main activities. Third, we present structured red flags of corruption used in prior research. Our models incorporate widely used corruption risk indicators such as submission period length, no call for tender publication, non-open procedure type, and decision period length. No call for tenders is the most widespread risk factor in our dataset. Third, we describe the 3 main text field types in public procurement announcements, highlighting that product descriptions are the most widely available text field. Finally, we show the text-based meta variables such as uniqueness, lexical diversity, the mention of numbers and laws.

Table 2. Variables used in predictive models, variable definitions, and distributions, Ncontracts=119,397

Variable	Definition	Distribution	
Outcome variable			
Single bidding	0 = more than 1 bid received 1 = 1 bid received	0 1	91046 28351
Control variables			
Bid price	Contract values coded into deciles are used instead of actual contract values because the contract value distribution is highly skewed with a few large contracts distorting results.	Mean (bid price) Missing	413113. 95 12207
Tender CPV divisions	CPV = Common procurement vocabulary CPV stands for the market division of the tender Top 20 CPV + other	top 5 largest categories 45 33 Other 79 71 Missing	 49312 9823 9030 8615 6870 214
Tender year	Year of contract (2011-2020)	2011 2012 2013 2014 2015 2016 2017 2018 2019 2020	8316 13819 14456 16030 14813 9202 10146 12651 11112 8852
Buyer NUTS	NUTS = Nomenclature of territorial units for statistics. Hungary's geographical regions: West, Central and Eastern Hungary plus the whole country for national markets	HU1 HU3 HU2 HU Missing	45702 40079 27863 5711 42
Buyer type	Main type of the buying organisation as defined by EU's Procurement Directive	Regional agency Regional authority Public body National authority Other Missing	50374 3732 2997 2030 58263 2001

Buyer main activities	Main sector of the buying organisation	Other Health General Public Services Urban Transport Education Recreation Culture & Rel. Economic & Fin.Affairs Environment Water Social Protection Railway Public Order and Safety Defence Gas & Heat Production Postal Electricity Missing	107075 12845 1940 1047 636 601 278 221 175 109 107 71 60 56 42 40 986
Red Flags based on structured data			
Submission period length	number of days between publication of call for tenders and submission deadline 0 = long submission period (no risk) (>= 38 days) 1 = short submission period (risky) (2 to 37 days)	0 1 Missing	1164 16381 101852
No call for tender publication	0 = call for tenders published on publication portal (no risk) 1 = call for tenders not published on publication portal (no notice URL) (risky)	0 1	17588 101809
Decision period length	number of days between submission deadline and announcing the contract award 0 = long decision period (no risk) (>= 41 days) 0.5 = moderately short decision period (medium risk) (21 to 40 days) 1 = very short decision period or missing decision period (high corruption risk) (<= 20 days)	0 0.5 1	9809 5158 104430
Procedure Type	0 = open (no risk) 1 = non-open (high risk)	0 1 Missing	52915 65881 601
Unstructured text fields used		**Missing ratios**	

Tendering requirements (Eligibility criteria)	Eligibility criteria define which companies are allowed to bid, what prior experience is required from eligible bidders. **Personal** requirements (e.g. compliance with laws) **Technical** requirements (e.g. different qualifications, professional knowledge) **Economic** requirements (e.g. financial resources)	Personal requirements Technical requirements Economic requirements	85% 85% 85%
Product description	Tender and/or lot **title**: subject of the contract Tender **description**: short description of the subject of the contract	Title Description	12% 12%
Award criteria	After the bids are received, the procuring body evaluates them and selects the winner according to the award criteria. The result is published in a contract award announcement. If no valid bids were received or the prices were too high for the institution, an announcement is published about the reason of the failure	Price only ratio Price and quality ratio Missing award criteria information	41% 43% 16%
Text-based meta variables			
Uniqueness score	This variable provides a measure of how unique or distinctive the language is in each document relative to the entire collection. It is calculated for each text type and aggregated on tender level. Higher scores indicate that the document contains words or phrases that are relatively unique within the dataset. Theoretical minimum: 0 Theoretical maximum: 1	mean	0.00014
Normalized length	The variable represents the normalized length of each document's combined text content, considering the total length of text in various columns and normalizing it with respect to the mean length of documents within the same CPV category.	min mean max	0 1 41
Law reference	0 = text does not contain a reference to any laws 1 = text does contain a reference to a law	Title 0 1 Description 0 1 Personal requirements 0 1 Technical requirements	 108589 10808 71654 47743 102104 17293

		0 1 Economic requirements 0 1 Award criteria 0 1	102623 16774 106185 13212 108856 10541
Number ratio	Number ratio shows the ratio of the count of numbers in each text to the sum of counts of numbers for the corresponding CPV sector. Theoretical minimum: 0 Theoretical maximum: 1	Mean of the variable: Title Description Personal requirements Technical requirements Economic requirements Award criteria	 0.000193 0.000193 0.000193 0.000193 0.000193 0.000192
Lexical diversity	Lexical diversity is defined here as the ratio of the number of unique words to the total number of words in a text (considering CPV category). A high value means higher lexical diversity score, which indicates that the text has a greater variety of unique words relative to the total number of words, suggesting a richer and more varied use of language. Theoretical minimum: 0 Theoretical maximum: 1	Mean of the variable: Title Description Personal requirements Technical requirements Economic requirements Award criteria	 0.98 0.79 0.94 0.94 0.96 0.96

