Predicting outcomes of Italian VAT decisions [1]

Federico GALLI [b], Giulia GRUNDLER [c], Alessia FIDELANGELI [b], Andrea GALASSI [c,2], Francesca LAGIOIA [a,b,2], Elena PALMIERI [c], Federico RUGGERI [c], Giovanni SARTOR [a,b] and Paolo TORRONI [c]

[a] *European University Institute, Law Department, Italy*
[b] *CIRSFID Alma-AI, Faculty of Law, University of Bologna, Italy*
[c] *DISI, Alma-AI, University of Bologna, Italy*

Abstract. This study aims at predicting the outcomes of legal cases based on the textual content of judicial decisions. We present a new corpus of Italian documents, consisting of 226 annotated decisions on Value Added Tax by Regional Tax law commissions. We address the task of predicting whether a request is upheld or rejected in the final decision. We employ traditional classifiers and NLP methods to assess which parts of the decision are more informative for the task.

Keywords. Predictive Justice, Machine Learning, Natural Language Processing, Case Law, Tax Law

1. Introduction

Outcome prediction has recently enjoyed renewed interest thanks to the availability of judicial data and breakthroughs in machine learning and NLP techniques [1,2,3]. Current approaches rely either on features describing aspects of the cases [4,5], which could be unrelated to their merit [6,7]; or on the textual content of the case decisions [8,9]. Our study falls under the second approach, which applies analytics techniques to automatically identify correlations between the textual content of decisions and their outcomes. In particular, we aim to determine the correlations between the requests by the parties and the uphold/rejection of such requests by the Regional Tax Commissions (second–instance Tax Courts).

Recent advances in outcome prediction include work by Aletras et al. [8], who predicted violations of some articles of the European Convention on Human Rights, using a dataset of 584 European Court of Human Rights decisions using Support Vector Machine (SVM), Bag-of-Words (n-grams) and topical features and later by [9], who expanded said dataset to obtain higher performance. Several works focused on national case law. For example, [10] applied a linear SVM classifier trained on lexical features to predict the legal area and the outcome of cases by the French Supreme Court. [11] used logistic regression and SVM to predict the outcomes of Bavarian court decisions. Chinese case law was addressed by [12,13,14] among others. In the field of tax law, [15] applied Naive

[1] This work has been partially supported by the H2020 ERC Project "CompuLaw" (G.A. 833647); the ADELE project (G.A. 101007420) under the European Union's Justice programme, the LAILA project (G.A. 2017NCPZ22) under the Italian Ministry of Education and Research's PRIN programme

[2] Corresponding authors: Francesca Lagioia: francesca.lagioia@eui.eu, Andrea Galassi: a.galassi@unibo.it

Bayer classifiers to predict the outcome of second-instance decisions based on metadata and case features extracted automatically. To the best of our knowledge, ours is the first study on outcome prediction of Italian decisions, and also the first one in the specific domain of VAT. We focus on appeal (second-instance) decisions. We model this as a binary classification task, whose goal is predicting whether a given request by the parties is accepted or rejected by the appeal court. A distinctive aspect of our work consists in covering requests and decisions addressing different aspects of VAT (e.g., taxable transactions, exemptions, out-of-scope transactions) rather than a single specific issue.

2. The corpus

The source corpus consists of 226 Italian second-instance decisions on Value Added Tax (VAT) by the Regional Tax Commissions from various judicial districts.[3] The decisions, downloaded from the *Giustizia Tributaria* database,[4] range from 2010 to 2022 and concern taxable transactions, exemptions, out-of-scope transactions, and the right to obtain a deduction. They contain 303 first-instance requests, of which 84 rejected, 126 upheld, and 5 with other outcomes, and 490 second-instance requests, of which 129 rejected, 99 upheld, and 22 with other outcomes. The number of requests is higher than the number of outcomes since a decision on a particular request may imply the uphold or rejection of other requests, which thus are not explicitly addressed. We chose to focus on VAT Italian cases since: (a) though some AI applications exist within the Italian Tax Administration, they do not yet address the case law; (b) VAT is harmonised at the European level, governed by the VAT Directive (Directive 2006/112/EC);[5] (c) the CJEU case law on this matter favours the uniform and consistent interpretation of legal norms, principles and concepts; (d) VAT is a relatively narrow self-contained branch of the law; (e) Italian VAT decisions have a rather consistent structure; (f) they affect –apart from lawyers– accountants, public servants and millions of taxpayers; (g) we have domain expertise.

