# Exploring the Language of Data

**Gábor Bella**
University of Trento
via Sommarive, 5
38123 Trento, Italy
gabor.bella@unitn.it

**Linda Gremes**
University of Trento
via Sommarive, 5
38123 Trento, Italy
linda.gremes@alumni.unitn.it

**Fausto Giunchiglia**
University of Trento
via Sommarive, 5
38123 Trento, Italy
fausto.giunchiglia@unitn.it

## Abstract

We set out to uncover the unique grammatical properties of an important yet so far under-researched type of natural language text: that of short labels typically found within structured datasets. We show that such labels obey a specific type of *abbreviated grammar* that we call *the Language of Data*, with properties significantly different from the kinds of text typically addressed in computational linguistics and NLP, such as 'standard' written language or social media messages. We analyse orthography, parts of speech, and syntax over a large, bilingual, hand-annotated corpus of data labels collected from a variety of domains. We perform experiments on tokenisation, part-of-speech tagging, and named entity recognition over real-world structured data, demonstrating that models adapted to the Language of Data outperform those trained on standard text. These observations point in a new direction to be explored as future research, in order to develop new NLP tools and models dedicated to the Language of Data.

## 1 Introduction

Structured data, such as database tables or XML trees, frequently contain short natural language labels that either describe the data structure itself (attribute names) or provide content (attribute values). In high-precision data integration and analysis applications, where mistakes in interpretation pose a significant risk, such as in health data analytics or emergency response, an unambiguous parsing of such pieces of text is crucial. Conventional NLP tools, such as supervised sequence labellers or embeddings trained on full sentences, do not, however, perform well on structured data. Firstly, data labels may be very short and thus devoid of context necessary for adequate NLP performance. Secondly, they are not self-contained: their interpretation relies on the data structure surrounding them, such as other values of the same attribute or record. And thirdly, their grammar shows significant difference with respect to the language on which NLP tools are typically trained.

This paper explores the grammar of structured data labels—that we succinctly designate as the *Language of Data* or *LoD* for short—according to three main hypotheses: (1) *uniqueness:* it is markedly different from other forms of language typically studied in computational linguistics, such as short social media messages, encyclopaedic articles, literature, or newswire; (2) *uniformity:* it displays coherent grammatical characteristics across domains, data sources, and even languages; and (3) *usefulness:* its distinctive features can be exploited to improve the performance of computational language analysis.

Following an empirical approach, our paper verifies the first two hypotheses through an extensive corpus analysis, and the third one through actually building and evaluating an initial set of NLP tools tailored to the grammar of short text labels. Accordingly, the contributions of this paper are (1) a freely available, expert-annotated, 120k-token, bilingual (English–Italian) dataset on which most of our analysis is based; (2) results of the comparative analysis of the annotated dataset in terms of orthography, parts of speech, and syntax, with respect to other types of text; (3) tokeniser, part-of-speech (PoS) tagger, and named entity recogniser (NER) models trained and evaluated on structured data labels.[1]



[1]The resources are downloadable from http://www.languageofdata.science.



*Proceedings of the 28th International Conference on Computational Linguistics*, pages 6638–6648
Barcelona, Spain (Online), December 8-13, 2020

The rest of the paper is organised as follows. Section 2 explores research relevant or related to our study. Section 3 provides an overview of the principal characteristics of the Language of Data, as well as the methodology of our analysis. Sections 4 to 6 analyse the language of data in terms of orthography, part-of-speech distributions, and syntax. Section 7 provides first results in building NLP tools for the language of data, evaluating them against tools trained on standard text. Finally, section 8 presents conclusions and sets the stage for what we consider as a new field of research on the Language of Data.