Methods: Predictive Models

In order to test our hypotheses we build on a small, but quickly growing literature making use of texts and specific terms in public procurement and more broadly project finance (*e.g. Winters, 2014; Modrusan et al, 2020*). Specifically, we estimate supervised learning models which predict single bidding (1=one bid submitted; 0=more than 1 bid submitted) with the help of structured data as well as unstructured textual information. We estimate a wide range of models with the goal of identifying the most accurate models including text-as-data on top of already tried and used structured information. When deciding which kinds of predictive models to estimate, we aimed to balance interpretability with predictive power. Hence, we opted for estimating i) binary logistic regression, ii) random forest, and iii) XGBoost models (*Gareth et al, 2021*).[4] Specifically, we used the built-in logistic regression and random forest models from the scikit-learn package in Python. We divided our data into two parts: 80% of the observations belong to the train and 20% to the test set, we used the default settings of the models.

Logistic regression is a statistical method used for binary classification problems, where the outcome variable is categorical with two possible values. The model estimates coefficients for the input features to make predictions, which makes the interpretation of the model easy and understandable. Random Forests and XGBoost are ensemble models built on multiple decision trees. Ensemble techniques let us train more accurate and more stable predictive models, by combining the output from a large number of individual models (in this case each individual decision tree), each of which have been estimated on a slightly different sample with somewhat different parameters. In Random Forest models each tree is built independently, and the final prediction is a majority vote of the individual tree predictions. With a higher number of trees we can achieve better performance but running a model with a high number of trees is very resource intensive (i.e. takes long to run on ordinary machines). An advantage of Random Forest models is, that they handle imbalanced datasets more effectively than regression models. In XGBoost models, trees are built sequentially, and each new tree corrects the errors of the combined ensemble of the previous trees.

To choose the most accurate predictive model, we compare the observed single bidding outcome with the predicted value on a dataset 'not seen' by the model (test set). We calculate 4 different goodness-of-fit metrics:

- **Precision**: Correct positive predictions relative to total positive predictions
  Precision = True Positive / (True Positive + False Positive)
- **Recall**: Correct positive predictions relative to total actual positives
  Recall = True Positive / (True Positive + False Negative)
- **F1-score**: Harmonic mean of precision and recall
  F1-score = 2 * (Precision * Recall) / (Precision + Recall)
- **Accuracy**: Percentage of all correctly classified observations
  Accuracy = (True Positive + True Negative) / (Total Sample Size)

[4] We also run Support Vector Machines models, but they did not perform well compared to the other methods, so the results are not reported in this paper.

Results

Understanding the role of text fields

The results section starts by estimating baseline regressions including only structured data on control variables as well as red flags (**Table 2**). Then we test H1 by adding different text-based indicators and H2 by adding textual information all together. At each step, we show performance of logistic regression, random forest and XGBoost models.

The baseline models largely replicate Fazekas et al, 2016, but using longer time series and fewer red flags which are more readily applicable in a wide range of countries (*see for example, Fazekas and Kocsi, 2020*). The baseline models achieve moderate accuracy (ranging between 76% and 78%) and F1-score (68% – 76%) of outcomes correctly classified (**Table 3**). While these appear high, given that single bidding is observed for about 23% of contracts, a naive estimation classifying all contracts as non-single bidding would achieve over 77% accuracy. As it will be systematically shown, random forest models perform generally better than traditional logistic regression and XGBoost models.

Table 3. Baseline models, binary logistic regressions, Random Forests, XGBoosts using controls and red flags, Hungary, 2011-2020

	precision	recall	f1-score	accuracy
Logistic Regression: control variables only	0.70	0.76	0.68	0.76
Random Forest: control variables only	0.74	0.77	0.74	0.77
Boosting: control variables only	0.74	0.77	0.69	0.77
Logistic Regression: control variables and red flags	0.71	0.77	0.69	0.77
Random Forest: control variables and red flags	**0.75**	**0.78**	**0.76**	**0.78**
Boosting: control variables and red flags	0.75	0.77	0.70	0.77

Note: the model closest to Fazekas et al, 2016 is highlighted

Now we turn to testing H1 by adding different text-based variables to the baseline models (**Table 4**). The textual information referring to different parts of the tendering documents - and hence different types of constraints on competition imposed - are added separately: eligibility criteria texts, product description texts, and award criteria texts. Each of these models outperform the baseline models, albeit typically not by far. The best Random Forest models using textual information achieve accuracy between 79% and 82% (f1-score of 0.77-0.80). Logistic regressions and XGBoost models are typically somewhat less accurate, but similarly above the baseline models. As these models are clearly above the baseline models which already include controls and validated red flags, we conclude that texts confer additional explanatory power for single bidding models in addition to structured information. This

suggests that constraining conditions and criteria in tender texts allow for limiting competition, that is increasing single bidding, when formalistic, visible procurement process biases are not present.