Appeal VAT decisions have a standard structure consisting of the following parts:

1. *Introduction*, reporting (i) the number of the decision; (ii) the composition of the judicial panel, (iii) the parties and their lawyers (if present);
2. *Account of the Proceeding*, reporting facts related to both the pre-litigation phase and the first-instance proceedings (e.g., the parties' requests, claims and arguments as well as first-instance decisions by the Provincial Tax Commission);
3. *Parties' Requests* in second-instance proceedings, often presented with the related claims and arguments;
4. *Justification*, the statement of reasons in fact and in law supporting the decisions;
5. *Final Ruling*, by the Regional Tax Commission, including the decision on costs.

Annotation guidelines were defined through an iterative refinement process of validation, evaluation of the agreement, and discussion. The labelling was done by two VAT experts. The conflicts between annotators have been discussed and solved with a third legal expert. We focused on the identification of the following elements: (i) the parties, (ii) their first and second-instance requests, (ii) the related claims and (iii) arguments; (iv) the Provincial and Regional Tax Commissions' justifications, and (v) first and second-

[3] The corpus and our code are available at `https://github.com/adele-project/italianVAT`

[4] Tax Justice database accessible at: `https://www.giustizia-tributaria.it/`.

[5] In Italy, the EU VAT Directive has been implemented by the Presidential Decree 633/1972.

instance decisions, as reported in the different parts of the analysed documents. Such information can be of different lengths and details. Moreover, it is often enclosed within the same portion of text. For this reason, we identified hierarchical levels of annotation.

The parties to the proceeding (*part*) – i.e., taxpayers and tax authorities, being appellants or respondents in the appeal proceedings, plaintiff or defendants in first-instance proceedings – are mentioned in the introductory section and are identified through their names and residential addresses. The parties' requests, claims, and arguments concerning the first-instance proceeding are presented in the *Account of the proceeding* section, while those concerning the second-instance are reported in the *Parties' Requests* section. Claims and arguments may be missing for certain requests, especially first-instance requests. Requests, claims, and arguments are often included in the same sentence. To identify the relevant segments we relied on (a) recurrent linguistic indicators, including keywords and word patterns; and (b) context indicators, as detailed in the following.

Requests (*req*) may be distinguished in *main requests* and responses to them, i.e. *counter-requests*. They are often characterised by different linguistic indicators, which may help the annotators in correctly labelling the relevant textual fragments. In first-instance proceedings, main requests are made by taxpayers and often concern the annulment of the Tax Administration's acts. Those made in the second–instance can be presented either by taxpayers or by tax authorities and are often aimed at reversing first-instance decisions. The set of keywords and word patterns signalling the main request includes (a) verbs expressing the action of requesting or concluding with a request; (b) nouns identifying the measure requested, such as the reversal of the first-instance decision; (c) word patterns specifying these ideas. Counter request(s) are usually signalled by word patterns referring to requests for the rejection of the appellant's claim or the acceptance of the respondents' claim. Each request is denoted by (i) a unique id, (ii) the degree of judgement in which it has been made and (iii) the party making the request.

Claims (*claim*) are the ultimate reasons for grounding a request, usually supported by premises. They may concern (a) substantive facts (e.g., the lack of competence of the administrative tax office in adopting a particular pre-litigation decision), or (b) procedural facts (e.g., the violation of a procedural norm). Each claim is denoted by 3 mandatory attributes (id, degree and party making the claim), as well as 2 optional attributes used to identify whether a claim is supporting or attacking a request. Recurrent linguistic indicators include: (a) a set of terms, and in particular, certain verbal forms indicating an argumentative attitude; and (b) word patterns having the same function.

Arguments (*arg*) are statements that support or attack a claim. Arguments can be legal or factual. Each argument has the usual three mandatory attributes (id, degree and party making the argument), plus two optional attributes specifying whether an argument supports or attacks one or more claims. An argument is often denoted by word patterns referring to a grounding relation.

Justifications (*mot*) report the inferences made by the Court, leading to decisions on claims or requests raised by the parties. Each justification is characterised by: (i) a unique id, (ii) the degree of the proceeding, and (iii) its object, which can be a request or a claim. Each justification is generally delimited by a heading and includes word patterns indicating the different requests/claims raised by the parties.

First-instance decisions (*dec*) are concisely presented in the *Account of the Proceedings*. Second-instance decisions are reported in the *Final Ruling* section. Each decision is denoted by (i) a unique decision id, (ii) the degree of judgement in which the deci-

Table 1. Cohen's κ for each element, attribute, and link.