## 2 State of the Art

The analysis of the grammar of short pieces of text dates back to (Straumann, 1935) and (Leech, 1966) who first described the *block language* of English newspaper headlines and of the *disjunctive language* of advertising in general. These early works have inspired (Cook, 2013) to analyse the 'language of the street' (i.e. signs and inscriptions), as well as ourselves to regard the short sentences of structured data as belonging to a specific form of these grammars. Nevertheless, as we show later, the Language of Data also presents specificities with respect to 'headlinese' and advertising slogans.

More recent research regarding *short text* concentrates on *social media messages* (e.g. tweets) often with a focus on opinion mining, *web search queries* with a focus on information retrieval, and on *schema* or *classification labels* with a focus on disambiguation and/or downstream semantic alignment tasks. Among these, works on understanding short social media messages (see (Song et al., 2014) for a survey) are mostly irrelevant to us due to fundamental linguistic differences in terms of syntactic structure, noisiness, discursivity, use of an extremely informal spoken register, etc., as demonstrated in our paper.

Works on the matching or disambiguation of data schemas or classification labels are, on the other hand, highly relevant to us in terms of the underlying objectives. (Autayeu et al., 2010) analyse the grammar of short labels used in classification hierarchies, the results of which can be exploited for the optimisation of NLP tasks, such as in cross-lingual alignment (Bella et al., 2017). Classification labels, however, are usually expert-curated and thus follow strict and regular grammars exclusively constituted of noun phrases with coordinating conjunctions. We therefore consider the language of such labels as a narrow, specific subset of the LoD. Works on the disambiguation of data schemas and underlying data values (Tekli, 2016; Tagarelli et al., 2009) typically approach the problem from a point of view of lexical semantics—i.e. the label is considered as a bag of words or word meanings—and concentrate on the exploitation of the context provided by the tabular or tree-based data structure for the purposes of disambiguation. They do not consider the grammar of the labels, which is the focus of our research. Finally, in preliminary work (Bella et al., 2016), the authors explore the problem of word sense disambiguation over text in structured data, and mention some of the key characteristics of the LoD in order to motivate the development of adapted NLP pipelines. They do not, however, provide any corpus-based empirical proof of the linguistic characteristics assumed, which is the main objective of our work.

Results on the analysis of search queries are also relevant given that, as we demonstrate, they are similar in certain aspects to structured data contents, such as in average length or their limited use of parts of speech. (Li, 2010) considers search queries as noun phrases and defines their semantic constituents as *intent heads* and *modifiers*. Still, they use the results of regular PoS taggers and chunkers as input features for learning-based methods, all the while being aware of their sub-optimal performance on short text. (Wang et al., 2014) instead consider that such queries *'do not follow grammar rules'* and attempt to identify the former constituents without relying on any syntactic feature.

We situate our work in addition, rather than in opposition, to those described above. We believe that the grammar-agnostic approaches that rely on keywords, lexical semantics, and structured context can be successfully complemented by tools and resources built in awareness of the underlying specific grammar.

## 3 The Language of Data: Characteristics and Methodology

Through an empirical, bottom-up corpus analysis of data labels, the results of which we present in sections 4–6, we have synthesised the main characteristics of the LoD as follows: (1) *abbreviated grammar* used for brevity, which manifests itself in the shortness of labels, the frequent use of abbreviations, non-standard word separation or lack thereof (*'firstname'*), as well as a *block language* syntax that is based

on the omission of words (*'age [of] pension'*); (2) *disjunctive grammar*, which consists of independent non-finite and minor sentences (*'road closed'*); (3) *named entities* appear in a large proportion of labels; and (4) *formal writing*, which manifests itself in orthography inspired from formal languages (*'zipCode'* or *'zip_code'*), change of word order (*'Smith John'*), as well as the presence of abstract expressions (*'Income <= 133%'*) and clues for the correct interpretation of labels (*'area (sqm)'*).