Table 4. Text based models, binary logistic regressions, Random Forests, and XGBoost using red flags, control variables, and texts, Hungary, 2011-2020

	precision	recall	f1-score	accuracy
PRODUCT DESCRIPTION				
Logistic Regression: texts, red flags and control variables	0.72	0.77	0.70	0.77
Random Forest: texts, red flags and control variables	**0.81**	**0.82**	**0.80**	**0.82**
Boosting: texts, red flags and control variables	0.79	0.77	0.69	0.77
ELIGIBILITY CRITERIA				
Logistic Regression: texts, red flags and control variables	0.72	0.77	0.69	0.77
Random Forest: texts, red flags and control variables	**0.77**	**0.79**	**0.77**	**0.79**
Boosting: texts, red flags and control variables	0.76	0.77	0.69	0.77
AWARD CRITERIA				
Logistic Regression: texts, red flags and control variables	0.71	0.76	0.69	0.76
Random Forest: texts, red flags and control variables	**0.78**	**0.80**	**0.78**	**0.80**
Boosting: texts, red flags and control variables	0.75	0.77	0.69	0.77

Note: Best model is highlighted in bold.

Now, we turn to testing H2 by adding the different text-based variables all at once and comparing model performance with previous models containing textual information (**Table 5**). The random forest model achieves the best prediction accuracy, correctly classifying 82% of contracts (f1-score=0.70). The improving model fit suggests that hard-to-identify constraints to competition in different parts of the tender documentation offer complementary avenues to achieving corrupt ends.

Table 5. Text based models, binary logistic regressions, and Random Forests using red flags, control variables, and all three text parts, Hungary, 2011-2020

	precision	recall	f1-score	accuracy
Logistic Regression: texts, red flags and control variables	0.72	0.77	0.70	0.77
Random Forest: texts, red flags and control variables	**0.81**	**0.82**	**0.80**	**0.82**
Boosting: texts, red flags and control variables	0.79	0.78	0.69	0.78

Note: Best model is highlighted in bold.

Next, we investigate H2 further by looking at the impact of text fields under 2 competitive scenarios: 1) when there is no call for tenders published - little to no open competition is expected, and 2) when a call for tenders is published - open competition is expected. We argue that texts can be used to justify the use of non-competitive procedure types (restrictive product description), but they can also bias the formally open competition (subjective award criteria). Hence, the role of product descriptions and award criteria may wary by call for tenders publication.

As a starting point for this analysis, we establish the different availability of different text fields in the 2 sub-samples: with/without call for tenders publication (**Table 6**). Clearly, eligibility criteria are only available in tenders with call for tenders publication, so they are not suitable for comparing across the 2 subsamples. As for product description and award criteria, missing rates vary somewhat, still allowing for comparisons across subsamples. Moreover, price-only award criteria are used approximately to the same degree in the 2 subsamples too, warranting meaningful comparisons.

Table 6. Share of missing and price only texts by text type and publication of call for tenders status (yes/no), Hungary, 2011-2020

Text type	Share of	Call for tenders published	No call for tenders published
product description	no text in title	0,08%	15%
product description	no text in description	0,1%	15%
eligibility criteria	no text in personal requirements	0,8%	99%
eligibility criteria	no text in economic requirements	0,2%	99%
eligibility criteria	no text in technical requirements	0,2%	99%
award criteria	no text in award criteria	0,9%	17%
award criteria	price only in award criteria	48%	40%

Turning to the impact of text fields on model performance in the 2 subsamples, we see marked differences (**Table 7**). Product descriptions improve model performance both with and without call for tenders, but the impact is larger in the no call published subsample with accuracy increasing from 78% to 82%. This suggests that specific or tailored product descriptions are more often used as a justification for non-competitive tenders rather than to bias open tenders. Still, the differences are not large, so the latter use is also somewhat prevalent in our data.

The inclusion of award criteria texts also improves model performance compared to the baseline without texts in the 2 subsamples (**Table 7**). Here, the improvement in model performance is somewhat larger for competitive tenders with call for tenders published: from 76% to 79%. This suggests, albeit only tentatively, that award criteria biases are more typically used when the competition is open in order to favour certain bidders. As a result, potential bidders stay away from the tender altogether.

Table 7. Random Forest models using red flags, control variables, and texts, for tenders with and without a published call, Hungary, 2011-2020

	Call published		No call published	
	f1-score	accuracy	f1-score	accuracy
NO TEXT: red flags and control variables only	0.75	0.76	0.75	0.78
PRODUCT DESCRIPTION: texts, red flags and controls	0.77	0.79	0.80	0.82
AWARD CRITERIA: texts, red flags and controls	0.78	0.79	0.78	0.80

Unpacking text-based models

While the overall performance of different models offer supporting evidence for H1 and H2, the models have remained a black box so far. In order to improve the interpretability of our results and link them back to our theoretical framework, we explore the role specific n-grams as well as full texts play in predicting single bidding. We hope to identify specific terms and conditions which are tailored to favoured firms by looking into our models.

First, we track the *individual impact of word n-grams in specific text fields* on the probability of single bidding. Specifically, we compare the terms with the largest positive and largest negative coefficients in the most complete logistic regression model. While Logistic regression models were inferior in terms of prediction accuracy, they offer straightforward coefficients which we can interpret and hence understand the relationships uncovered by our models more broadly (we follow in this approach (Rabuzin–Modrušan 2019). We highlight the most relevant patterns translated into English below, while delegating the detailed, Hungarian language tables to the Annex.

Regarding terms with highest positive and negative impact on single bidding probability in product descriptions, we can see a range of specific goods and services mentioned (Table A1). Both in the positive and negative impact groups, product specificity seems to be high; indicating that it is not how concrete the products are what matters, rather their specificity with regards to the product market context. In other words, some very specific products can be delivered by a wide set of suppliers while others are unique to one supplier. Without understanding these contexts, these results offer little for theory testing.