Element	κ
part	0.87
req	0.85
arg	0.94
claim	0.86
mot	0.97
dec	0.91
avg	0.90

Attribute	κ
instance	0.93
party	0.70
outcome	1.00
avg	0.88

Link	κ
req-claim	0.66
claim-arg	0.78
req-mot	0.73
req-dec	0.77
claim-mot	0.37
avg	0.66

sion was taken, (iii) its object and (iv) outcome. Possible outcomes are: uphold, reject, or other (inadmissibility of the parties' requests, extinction of the proceeding, referral to the first–instance Court, or absent decision since implicit in other decisions).

2.1. Inter-Annotator Agreement

Agreement was measured on 10 documents tagged by 2 annotators. Because a marked element may consist of a sentence fragment, and each fragment can be labelled with multiple tags, we modelled the task as a multi-label binary classification task at the word level. Accordingly, we separately measured the agreement for each type of element and attribute. Table 5 shows the Cohen's κ [16] of each category. An average κ of 0.90 indicates a strong agreement. To properly evaluate the agreement on the attributes, we considered only cases with an agreement on the annotation of elements. An average κ of 0.88 indicates good agreement in all attributes. To measure the agreement on the links (i.e., the presence of attributes that express a relation between two elements), we considered each pair of element types as a separate case. For a given pair of element types, we considered for each decision all the possible pairs of elements that belong to such types (e.g., the first element must be a request, the second must be a claim). We treated agreement on links as a binary classification problem with the aim of predicting whether there is a link between a pair of elements. Results are reported in Table 5. We obtained a good agreement in almost all categories with an average κ of 0.66. For the specific cases of *req-claim* and *claim-arg* links, we also computed the agreement on the type of link, by only considering pairs where the two annotators agreed on the presence of a link, reaching the perfect score of κ=1.0 for both classes.

3. Methods

This study aims at *(i)* predicting the outcomes of second-instance decisions and *(ii)* assessing the extent to which different parts of the decision are informative for this task. Given that each decision can contain multiple requests, and each request can have a separate outcome (different from the general outcome of the case), we considered each request separately. For each of them, we identified claims (*claim*) and arguments (*arg*) as the basic information needed to predict the outcome. For this reason, we filtered out requests not associated with a claim as well as those not explicitly decided. Furthermore, we excluded those few decisions which do not reject or uphold a request (*other outcomes*) due to the lower number of samples. Thus, we addressed the task as a binary classification (reject/uphold). Our final dataset is composed of 112 rejected decisions and 71 uphold decisions.

Our aim is to predict the court's decisions on the basis of the information provided by the parties before the case. Such information are partially present in the decision, which reports the parties' *request*(s), *claim*(s), and *argument*(s). Nonetheless, in this study, we

Table 2. Results on the second-instance requests.

	Inputs	req + arg + claim			r + a + c + mot			r + a + c + dec			r + a + c + m + d		
Embedding	**Classifier**	**Avg**	**rej**	**uph**	**Avg**	**rej**	**uph**	**Avg**	**rej**	**uph**	**Avg**	**rej**	**uph**
-	Random	0.49	0.57	0.43									
-	Majority	0.38	0.76	0.00									
TF-IDF	Linear SVC	0.64	0.77	0.50	0.59	0.75	0.44	0.64	0.78	0.50	0.61	0.75	0.46
TF-IDF	Random Forest	0.49	**0.78**	0.20	0.51	0.77	0.25	0.54	**0.79**	0.30	0.51	0.77	0.24
TF-IDF	Gaussian NB	0.57	0.75	0.39	0.56	0.70	0.41	0.55	0.73	0.37	0.58	0.71	0.44
TF-IDF	K Neighbors	0.57	0.72	0.41	0.58	0.71	0.45	0.58	0.73	0.43	0.59	0.71	0.47
TF-IDF	SVC	0.47	**0.78**	0.16	0.47	**0.78**	0.16	0.47	0.78	0.16	0.47	**0.78**	0.16
SBERT	Linear SVC	**0.68**	0.77	**0.60**	**0.66**	0.76	**0.57**	**0.68**	0.76	**0.60**	**0.66**	0.75	**0.56**
SBERT	Random Forest	0.58	**0.78**	0.38	0.56	0.77	0.34	0.58	0.77	0.40	0.59	**0.78**	0.40
SBERT	Gaussian NB	0.61	0.73	0.50	0.62	0.72	0.52	0.64	0.73	0.54	0.60	0.71	0.50
SBERT	K Neighbors	0.59	0.68	0.50	0.58	0.67	0.49	0.61	0.69	0.53	0.57	0.66	0.48
SBERT	SVC	0.47	0.75	0.19	0.54	**0.78**	0.31	0.52	0.76	0.28	0.54	0.77	0.31