The term *block language* was first used by (Straumann, 1935) who examined the sparing syntax of newspaper headlines that frequently omit function and other words. Later, (Leech, 1966) used the term *abbreviated grammar* to define more generally the grammar of public signs and the linguistic register of advertising. In the Language of Data, the frequent use of abbreviated forms on both word and phrase level may be motivated partly by space saving concerns (e.g. in data entry forms), partly by an inspired adoption of the register of informative signs, e.g. in street language.

The term *disjunctive grammar*, also coined by Leech, refers to a syntax consisting of non-finite or minor (verb-less) sentences as independent clauses.[2] We explain the dominance of such constructs in the LoD by the structured nature of data itself: a *data value* is always interpreted in the context of a *data record* (i.e. the 'row') and an *attribute* (the 'column'). In terms of semantics, the data value, the record, and the attribute correspond to the object, subject, and predicate of a full sentence, respectively. Thus, a data value *'dark brown'* within a record describing *'Mr Jones'* and under the attribute *'hair'* can be transformed into the full sentence *'The hair of Mr Jones is dark brown.'* None of the three labels needs to be fully grammatical, as the data structure implies the syntactic 'glue' that joins them.

As to the presence of formal writing in the LoD, we attribute it to the computational context. The use of word separators such as underscores or *camelCasing* is likely inspired by computer programming practices. The change of word order often happens due to sorting or classification requirements. The frequency of interpretative clues (such as providing the unit of measure in attribute names) is explained by disambiguation needs for data entry or subsequent interpretation and analytics. Finally, the fact that the primary purpose of structured data is to describe or refer to real-world entities (such as *Mr Jones*) explains the high proportion of named entities within data.

Our analysis of the LoD was based on a new, human-annotated 120k-token corpus of natural language labels extracted from structured data, that we used to compute statistics and insights on the characteristics of the LoD, as well as to train a first batch of dedicated NLP tools. Annotation and subsequent validation were carried out by two language experts with adequate knowledge of both English and Italian. They annotated: (a) token and sentence boundaries; (b) parts of speech; (c) named entity categories; (d) the presence of abbreviations; (e) the presence of non-standard syntax and its categorisation (see section 6). For cross-comparability, the annotators used the standard Penn Treebank PoS tagset (Santorini, 1990) for English and the widely used ISST-TANL tagset[3] for Italian. For NER tagging, we used the following categories: *person, organization, geopolitical entity, date, address, email address, website, phone number*, and *misc*. The annotated corpus is downloadable from the web.[4]

In order to reduce bias (and thus reduce the generality of our study) with respect to any single language, domain, or data source, we crawled *open data* from the web: not only are these datasets large and free for use, they also exist in multiple languages, cover many domains, and are produced by authorities of all kinds. We downloaded and preprocessed data in two languages (English and Italian) from five open data catalogues in four countries (UK, USA, Australia, Italy), using a custom-built crawler tool. We downloaded only CSV (i.e. tabular) data files: CSV is by far the most widely used, simplest, and most robust structured open data format. The tool separately collected attribute names (that we will call *headers* in the rest of the paper) and textual data values, where the preliminary heuristic for a string to be textual was to contain at least three consecutive alphabetical Unicode characters. In order to ensure variety within the corpus, repetitions of labels were eliminated, and only the first five data records of every dataset were extracted, in order to avoid bias due to dataset size. The tool filtered out labels longer than 300 characters: such labels were found to contain, without exception, regular grammatical text that

[2]The sign 'Conference Registration Desk' is an example of a minor and 'Road closed' of a non-finite sentence.