Turning to eligibility criteria texts, referring to personal, technical and economic requirements, notable differences emerge in line with theoretical expectations. While there is considerable variation, we find that n-grams of personal requirements decrease single bidding probability when they refer to general rules and guidance notes (e.g. “guidance note of the public procurement authority”) or when they allow

bidders to prove sufficient capacity using external resources such as subcontractors (e.g. “related to subcontractor eligibility certificate). Personal requirements terms increase single bidding, when they refer to exclusion conditions (e.g. “section subsection exclusion”) or when they require detailed data and conditions to be met (e.g. “specific detailed data”). When it comes to n-grams in technical requirements, we see a mixed picture. On the one hand, a number of specific products are mentioned which, similarly to product descriptions, may increase or decrease single bidding probability. This again, underlines the importance of market context. One notable example of this are the terms including “lighting installation” which increase single bidding probability. It is likely that they refer to the infamous Elios corruption case related to Viktor Orbán’s son-in-law: István Tiborcz[5]. On the other hand, when specific expertise is required, single bigging is more likely (e.g. “available expert higher-education degree electricity”). Conversely, when general skills or expertise are required single bidding probability goes down (e.g. “experienced project manager expert”). When it comes to economic requirements, an already familiar pattern emerges. Terms allowing bidders to prove sufficient capacity and guarantees using external resources are associated with lower single bidding probability (e.g. “organisation enables capacity for case”). Moreover, lighter bureaucracy, such as allowing for a self-declaration instead of an official certificate, is also associated with lower single bidding probability (this also showed up as lowering single bidding probability in the other eligibility criteria fields). However, when n-grams relate to minimum requirements, especially those related to financial performance such as turnover, we see a higher predicted single bidding probability (e.g. “offer income in bid”). Once again, specific products show up among the most impactful predictors, but their correct interpretation requires a more comprehensive understanding of the market environments.

Considering award criteria, terms considering guarantee periods and warranty clauses tend to lower the probability of single bidding (e.g. “guarantee period month”), which suggests that a longer-term, quality-oriented perspective of public investments tend to decrease corruption risks. Similarly, award criteria n-grams referring to total prices decrease single bidding probability in our models (e.g. “single amount net offer”), confirming prior research using key-word based, and hence simpler text-mining methods (Fazekas et al, 2016; Fazekas and Kocsis, 2020). When award criteria include delivery timeliness-related terms, single bidding probability increases which points at the corruption enhancing effect of emergencies, expedited, and urgent procurement (Schultz and Soreide, 2008).

Second, we look at the *full, raw texts of description, eligibility criteria and award criteria fields* in tenders where the best model without texts incorrectly predicts no single bidding, while the best model with texts correctly predicts single bidding[6]. Such tenders and contracts should capture those cases when the inclusion of textual information crucially contributes to a more precise identification of competition restrictions, in spite of formally open procedures.

While there is a considerable amount of noise, especially a large volume of texts which are short and vague (recall we only work with official announcement texts, while full technical details are in separate tender documentations we do not have access to), we can confirm theoretically sound tendencies already identified using n-grams. The frequent, lengthy and complicated exclusion criteria appear to co-occur with single bidding in spite of formally open procedural features (e.g. there are personal requirement fields which are 2-3 standard deviation above the mean length of such fields and contain a bewildering array of legal references to certificates and proofs to be submitted). Moreover, criteria and requirement specificity also comes up as associated with correctly predicting the incidence of single bidding. Among eligibility criteria, some surprisingly specific conditions are associated with single

[5] For the detailed investigative report of OLAF see: https://tasz.hu/wp-content/uploads/2024/01/Final_Report_OCM201726804_redacted.pdf and the broader context of the case see: https://atlatszo.hu/kozpenz/2022/02/04/vegre-nyilvanos-az-elios-ugyrol-szolo-olaf-jelentes-bar-tiborcz-istvan-es-az-elios-nevet-kitakartak-benne/

[6] For tractability, our review concentrates on contracts with full tender text information in all text fields from major buyers.

bidding in spite of no apparent, structural red flag for competition limitation: "*Has reference(s) with CPV code 33111720-4 for angiographic equipment within the last 3 years from the date of dispatch of the invitation to tender (see point VI.5 of this notice), with the following quantities of a reference delivered in accordance with the specifications and the contract: 19 for the 1st part, 193 for the 2nd part, 193 for the 3rd part 23 for the 4th part […] 28 for 20th part, 19 for the 21st part, 19 for the 22nd part, 8 for 23rd part. [etc.]*". It is hard to fathom why so specific numbers of delivery references are required instead of a general minimum amount or interval. Finally, when it comes to award criteria texts, the overwhelming majority of tenders award contracts based on price with some further tenders also considering objective, numerical award criteria such as delivery speed, guarantee length, or damages payments. In a few cases, arguably subjective criteria co-occurs with single bidding even in the absence of formal, visible competition restrictions. For example, a contract award was partially based on the "quality of the organisational plan", coinciding with earlier examples cited in Fazekas and Kocsis (2020).