are also interested in assessing which part of the document can provide a valuable contribution to predicting the outcome. Since the *justification* and *decision* sections may hold important information about the outcome of a case, we decided to include those sections in the experiments, obtaining four experimental settings. In the first one, the inputs are *req*, *args*, and *claims*; the second and third are similar, but, respectively, also *mot* and *dec* are included; the fourth uses both *mot* and *dec*.

We pre-processed the decisions by removing stopwords and punctuation symbols. For each experimental setting, we concatenated together the representation obtained for the request and the ones obtained for the other sections. We adopted two representations of the input text: **TF-IDF** vectorization, which is based on the term frequency-inverse document frequency statistic; Sentence-BERT (**SBERT**) [17], a modification of the BERT model that produces semantically meaningful sentences' embeddings, mapping sentences with similar semantic content into vectors close to each other. As classifiers, we have chosen the following set of traditional machine learning models that have low computational requirements: Linear SVC, SVC, Random Forest, Gaussian Naive Bayes and K-Neighbours.[6] Experiments were conducted using 5-fold cross-validation with folds determined at the document level so that all the requests of the same decision belong to the same fold. The folds were created manually to balance their composition with respect to the reject/uphold and the first/second-instance distinctions.

4. Results

Tables 2 shows the results obtained through each combination of embeddings and classifiers in each setting, as well as two baselines (random and majority class). We measure the F1 score obtained for each class and their macro-average. The task of determining the decision outcome based only on the *claims* and *arguments* of the parties reaches a maximum score of 0.68 with Linear SVC and SBERT. The use of SBERT embeddings instead of TF-IDF is beneficial with all classifiers, leading to results not worse than the baselines. Overall, the Linear SVC classifier seems to give the best result in almost all the settings. There is a wide gap between the scores obtained in the two classes, which we speculate may be caused by their unbalanced distribution in the dataset. The *justification* section seems to give conflicting results: it slightly worsens the performance of Linear SVC, but it improves other classifiers. The introduction of the *decision* section has a

[6] We used the default hyper-parameters offered by the `sci-kit learn` library.

limited impact, slightly improving some classifiers but without improving the best case. The results obtained by the use of both sections are unexpected: most classifiers perform worse than when adding only the *decision* information. We speculate that, since in the *justification* the Court retraces the arguments of the parties, mentioning all the clues towards each possible outcome, its use may introduce noise that lowers the performance.

5. Conclusion

Ideally, one would aim to predict the decision of the court based on the information provided by the parties before the case. Our experiments approximate the ideal setup, by focusing on outcome prediction based on fragments in the narrative provided by courts, which we captured through the requests, claims, and arguments marked elements. To this end, we built a first-of-a-kind dataset, on Italian decisions and on the VAT domain. In the future, we plan to include information provided by the parties before the case. From the machine learning viewpoint, we plan to adopt oversampling or augmentation to balance the distribution of the classes in the dataset, and we are investigating more advanced neural architecture for classification or domain-specific embeddings.

References

[1] Ashley KD. A brief history of the changing roles of case prediction in AI and law. Law Context: A Socio-Legal J. 2019;36:93.

[2] Feng Y, Li C, Ng V. Legal Judgment Prediction: A Survey of the State of the Art. In: IJCAI. ijcai.org; 2022. p. 5461-9. Available from: `https://doi.org/10.24963/ijcai.2022/765`.

[3] Zhong H, Xiao C, Tu C, Zhang T, Liu Z, Sun M. How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence. In: ACL. Association for Computational Linguistics; 2020. p. 5218-30.

[4] Ashley KD, Brüninghaus S. Automatically classifying case texts and predicting outcomes. Artificial Intelligence and Law. 2009;17(2):125-65.

[5] Ashley KD, Keefer M. Ethical reasoning strategies and their relation to case-based instruction: some preliminary results. In: Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society. Routledge; 2019. p. 483-8.

[6] Surdeanu M, Nallapati R, Gregory G, Walker J, Manning C. Risk analysis for intellectual property litigation. In: ICAIL-2011. ACM; 2011. p. 116-20.