[3]`http://www.italianlp.it/docs/ISST-TANL-POStagset.pdf`

[4]`http://www.languageofdata.science`

| Data Source | Lang | Provenance | Hdr Labels | Hdr Tokens | Val Labels | Val Tokens |
|---|---|---|---|---|---|---|
| dati.trentino.it | Italian | Trentino, Italy | 1,308 | 4,213 | 3,739 | 25,251 |
| dati.comune.milano.it | Italian | Milano, Italy | 2,228 | 5,510 | 2,789 | 14,266 |
| data.gov.uk | English | UK | 3,000 | 13,684 | 3,000 | 16,108 |
| data.gov.au | English | Australia | 3,003 | 9,658 | 3,000 | 13,166 |
| data.illinois.gov | English | Illinois, US | 2,555 | 7,785 | 2,731 | 10,424 |
| LANGUAGE OF DATA TOTAL | | | 12,251 | 40,850 | 18,470 | 79,215 |

Table 1: Annotated data sources and sizes in terms of number of (distinct) labels and number of tokens.

falls outside of the scope of our study. Finally, labels underwent human filtering to eliminate pseudo-textual labels, such as British postcodes. From each catalogue, roughly 3,000 header and 3,000 data labels underwent expert-annotation (see Table 1), amounting to 120k annotated tokens overall.

In order to test the *uniqueness* of the LoD, we compared our results to (subsets of) well-known annotated *control corpora*: the Brown Corpus (genres used: newswire, adventure, government, belles lettres, 413k tokens), the PoS-tagged English and Italian Wikipedia (4.3M and 15k tokens, respectively) (Bosco et al., 2000; Reese et al., 2010), a PoS-tagged Twitter corpus (1.55M tokens) from (Derczynski et al., 2013), and the *TREC 2007 million-query corpus* of search query logs (Carterette et al., 2009), a 2.6k-token subset of which we manually annotated in the same way as described above. To test for *uniformity*, we performed comparisons among the open data corpora: across languages, datasets, as well as between headers and values. Finally, to test for *usefulness*, we used the annotated datasets to train NLP tools and evaluate them on common tasks, confronting them with models trained on state-of-the-art corpora.

## 4 Surface Statistics and Orthography

Orthographic and other surface aspects of texts play a crucial role in computational tasks: as they are often dealt with as part of preprocessing, their incorrect handling affects the entire downstream pipeline.

**Label length.** Figure 1 shows token count distributions for all header (dotted), value (continuous), and unstructured (dashed) corpora, while Table 2 provides aggregate statistics.We observe a remarkable coherence among header labels from all five catalogues in terms of average, median, and most frequent lengths, and the proportion of long ($>$ 9-token) labels. The query corpus proves to be the most similar to headers, although typically they are 1–2 tokens longer. Long queries are just as rare as long headers (1.3–2.7%, the UK headers being somewhat of an outlier), while long data labels are more frequent (5.2–24.4%). Tweets and sentences in 'standard' language show entirely different characteristics. Overall, the majority of labels in the LoD tend to be very short (the bulk being in the 1–5 token range), even if longer labels are still present especially within data values. The shortness of LoD labels has to be taken into account when designing NLP solutions for structured data.

**Phrase and word separation.** In the LoD, non-standard phrase and word separation practices are frequent: the *use_of_punctuation to/separate/words*, *camelCasing*, *TitleCasing*, or the overall *omissionoftheseparator*. The characters '/', '|', and '-' are also frequently used to separate distinct phrases and sentences, a practice very rare outside of the LoD. Table 2 shows that non-standard separation is extremely frequent in headers (19–63% of all header labels), while it plays a smaller but still non-negligible role in data values (5–8%). NLP on headers (attribute and element names, etc.) thus requires non-standard tokenisation techniques that are able to cope with the phenomena described above. In contrast, search queries mostly consist of a single phrase, punctuation is rare, and words are separated by spaces.