Conclusions and further work

We have built a comprehensive database of public procurement tenders and contracts for Hungary, spanning over a decade worth of public contracts, accounting for roughly one third of public spending. The analysis identified subtle, text-based strategies for limiting competition by tailoring the bidding conditions to a potentially favoured company. This can include specifying the purchase of a unique product, the excessive use of exclusion conditions, putting high weight on idiosyncratic, specific experience, and the requirement of unreasonably extensive prior experience (e.g. past turnover). We found evidence that these corruption strategies are mostly employed by corrupt actors when more visible strategies for favouritism, such as non-competitive tendering procedures, are undesirable or impractical. The inclusion of text based information improves overall model prediction accuracy from 77% to 82%, with each text field type additionally contributing to model performance. Nevertheless, text-based strategies can also support visible corruption strategies, such as the use of non-competitive procedure types justified by the requirement for purchasing a highly specific product.

While we gathered supporting evidence for the importance and strategic use of textual information for furthering corruption and hence we contributed to this small but growing literature, our analysis is limited in a number of ways. Most importantly, our dataset has contained a high rate of missing data which is most likely due to lots of texts being delegated to full tender documents rather than the official tender announcements (recall, we collected data from the latter but could not access the former due to its irregular structure and varying formats such as scanned pdfs).

Further work along the lines of this paper should improve model prediction accuracy, for example by drilling deeper into sectoral differences. If the approach turns out to be fruitful and valuable, it should be extended to other countries using different languages in public procurement (following for example, Gorgun et al (2020)). The long-term ambition is to extend the regular toolkit of corruption red flagging by researchers and policy actors using text-as-data.

Further work could be extended to predict other proxies for corruption and limited competition, such as spending and market concentration or contract award to a politically connected firm. These new dependent variables could in particular unpack the complex dynamics in multiple bid tenders where favouritism is at play and hence, among others, award criteria is applied in a biased manner to favour the connected bidder over the other bidders in the tender.

References

Acikalin, U. U.; Gorgun, M. K.; Kutlu, M.; & Tas, B. K. O. (2023). *How you describe procurement calls matters: Predicting outcome of public procurement using call descriptions. Natural Language Engineering, 1–22.* doi:10.1017/S135132492300030X

Adam, Isabelle; Dávid Barrett, Elizabeth; and Fazekas, Mihály (2020) *Modelling Reform Strategies for Open Contracting in Low and Middle Income Countries.* Transparency International, London, UK.

Auriol, E., Straub, S., and Flochel, T. (2016). *'Public Procurement and Rent-seeking: The Case of Paraguay'*, World Development, 77: 395–407.

Bandiera, Oriana, Andrea Prat, and Tommaso Valletti. (2009) "*Active and Passive Waste in Government Spending: Evidence from a Policy Experiment*." American Economic Review 99(4): 1278–1308.

Nicholas Charron, Carl Dahlström, Mihály Fazekas, and Victor Lapuente, (2017), *Careers, Connections and Corruption Risks In Europe.* Journal of Politics, 79(1): 89-104.

Chong, Eshien, Michael Klien, and Stéphane Saussier. (2015) *The Quality of Governance and the Use of Negotiated Procurement Procedures: Evidence from the European Union*. Paris.

Coviello, Decio, and Stefano Gagliarducci. (2017) "*Tenure in Office and Public Procurement.*" American Economic Journal: Economic Policy, 9 (3): 59-105.

Decarolis, F., and Giorgiantonio, C. (2022) *Corruption red flags in public procurement: new evidence from Italian calls for tenders*. EPJ Data Sci. 11, 16 https://doi.org/10.1140/epjds/s13688-022-00325-x

Denny, M., and Spirling, A. (2018). *Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It*. Political Analysis, 26(2), 168-189. doi:10.1017/pan.2017.44

Di Tella, R., and Schargrodsky, E. (2003). '*The Role of Wages and Auditing During a Crackdown on Corruption in the City of Buenos Aires'*, Journal of Law and Economics, 46(1): 269–92.

Fazekas, Mihály, and Kocsis, Gábor, (2020), *Uncovering High-Level Corruption: Cross-National Corruption Proxies Using Public Procurement Data*. British Journal of Political Science, 50(1).

Fazekas, Mihály, Luciana Cingolani, & Bence Tóth (2018), *Innovations in Objectively Measuring Corruption in Public Procurement*. In Helmut K. Anheier, Matthias Haber, and Mark A. Kayser (eds.) Governance Indicators. Approaches, Progress, Promise. Ch. 7. Oxford University Press, Oxford.

Fazekas, Mihály & Stéphane Saussier (2018), *Big Data in Public Procurement*. Colloquium, in Gustavo Piga & Tünde Tátrai (Eds.) Law and Economics of Public Procurement Reform. ch. 3. Routledge, London.

Fazekas, Mihály, Tóth, István János, and King, Peter Lawrence, (2016), *An Objective Corruption Risk Index Using Public Procurement Data*. European Journal of Criminal Policy and Research, 22.

Fazekas, Mihály, Tóth, István János, and King, Peter Lawrence, (2013), *Corruption manual for beginners: 'Corruption techniques' in public procurement with examples from Hungary*. GTI-WP/2013:01, Budapest: Government Transparency Institute.

Ferwerda, J., Deleanu, I., and Unger, B. (2017). *'Corruption in Public Procurement : Finding the Right Indicators'*, European Journal on Criminal Policy and Research, 23(2): 245–67.