[7] Katz DM, Bommarito MJ, Blackman J. A general approach for predicting the behavior of the Supreme Court of the United States. PloS one. 2017;12(4):e0174698.

[8] Aletras N, Tsarapatsanis D, Preoţiuc-Pietro D, Lampos V. Predicting judicial decisions of the European Court of Human Rights: A natural language processing perspective. PeerJ Computer Science. 2016;2:e93.

[9] Medvedeva M, Vols M, Wieling M. Using machine learning to predict decisions of the European Court of Human Rights. Artificial Intelligence and Law. 2020;28(2):237-66.

[10] Sulea OM, Zampieri M, Malmasi S, Vela M, Dinu LP, Van Genabith J. Exploring the use of text classification in the legal domain. arXiv preprint arXiv:171009306. 2017.

[11] Urchs S, Mitrovic J, Granitzer M. Design and Implementation of German Legal Decision Corpora. In: ICAART (2); 2021. p. 515-21.

[12] Xiao C, Zhong H, Guo Z, Tu C, Liu Z, Sun M, et al. Cail2018: A large-scale legal dataset for judgment prediction. arXiv preprint arXiv:180702478. 2018.

[13] Luo B, Feng Y, Xu J, Zhang X, Zhao D. Learning to predict charges for criminal cases with legal basis. arXiv preprint arXiv:170709168. 2017.

[14] Long S, Tu C, Liu Z, Sun M. Automatic judgment prediction via legal reading comprehension. In: China National Conference on Chinese Computational Linguistics. Springer; 2019. p. 558-72.

[15] Waltl B, Bonczek G, Scepankova E, Landthaler J, Matthes F. Predicting the outcome of appeal decisions in germany's tax law. In: International conference on electronic participation. Springer; 2017. p. 89-99.

[16] Cohen J. A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement. 1960;20:37 46.

[17] Reimers N, Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: EMNLP/IJCNLP (1). Association for Computational Linguistics; 2019. p. 3980-90.

Appendix

A. Detailed Corpus Information

The source corpus consists of 226 Italian second-instance decisions on Value Added Tax (VAT) by the Regional Tax Commissions from various judicial districts. [7] The decisions concern taxable transactions, exemptions, out-of-scope transactions, and the right to obtain a deduction. In Table 3, we report some statistical data on the distribution of the parties' requests according to their outcomes, i.e., reject, uphold, and other outcomes (e.g., inadmissibility or extinction of the proceedings). Note that the number of requests is higher than the number of decisions since a decision on a particular request may imply the uphold or rejection of other requests (which thus are not explicitly addressed).

Table 3. Distribution of the decision's requests.

First–instance			Second–instance		
303			490		
Reject	Uphold	Other	Reject	Uphold	Other
84	126	5	129	99	22

The selected decisions range from 2010 to 2022. Documents have been downloaded from the *Giustizia Tributaria* database[8]. We chose to focus on VAT Italian cases since: (a) though some AI applications exist within the Italian Tax Administration, they do not yet address the case law; (b) VAT is harmonised at the European level, governed by the VAT Directive (Directive 2006/112/EC)[9]; (c) the CJEU case law on this matter favours the uniform and consistent interpretation of legal norms, principles and concepts; (d) VAT is a narrow domain and a self-contained branch of the law (even within general taxation law); (e) Italian VAT decisions present a relatively consistent structure; (f) they affect –apart from lawyers– accountants, public servants and millions of taxpayers; finally (g) our annotators have expertise in this domain.

A.1. Annotation process

Even though appeal VAT decisions are not organised in clearly separated sections, they have a standard, although not fixed, structure. In each decision it is possible to distinguish the following parts:

1. *Introduction*, reporting (i) the number of the decision; (ii) the composition of the judicial panel, (iii) the parties and their lawyers (if present);
2. *Account of the Proceeding*, reporting facts related to both the pre-litigation phase and the first-instance proceedings (e.g., the parties' requests, claims and arguments as well as first-instance decisions by the Provincial Tax Commission);
3. *Parties' Requests* in second-instance proceedings, often presented with the related claims and arguments;
4. *Justification*, namely the statement of reasons in fact and in law supporting the decisions;

[7] For reviewing purposes the corpus is available at `https://www.dropbox.com/sh/aohi2sy7qdvus7a/AACzW8OXO2vX8ODGn4X9wHRga?dl=0`

[8] The Tax Justice database is accessible at the following address `https://www.giustizia-tributaria.it/`.