**Capitalisation.** A major feature of the LoD with respect to standard text is frequent non-standard capitalisation. In regular text, initial capitals mark the beginning of a sentence, names, roman numerals, etc. In search queries, capitalisation is rare as users tend to type queries in all-lowercase. In the LoD, the use of capitals is more fuzzy: standard usage of capitalisation coexists with text in *IN ALL CAPITALS* or all-lowercase, and *Initial Capitals* abound even outside names. A significant proportion (in some

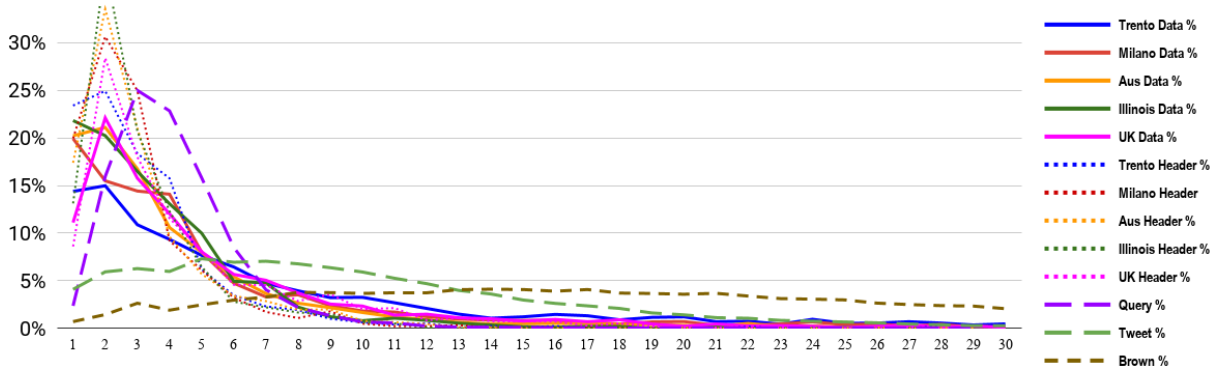


Figure 1: Distribution of the number of tokens per label/sentence. Continuous lines: data values. Dotted lines: headers. Dashed lines: the query, tweet, and Brown corpora.

| | Label length | | | | Non-std separation | | | †Non-std | |
|---|---|---|---|---|---|---|---|---|---|
| **Corpus** | **mean** | **med** | **mfreq** | **>9tok** | **punct** | **absent** | **casing** | **capitals** | **†Abbr** |
| Trento headers | 3.22 | 2 | 2 | 2.7% | 14.0% | 0% | 3.1% | 16.8% | 30.4% |
| Milano headers | 2.95 | 2 | 2 | 1.3% | 34.7% | 2.1% | 2.9% | 26.0% | 14.7% |
| Aus headers | 3.23 | 2 | 2 | 2.5% | 31.1% | 3.5% | 5.5% | 33.0% | 5.7% |
| Illinois headers | 3.05 | 2 | 2 | 1.3% | 49.9% | 4.7% | 8.5% | 62.0% | 16.7% |
| UK headers | 4.54 | 3 | 2 | 9.8% | 15.3% | 1.3% | 14.0% | 71.0% | 5.2% |
| Trento values | 7.05 | 4 | 2 | 24.4% | 6.4% | 0% | 0.1% | 18.5% | 3.6% |
| Milano values | 5.28 | 3 | 1 | 14.2% | 5.3% | 0% | 0.3% | 31.0% | 4.5% |
| Aus values | 4.47 | 3 | 2 | 9.5% | 7.0% | 0.6% | 0.4% | 14.7% | 4.2% |
| Illinois values | 3.84 | 3 | 1 | 5.2% | 4.2% | 0% | 0.4% | 55.7% | 8.2% |
| UK values | 5.38 | 3 | 2 | 14.0% | 6.2% | 0% | 0.2% | 45.3% | 4.6% |
| Query | 4.13 | 4 | 3 | 2.0% | 0.6% | 0% | n/a | n/a | 3.8% |
| Twitter | 9.67 | 8 | 5 | 43.3% | *0.2% | n/a | *0.4% | 39.7% | 0.8% |
| Brown | 16.95 | 16 | 14 | 77.1% | 0.5% | n/a | 0% | 2.7% | 0.6% |

Table 2: Label length and orthography statistics: mean / median / most-frequent length, proportion of labels longer than 9 tokens, non-standard separation by punctuation / absent / through casing, labels with non-standard capitalisation, proportion of abbreviated tokens. *In the Twitter corpus, non-standard word separators were counted excluding handles and hashtags. †Based on annotations of 300 labels per corpus.

datasets the majority) of headers and data values contain non-standard capitalisation. Such practices obviously reduce the performance of NLP models that rely on consistent capitalisation, such as named entity recognition, even if on the whole capitals remain useful predictors of certain phenomena.