Gnaldi, Michaela, Del Sarto, S., Falcone, M., Troìa, M. (2021). *Measuring Corruption*. In: Carloni, E., Gnaldi, M. (eds) Understanding and Fighting Corruption in Europe. Springer, Cham. chapter 4, pp. 43-71

Gorgun, Mustafa Kaan; Kutlu, Mucahid; Taş, Bedri Kamil Onur (2020) *Predicting The Number of Bidders in Public Procurement*, 2020 5th International Conference on Computer Science and Engineering (UBMK), 2020, pp. 360-365, doi: 10.1109/UBMK50275.2020.9219404.

Hyytinen, A., Lundberg, S., and Toivanen, O. (2008). *Politics and Procurement. Evidence from Cleaning Contracts.* Discussion Paper No. 233. Helsinki: Helsinki Center of Economic Research.

James, Gareth; Witten, Daniela; Hastie, Trevor; and Tibshirani, Robert (2021) *An Introduction to Statistical Learning: With Applications in R. 2nd edition*, Springer, London.

Klašnja, M. (2015). '*Corruption and the Incumbency Disadvantage: Theory and Evidence'*, Journal of Politics, 77(4): 928–42.

Lisciandra, M., Milani, R., and Millemaci, E. (2022) *A Corruption Risk indicator for Public Procurement*. European Journal of Political Economy, Vol. 73, No. 102141, 2022, Available at SSRN: https://ssrn.com/abstract=4045108

Lyra, M.S., Damásio, B., Pinheiro, F.L., and Bacao, F. (2022) *Fraud, corruption, and collusion in public procurement activities, a systematic literature review on data-driven methods*. Applid Network Science, 7, 83.

Modrušan, N.; Rabuzin, K.; Mrsic, L. (2020) *Improving public sector efficiency using advanced text mining in the procurement process*. 9th International Conference on Data Science, Technology and Applications pp. 200–206 doi:10.5220/0009823102000206.

North, D. C., Wallis, J. J., and Weingast, B. R. (2009). *Violence and Social Orders: A Conceptual Framework for Interpreting Recorded Human History*. Cambridge: Cambridge University Press.

OECD. 2007. Integrity in Public Procurement. Good Practice from A to Z. Paris: OECD.

Olken, B. A. (2007). '*Monitoring Corruption: Evidence from a Field Experiment in Indonesia'*, Journal of Political Economy, 115(2): 200–49.

Rabuzin, Kornelije; Modrušan, Nikola (2019) *Prediction of Public Procurement Corruption Indices using Machine Learning Methods.* KMIS.

Rothstein, Bo, & Teorell, Jan (2008). *What is quality of government? A theory of impartial government institutions.* Governance, 21(2), 165–190.

Schultz, Jessica & Soreide, Tina (2008). *Corruption in emergency procurement*. Disasters, 32(4), 516-536.

Villamil, Isabella; Janos Kertész, and Johannes Wachs (2023) *Computational Approaches to the Study of Corruption.* in Taha Yasseri (ed.) 2023, Handbook of Computational Social Science, Edward Elgar Publishing.

Winters, Matthew S. (2014) *Targeting, Accountability and Capture in Development Projects, International Studies Quarterly*, 58(2), Pages 393–404

Yukins, Christopher (2007). *Integrating Integrity and Procurement: The United Nations Convention Against Corruption and the Uncitral Model Procurement Law*. Public Contract Law Journal, 36(3), 307–329.

Annex A. Detailed tables

Table A1. Terms in the description field with highest and lowest coefficients (log odds), logistic regression model using red flags, control variables, and all texts, Hungary, 2011-2020

Decreasing single bidding probability		**Increasing single bidding probability**	
Term	**Coeff.**	**Term**	**Coeff.**
bővítés meglévő miskolc mechatronikai park	-0.15	családi öltöző építés gyalul faanyag	0.04
bontás törmelék leőrlés ágyazati anyag	-0.08	bontás kif szabadvezetékes	0.04
ajánlatkérő támasztott további műszaki	-0.06	bonta homlokzati panél	0.04
berendezés bérelt tulajdon	-0.05	bonta gipszkarton mennyezet bonta dogoza	0.04
bevezetés megelőző rendszer megfelelőségi nyilatkozat	-0.04	betárolás kerül sor tényleges mennyiség	0.04
belső tér nem	-0.04	betárolás kerül sor tényleges	0.04
belsőterű helyiség terápiás	-0.04	audio outpu hdmi buil speaker	0.04
ajánlattevő ütem megvalósít nyertes ajánlattevő	-0.04	beszerzés alábbi rész egyszerhasználatos fecskendő	0.04
ajánlatkérő pályázati tevékenység érintett intézmény	-0.04	audio outpu hdmi buil	0.04
altemplom elhelyezkedő két	-0.04	beszerzés alábbi rész dró	0.04
bővítés átalányáras kivitelezési szerződés jelű	-0.04	audio out xlr timecode and	0.04
azonosító ajánlatkérő rendelet vidéki	-0.04	beszerzés alábbi rész darab szállítás	0.04
azonosító ajánlatkérő önkormányzati feladat	-0.04	betárolás kapcsolatos feladat	0.04
ajánlatkérő rész hrsz partfal	-0.04	boumaz abdelkrim mintagazdasági terület öntözésfejlesztés	0.04
ajánlatkérő rész hrsz partfal helyreállítás	-0.04	csatlakozás buszváró felépítmény építés balatonmagyaród	0.04
bontás törmelék elszállítás hulladéklerakó elhelyezés	-0.04	ajánlattevő figyelem mindkét	0.04
alfaterv kft generáltervező	-0.03	beléptető rendszer alapcsomag kártyaolvasó	0.04
adattáblázat zárótanulmány módszertani	-0.03	beépítendő föld építhető gumi zárású	0.04
ablak méret minőségi	-0.03	beépítendő földanyag biztosítás	0.04
csap stb alapgépészet	-0.03	bemenet továbbfűzési lehetőség átkapcsolható mikrofonbemenet	0.04