[9] In Italy, the EU VAT Directive has been implemented by the Presidential Decree 633/1972

Table 4. Annotation scheme (* indicates optional attributes).

Elements and Tags	Attributes		
	Name	**Value**	**Tag**
Parties to the proceedings <part>	Party Identifier	A, B, C, D, L	P="L"
Requests <req>	Identifier	Req1, Req2, ReqN	ID="ReqN"
	Instance	1 or 2	G="N"
	Party	A, B, C, ...	P="A"
Claim(s) <claim>	Identifier	Claim1, Claim2, ClaimN	ID="ClaimN"
	Instance	1 or 2	G="N"
	Party	A, B, C, ...	P="A"
	Support *	Req1, Req2, reqN	PRO="ReqN"
	Attack *	Req1, Req2, ReqN	CON=""
Argument(s) of the parties <arg>	Identifier	Arg1, Arg2, ArgN	ID="ArgN"
	Instance	1 or 2	G="N"
	Party	A, B, C, ...	P="A"
	Support *	Claim1, Claim 2, ClaimN	PRO="ClaimN"
	Attack *	Claim1, Claim2, ClaimN	CON="ClaimN"
Motivation(s) of the Commission <mot>	Identifier	Mot1, Mot2, MotN	ID="MotN"
	Instance	1 or 2	G="N"
	Object	Req1, Req2, ReqN Claim1, Claim2, ClaimN	O="ReqN" O="ClaimN"
Decision(s) <dec>	Identifier	Dec1, Dec2, DecN	ID="DecN"
	Instance	1 or 2	G="N"
	Object	Req1, Req2, ReqN	O="ReqN"
	Outcome	1 (uphold), 0 (reject), -1, CMC, ABS, REM (other outcomes)	E="N"

5. *Final Ruling*, by the Regional Tax Commission, including the decision on costs.

As noted in Section 1, we assumed that a ML classifier can be trained on some input data (e.g., the parties' requests, claims and arguments) connected with labels (outcomes) so as to infer patterns characterising the relationship between the two.

In order to create a data set that we can use for our experiments, we defined a set of annotation guidelines. They were written and refined through multiple stages of validation, evaluation of the agreement, and discussion. The labelling was done by two annotators with expertise in the VAT domain. The conflicts between annotators have been discussed and solved with a third legal expert.

In particular, we focused on the identification of the following elements: (i) the parties, (ii) their first and second-instance requests, (ii) the related claims and (iii) arguments; (iv) the Provincial and Regional Tax Commissions' justifications, and (v) first and second-instance decisions, as reported in the different parts of the analysed documents. Such information can be of different lengths and details. Moreover, it is often enclosed within the same portion of text. For this reason, hierarchical levels of annotation were identified, as shown in Table 4. For each element, we established a set of mandatory and optional attributes and their possible values. If the value of an attribute is composed of multiple items, each item has been separated by the vertical bar (|).

A.1.1. The parties to the proceeding

The parties to the proceeding (<*part*>) – i.e., taxpayers and tax authorities, being appellants or respondents in the appeal proceedings, plaintiff or defendants in first-instance proceedings – are mentioned in the introductory section and are identified through their names and residential addresses.

In the annotation process, we denoted each party through an attribute (*P*=""), whose value (i.e., a letter of the Italian alphabet) uniquely identifies that party throughout the document. This makes it possible to connect the requests, claims and arguments by the same party. Consider the following example:

> *. . . <part P="A"> proposto dall'ufficio: AG. ENTRATE DIREZIONE PROVINCIALE MODENA</part> <part P="B"> controparte: INTERTECNICA SRL ARREDAMENTI VIA PER MODENA 125 FR.VILLAVARA 41030 BOMPORTO MO </part> . . .* (Comm. trib. reg. Emilia-Romagna, Sez. XIX, 17 July 2013, n. 59, lines 21-25).

A.1.2. Requests, claims and arguments

The parties' requests (<*req*>), claims (<*claim*>), and arguments (<*arg*>) concerning the first-instance proceeding are presented in the *Account of the proceeding* section, while those concerning the second-instance are reported in the *Parties' Requests* section. Note that claims and arguments may be missing for certain requests. This is more frequent for first-instance requests. Requests, claims, and arguments are often included in the same sentence. To identify the relevant textual fragments we relied on (a) recurrent linguistic indicators, including keywords and word patterns; and (b) context indicators, as detailed in the following.