**Abbreviations and acronyms** are pervasively used in the LoD. Table 2 shows that in header tokens their proportion is 5–30% while in data values it is 4–8%. In other kinds of corpora, this number is much lower;[5] only the query corpus comes close with 3.8%, although half of those cases were two-letter US state codes. In NLP tasks such as lemmatisation or WSD, abbreviations and acronyms typically behave as out-of-vocabulary words, and need to be expanded. In the LoD, however, due to their high frequency and inherently unbound nature (any word or expression can be abbreviated), simple dictionary, pattern, or supervised-learning-based expansion techniques are not likely to provide high recall. Recent efforts have used vector-space-based techniques in order to provide higher recall and robustness (Ciosici and Assent, 2018). Our hypothesis, to be verified in future work, is that in order for such solutions to be efficient on structured datasets, specific LoD-based word vector models would need to be trained.

[5]For tweets, we did not count handles and hashtags, nor the ubiquitous leading *RT* (retweet) acronym. Note that abbreviations and acronyms in tweets are mostly typical of spoken language and social media practices, e.g. *OMG* or *LOL*. Such constructs are completely absent from the LoD.

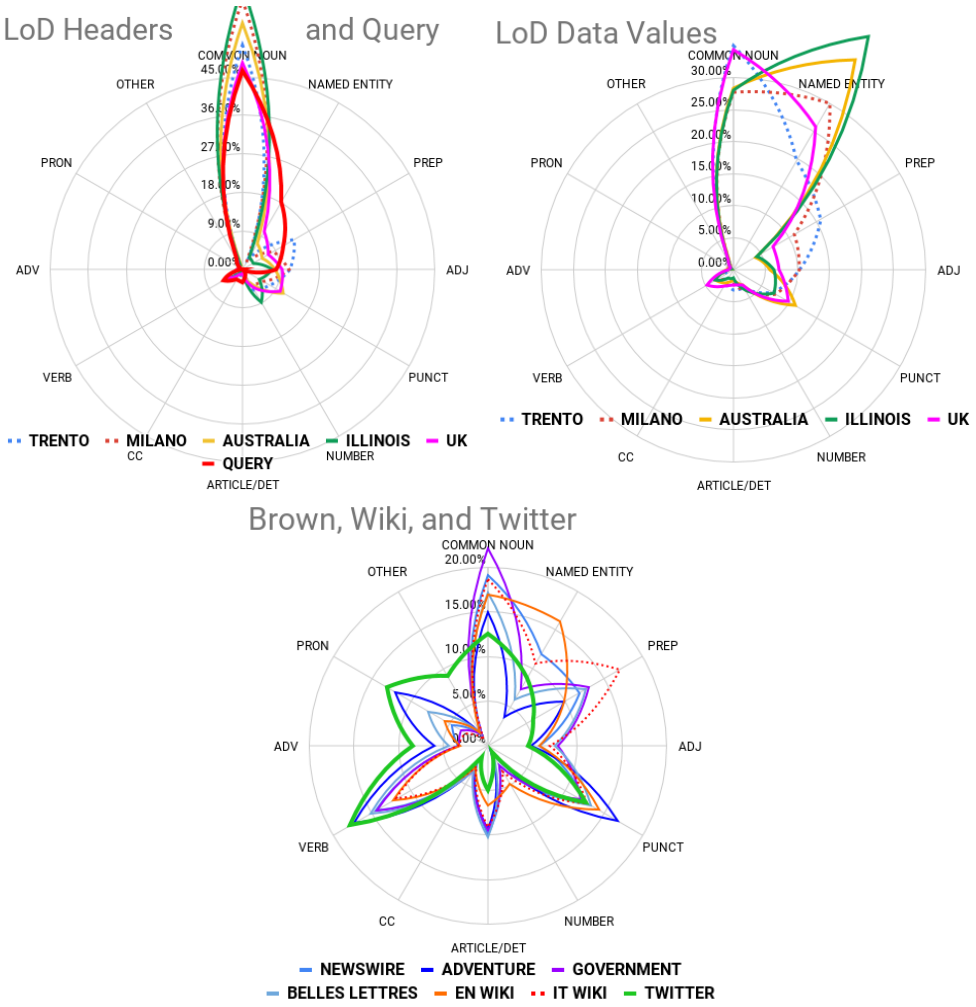


Figure 2: Comparison of part-of-speech distributions across corpora.

## 5 Parts of Speech

We analysed and compared the distribution of PoS and named entity categories in each annotated corpus shown in Table 1. The reason for this analysis is threefold: firstly, the PoS distribution of a corpus can be considered as a 'fingerprint' that is revelatory of its grammatical similarity or difference to other corpora. Secondly, it provides a basis for the subsequent analysis of syntactic constructs. And thirdly, the annotations can be used to train NLP components (e.g. PoS and NER taggers) dedicated to the LoD.

For cross-comparability across the Penn Treebank tagset (used for all English open data datasets), the Italian tagset, the Brown Corpus, and the Twitter tags, we mapped all tagsets to a coarse-grained tagset of the following 12 categories: *common noun, named entity, adjective, verb, adverb, pronoun, preposition, article or determiner, coordinating conjunction, punctuation, number*, and *other*. Figure 2 shows the distribution of coarse PoS tags for all corpora, that we synthesise here.