Table A2. Terms in the personal requirements field with highest and lowest coefficients (log odds), logistic regression model using red flags, control variables, and all texts, Hungary, 2011-2020

Decreasing single bidding probability		**Increasing single bidding probability**	
Term	**Coeff.**	**Term**	**Coeff.**
leírt ellenőriz ajánlat tartalmaznia	-0.06	nem elválasztható rész bontás	0.04
pont alap kizár eljárás	-0.04	meghatározott legkedvezőtlen érték ami	0.04
közb dokumentum foglalt nyilatkozat kizáró	-0.04	pont alpont hatály	0.04
közbeszerzési hatóság kiadott közbeszerzési értesítő	-0.04	nem ellenőrizhető szükséges	0.03
közbeszerzési hatóság kiadott útmutató	**-0.04**	**pont alpont foglalt kizáró**	0.03
engedély meglét pénzügyi szervezet	-0.03	**nem régebbi lesz ajánlatkérő kizár**	0.02
ajánlatkérő további hivatkozik kbt bekezdés	-0.03	ajánlatkérő változásbejegyzési eljárás kapcsolat	0.02
száj című útmutató közbeszerzési	**-0.03**	**nem régebbi lesz ajánlattevő**	0.02
ajánlatkérő részvétel jelentkezés érvényesség	-0.03	nem elég tesz	0.02
ajánlatkérő ajánlatkérő megajánlott szakember	-0.03	meghatározott megfelelő eljárás nem	0.02
pont kapcsán alvállalkozó alkalmasság igazolás	-0.03	**meghatározott megadott részletes adat**	0.02
pont ajánlattevő cégszerű nyilatkozik	-0.02	**ajánlatkérő részszempont eset pontozás módszer**	0.02
hiány részvétel jelentkező kizáró	-0.02	nem elválasztható rész bontás műszaki	0.02
nek korm rend rendelkezés	-0.02	meghatározott magyarország letelepedett	0.02
következik alap ajánlattevő köteles ajánlat	-0.02	meghatározott legkedvezőtlen érték ami minimális	0.02
nek benyújtás eekd rész	-0.02	**ajánlatkérő részszempont eset ajánlatkérő kedvező**	0.02
nek vmint rendelkezés igazol	-0.02	**ajánlatkérő részszempont eset ajánlatkérő**	0.02
nek zár eljárás airsz rész	-0.02	európai közbeszerzési dokumentum minta ekr	0.02
pont alap kizár eljárás ajánlattevő	-0.02	európai közbeszerzési dokumentum mely elegendő	0.02
rendelkezik adott rész rész kapcsolódó	-0.02	ajánlattevő ajánlattételi felhívás nyilatkozik	0.02

Table A3. Terms in the technical requirements field with highest and lowest coefficients (log odds), logistic regression model using red flags, control variables, and all texts, Hungary, 2011-2020

Decreasing single bidding probability		**Increasing single bidding probability**	
Term	**Coeff.**	**Term**	**Coeff.**
szolgáltatás kapcsolatos üzleti média	-0.15	**tapasztalat megfelelőség szakember nyilatkozat ajánlattevő**	0.05
rendszám forgalmi engedély műszaki	-0.08	**rendelkező szakértő felsőfokú gyengeáramú**	0.04
bevon kíván építész építőmérnök	**-0.06**	**rendelkezik szervezet kapacitás**	0.04
megkezdett építés vesz figyelem fenntartott	-0.05	rendelkezik szerkezetkész állapot kezdődő használatbavételi	0.04
nem mutatható rész történő	-0.04	natura hatásbecslés elkészítés szakember	0.04
magasépítési épület felújítási referencia részajánlat	-0.04	natura hatásbecslés elkészítés	0.04
megad csatolt szakmai	-0.04	kbt bekezdés korm rendelet vonatkozó	0.04
cél csatol köteles eljárás megindító	-0.04	**minimum lámpatest beépítés**	0.04
beruházás ismertetés alkalmassági	-0.04	kbt bekezdés korm rendelet további	0.04
gépjármű alkalmas hűtés igénylő	-0.04	minimum lábazati hőszigetelés	0.04
szolgáló lépcsőnjáró személyemelő	-0.04	kbt bekezdés korm rendelet meghatározott	0.04
kizárólag önálló értelmezhető alkalmassági követelmény	-0.04	minimum literes rakodótér rendelkező üzemű	0.04
kizárólag üzemképes mkeh hiteles regiszterbizonylat	-0.04	natrium klorid tartalom min nedvességtartalom	0.04
beruházás szóló nyilatkozat szerződés	-0.04	rész tervezési tervezői művezetési szolgáltatás	0.04
beruházás szóló nyilatkozat szerződés kötő	-0.04	tekintet mérnöki tervező modellező szoftver	0.04
rendszerű talajvíz kitermelő nap	-0.04	bőrgyógyászat kardiológia reumatológia	0.04
felül felsőfokú végzettség európai unió	-0.03	megfelel azaz amennyiben	0.04
alkalmassági előírás nak	-0.03	nettó értékű lökéshullám	0.04
adott szakember ellátott feladat tekintet	-0.03	nettó értékű lökéshullám terápiás	0.04
tapasztalat rendelkezik projektvezető szakember	-0.03	megfelelés vhr bekezdés pont alap	0.04