Requests. Requests may be distinguished in *main requests* and responses to them, i.e. *counter-requests*. They are often characterised by different linguistic indicators, which may help the annotators in correctly labelling the relevant textual fragments. In first-instance proceedings, main requests are made by taxpayers and often concern the annulment of the Tax Administration's acts. Those made in the second–instance can be presented either by taxpayers or by tax authorities and are often aimed at reversing first-instance decisions. The set of keywords and word patterns signalling the main request includes (a) verbs expressing the action of requesting ("chiede") or concluding with a request ("conclude"); (b) nouns identifying the measure requested, such as the reversal of the first-instance decision ("riforma integrale"); (c) word patterns specifying these ideas ("Chiede la riforma della sentenza", "Conclude, quindi, per", "In via subordinata chiede", "Conclusivamente l'appellante chiede", "…propone appello avverso tale sentenza chiedendo").

Counter request(s) are usually signalled by word patterns referring to requests for the rejection of the appellant's claim or for the acceptance of the respondents' claim ("Nell'odierno giudizio si è costituita l'Agenzia delle Entrate chiedendo il rigetto"; "chiede"; "Si costituiva l'Ufficio, confermando la regolarità del proprio operato e chiedendo il rigetto del ricorso"; "Si è costituita in giudizio l'appellata che chiede il rigetto dell'appello e la conferma della sentenza impugnata"; "L'Agenzia delle Entrate depositano controdeduzioni chiede la conferma").

Each request (<*Req*>) is denoted by (i) a unique identifier (*ID*), (ii) the degree of judgement in which it has been made (*G*) and (iii) the party making it (*P*) (see Table 4). The ID value is constructed by joining the *Req* abbreviation with a progressive number, which distinguishes the single request within the decision (e.g., *Req1, Req2, ReqN*). Consider the following examples:

> *<req ID="Req1" G="1" P="A"> Avverso tale provvedimento agenziale la contribuente proponeva ricorso.</req>* (Comm. trib. reg. Sicilia, Sez. II, 13 April 2021, n. 6074, lines 61-64).

. . . <req ID="Req4" G="2" P="A"> chiede pertanto il rigetto dell'appello.</req> (Comm. trib. reg. Sicilia, Sez. II, 13 April 2021, n. 6074, line 124).

Claims. Claims are the ultimate reasons for grounding a request, usually supported by premises. They may concern (a) substantive facts (e.g., the lack of competence of the administrative tax office in adopting a particular pre-litigation decision), or (b) procedural facts (e.g., the violation of a procedural norm).

Each claim is denoted by 3 mandatory and 2 optional attributes. *Mandatory attributes* include: (i) a unique identifier (*ID*), (ii) the degree of judgement in which the claim has been made (*G*) and (iii) the party making it (see Table 4). The ID value is constructed by joining the *Claim* word with a progressive number, which distinguishes each claim within the decision (e.g., *Claim1, Claim2, ClaimN*). *Optional attributes* are used to identify whether a claim, in turn, is (i) supporting (*PRO*) or (ii) attacking (*CON*) a request. The values of *PRO* and *CON* are the IDs of the concerned request(s).

Recurrent linguistic indicators include: (a) a set of terms, and in particular, certain verbal forms indicating an argumentative attitude ("lamentando", "eccependo" "sostendendo", "deducendo"); and (b) word patterns having the same function ("L'Amministrazione contesta", "Deduce la parte che", "la parte rileva che", "La parte propone appello per i seguenti motivi"). Consider the following examples:

<claim ID="Claim1" G="2" P="D" PRO="Req2"> lamentando l'infondatezza ed illegittimità della sentenza impugnata, senza, peraltro, indicare specificamente i motivi di gravame.</claim> (Comm. trib. reg. Basilicata, Sez. II, 19 June 2018, n. 205, line 72-73).

<claim ID="Claim2" G="2" P="A|B|C" PRO="Req3">In particolare, veniva eccepita la genericità dei motivi di doglianza</claim> (Comm. trib. reg. Basilicata, Sez. II, 19 June 2018, n. 205, line 78).

<claim ID="Claim7" G="2" P="B" PRO="Req2">a) Con il primo motivo dell'appello principale, 1'Ufficio si duole del fatto che la vendita dell'immobile ad Anna Abbate non è stata considerata sottofatturata,</claim>. . . (Comm. trib. reg. Emilia-Romagna, Sez. II, 24 May 2019, n. 1176, line 82).