**LoD vs 'standard' text and tweets.** We take as reference representing 'standard' text the four genre-specific subsets of the Brown corpus: *newswire, adventure, government documents*, and *belles lettres*. As observed in Figure 2 (chart to the right), the four genres are characterised by very similar patterns on the diagram, showing only minor differences in PoS distribution. The Wiki corpora are also similar, with fewer pronouns while more numbers and named entities, as it is expected from encyclopaedic text. The patterns of the LoD (right and middle), in contrast, are fundamentally different, marked by an abundance of nouns and the absence of pronouns (0.3–0.6% as opposed to 3–12% of standard text) and adverbs (0.3–1.5% as opposed to 3–6%). The LoD also has much fewer verbs (1–4% as opposed to 12–17%),

articles and determiners (0.2–3% as opposed to 7–10%), and prepositions (in English: 4–7% as opposed to 10–13%, in Italian: 8–16% as opposed to 17%).[6] As these differences are much more pronounced than cross-genre variations in the Brown corpus, the Language of Data cannot be considered as merely another *genre* of written language. Tweets, while also short labels, are even more dissimilar, having the largest proportion of verbs, adverbs, and pronouns among our corpora. These findings point in the direction that parsing short data labels is likely to be a very different problem with respect to both short social media messages and 'standard' text.

**LoD headers vs data values.** Both the five *header* and the five *data value* corpora show a high degree of internal uniformity. Headers are characterised by a major dominance of common nouns (48–66% as opposed to 15–22% in standard text) and a relatively high proportion of adjectives (7–11% as opposed to 5–8%). Common nouns also dominate data values, although to a lesser extent (28–35%). Values, however, are characterised by a very large number of named entities (20–42%), reflecting the fact that the typical function of structured data is to describe objects of the real world. Names do appear in headers as well, although much more rarely (2–10%).