Table A4. Terms in the personal economic field with highest and lowest coefficients (log odds), logistic regression model using red flags, control variables, and all texts, Hungary, 2011-2020

Decreasing single bidding probability		**Increasing single bidding probability**	
Term	**Coeff.**	**Term**	**Coeff.**
gazdasági szereplő egységes európai közbeszerzési	-0.06	**hónap vezet valamennyi pénzforgalmi**	0.04
illetőleg felmérés engedélyezési	-0.04	helyett alkalmasság minimumkövetelmény közbeszerzési eljárás	0.04
fordul elő műszaki szakmai	-0.04	állítás kivitelezés tevékenység	0.04
szervezet eset kapacitás rendelkezés bocsátó	**-0.04**	informatikai alkatrész karbantartási anyag szállítás	0.04
forint alkalmatlan ajánlattevő előző	-0.04	**hónap valamennyi pénzforgalmi**	0.03
forint alkalmatlan ajánlattevő előző három	-0.04	informatika terület oktatási képzési szolgáltatás	0.03
nyomatató értékesítés vonatkozás	-0.03	idő közbeszerzés tárgya kegyeleti	0.02
benyújtás helyett előző	**-0.03**	pont ajánlattevő jogelőd két	0.02
ajánlat tevő alábbi dokumentum	-0.03	**ajánlat árbevétel ajánlattétel**	0.02
vezet adott szám nap	-0.03	idő közbeszerzés tárgya kertészeti	0.02
korm bekezdés pont kapcsolat előírt	-0.03	hónap vezet valamennyi élő	0.02
ajánlat tartalmaznia rész szóló	-0.03	tárgy szerinti speciális nyomdai kivitelezési	0.02
vezet adott számla számla számla	-0.03	helyett alkalmasság minimumkövetelmény nem rendelkezik	0.02
mft alkalmatlan ajánlattevő	-0.03	helyett **alkalmasság minimumkövetelmény** mérleg	0.02
mft biztosítási időszak kártérítési limitet	-0.03	ajánlat tartalmaznia mind ajánlattevő mind	0.02
aelv rossz ajánlatkérő	-0.03	tárgy szerinti rádióberendezés	0.02
intézmény rész gépműszer	-0.03	hónap vezet valamennyi pénzforgalmi számla	0.02
igény vevő helyett helytáll	**-0.02**	helyett alkalmasság minimumkövetelmény mindhárom rész	0.02
nyilatkozik közbeszerzés tárgy temető	**-0.02**	helyett alkalmasság minimumkövetelmény mindhárom	0.02
eljárás folyamat ajánlat csatol cégbíróság	-0.02	tárgy szerinti rádió adó berendezés	0.02

Table A5. Terms in the award criteria field with highest and lowest coefficients (log odds), logistic regression model using red flags, control variables, and all texts, Hungary, 2011-2020

Decreasing single bidding probability		**Increasing single bidding probability**	
Term	**Coeff.**	**Term**	**Coeff.**
hónap max hónap	-0.06	**időtartam hónap teljesítés időtartam**	0.04
jótállás időtartam hónap	-0.04	**hónap teljesítés időtartam**	0.04
hónap ajánlati nettó	-0.04	vállal jótállás **időtartam hónap teljesítés**	0.04
szakmai tapasztalat hónap	-0.04	jótállás időtartam **hónap teljesítés időtartam**	0.04
hónap egyösszegű nettó	-0.04	**időtartam hónap teljesítés**	0.03
hónap egyösszegű nettó ajánlati	-0.04	jótállás **időtartam hónap teljesítés**	0.03
nettó ajánlati kwh	-0.03	időtartam nap nettó	0.02
egyösszegű nettó ajánlati	-0.03	nettó huf mennyiség	0.02
ajánlati nettó huf	-0.03	ajánlati nettó huf mennyiség	0.02
többlet jótállás időtartam	-0.03	**időtartam nap nettó ajánlati**	0.02
késedelmi kötbér mérték	-0.03	**időtartam hónap teljesítés időtartam nap**	0.02
ajánlati nettó forint	-0.03	**teljesítés időtartam nap nettó**	0.02
többlet jótállás időtartam hónap	-0.03	**hónap teljesítés időtartam nap nettó**	0.02
min hónap max	-0.03	**hónap teljesítés időtartam nap**	0.02
min hónap max hónap	-0.03	ajánlati nettó fizetési határidő nap	0.02
ajánlati fizetési határidő	-0.03	**teljesítés időtartam hónap nettó ajánlati**	0.02
jótállás vállal időtartam	-0.03	**időtartam hónap teljesítés időtartam hónap**	0.02
jótállás időtartam ajánlati	-0.02	**hónap teljesítés időtartam hónap nettó**	0.02
nettó ajánlati ban	-0.02	**hónap teljesítés időtartam hónap**	0.02
felelős műszaki vezető	-0.02	**teljesítés időtartam hónap nettó**	0.02