Arguments. By an "argument" we mean a statement that supports or attacks a claim. Arguments can be legal or factual.

We defined three mandatory and two optional attributes. Mandatory attributes include: (i) a unique argument identifier (*ID*), (ii) the degree of judgement in which the argument was made (*G*) and (iii) the party making it (*P*) (see Table 4). The ID value is constructed by joining the *Arg* abbreviation with a progressive number, which distinguishes each argument within the decision (e.g., *Arg1, Arg2, ArgN*). The *PRO* or *CON* attributes specify whether an argument supports or attacks one or more claims. Their values correspond to the IDs of the concerned claims.

An argument is often denoted by word patterns referring to a grounding relation such as *"in considerazione di"*, *"in forza di"*, *"in quanto"*, *"atteso che"*. Consider the following examples:

<arg ID="Arg1" G="2" P="B|C\D|E|F" PRO="Claim1">in quanto, in caso di contestazione, spetta all'ufficio dare prova della corretta delega che non è stata documentata neppure in corso di giudizio cosicchè la stessa è irregolare.</arg> (Comm. trib. reg. Umbria, Sez. III, 18 November 2019, n. 368, lines 114-155).

Table 5. Cohen's kappa for each element, attribute, and link.

Element	κ
part	0.87
req	0.85
arg	0.94
claim	0.86
mot	0.97
dec	0.91
avg	0.90

Attribute	κ
instance	0.93
party	0.70
outcome	1.00
avg	0.88

Link	κ
req-claim	0.66
claim-arg	0.78
req-mot	0.73
req-dec	0.77
claim-mot	0.37
avg	0.66

<arg ID="Arg2" G="1" P="A" PRO="Claim2">trattandosi di interessi relativi a finanziamenti effettuati da società asseritamente controllata e per un importo superiore a quello previsto dalla legge per la deducibilità;</arg> (Comm. trib. reg. Emilia-Romagna, Sez. II, 24 May 2019, n. 1176, lines 56-57).

A.1.3. The justification

Justifications (<*mot*>) report the inferences made by the Court, leading to decisions on claims or requests raised by the parties.

Each justification is characterised by: (i) a unique identifier (*ID*), (ii) the degree of the proceeding (*G*), and (iii) its object (*O*), which can be a request or a claim (see Table 4). The ID value is constructed by joining the *Mot* abbreviation with a progressive number, which distinguishes each justification (e.g., *Mot1, Mot2, MotN*). The object's value corresponds to the IDs of the addressed request(s) and/or claim(s).

Concerning the contextual indicators, each justification is generally delimited by a heading and includes word patterns indicating the different requests/claims raised by the parties ("Per quanto riguarda il primo motivo", or "Quanto alla richiesta di").
As an example, consider the following:

. . . 3) <mot ID="Mot3" G="1" O="Req1">quanto all'indebita detrazione IVA, "La tesi della ricorrente secondo cui l'operazione all'origine del credito d'imposta indebitamente detratto non sarebbe esente Iva, come rilevato invece dall'Ufficio, non merita all'evidenza accoglimento. ".</mot> . . . (Comm. trib. reg. Lombardia, Sez. XI, 23 April 2021, n. 2471, lines 114-140).

A.1.4. The decisions

First-instance decisions are concisely presented in the *Account of the Proceedings*, while second-instance decisions are reported in the *Final Ruling* section. Each decision is denoted by: (i) a unique decision identifier (*ID*), (ii) the degree of judgement in which the decision was taken (*G*), (iii) its object (*O*) and (iv) outcome (*E*) (see Table 4). The ID value is constructed by joining the *Dec* abbreviation with a progressive number (e.g., *Dec1, Dec2, DecN*). The object value corresponds to the request(s)' ID(s) addressed by each decision. Depending on the decision's outcome, the *E* value is equal to 1 (uphold), 0 (reject), -1 (inadmissibility of the parties' requests), CMC (extinction of the proceeding), REM (referral to the first–instance Court), ABS (absent decision since implicit in other decisions). The last four on the list have been generally considered as *other outcomes*.
As an example, consider the following:

. . . <dec ID="Dec2" G="2" E="0" O="Req2" D="Find6bis">la Commissione Tributaria Regionale di Bologna sez. XI - rigetta l'appello;</dec> (Comm. trib. reg. Emilia-Romagna, Sez. XI, 18 October 2019, n. 2227, line 220).