**LoD vs queries.** The part-of-speech distribution of the query corpus is remarkably similar to the LoD corpora, with a very low numbers of adverbs (0.5%), pronouns (0.9%), articles (3%), and verbs (5%). In terms of the proportion of common nouns (46%) and names (18%), it falls between the LoD header and value corpora. A closer inspection, however, also reveals differences. Firstly, queries generally lack punctuation signs (0.3%) while they play a major role in the LoD (5–11%). Secondly, fine-grained verb tags reveal that, in all LoD datasets, the most frequent verb form is the past participle (45–89% within headers, 28–59% within data values), well known often to play a non-finite, adjectival role. In the query corpus, most verbs are in indicative present (60%) while past participles are less frequent (11%). And thirdly, queries were found to contain a high proportion of prepositions, similarly to standard text (11%).

## 6 Syntax

Our analysis of the syntax of the LoD is motivated by both descriptive and practical goals: even if full syntactic parsing is becoming less common in state-of-the-art NLP, word context windows of varying length play a major role in the training of most machine learning models (from word embeddings to semantic taggers), the contents of which depend on the syntactic structure of the text.

Instead of an onerous full syntactic annotation, we based our analysis on a three-step classification of labels. First, we applied a coarse-grained classification into: (1) fully grammatical sentences that follow standard syntax; (2) sub-sentence phrases that still follow standard syntax; and (3) non-standard syntax. What to consider as 'standard syntax' was left to the judgment of the annotator as to whether the phrase in question would feel grammatically out of place within standard written media, such as a book or a magazine.[7] As a second step, a manual analysis of the non-standard labels has identified a set of recurrent syntactic constructs: *concatenation*, *compression*, *non-standard word order*, *interpretational clues*, and *abstract expressions*. We classified all such labels with respect to these inductive categories. The frequency of these constructs in each corpus is reported in the right-hand side of Table 3. Finally, we linked the constructs to the high-level linguistic categories of *disjunctive grammar*, *block language*, and *formal writing*, that we quantify in the left-hand side of Table 3. We also computed statistics on the proportion of labels containing named entities (middle of Table 3).

**Compression,** that we classify among the devices of block language, consists of the omission of certain grammatical elements, such as function words or verbs. This style is well known from news headlines, e.g. *'Officer [who was] fired for refusing to kill [a] bear wins [a] legal battle.'* In the LoD, the same principle is pervasively used, more in Italian where function words are more frequent (5–29% of all labels, such as: *'acquisto [al] bordo'*), *'imprese [che] prevedono assunzioni'*) but to a lesser extent also in English (1–5%: *'operate [an] uninsured vehicle'*, *'[the] year [the mine was] opened'*). In the last case,

[6] Italian has an inherently higher proportion of prepositions: where English would use a possessive or a noun adjunct (*director's cut*, *fish knife*), Italian uses prepositions (*versione del direttore*, *coltello da pesce*).

[7] When unsure or in corner cases, the annotator was asked to classify the phrase as standard.

| Corpus | Full | DisjGr | BlkLng | Name | Formal | Concat | Compr | Order | Clue | Abstr |
|---|---|---|---|---|---|---|---|---|---|---|
| Trento headers | 0% | 100% | 39.0% | 4.5% | 4.7% | 22.2% | 16.8% | 0% | 2.7% | 2.0% |
| Milano headers | 0% | 100% | 40.4% | 4.5% | 3.1% | 11.4% | 29.0% | 1.1% | 2.0% | 0% |
| Aus headers | 1.3% | 98.7% | 17.2% | 13.7% | 12.7% | 12.9% | 4.3% | 1.4% | 10.2% | 1.2% |
| Illinois headers | 0.3% | 99.7% | 27.8% | 7.0% | 9.1% | 25.1% | 2.7% | 1.7% | 0.6% | 6.8% |
| UK headers | 1.0% | 99.0% | 37.8% | 21.7% | 10.0% | 29.7% | 8.2% | 1.0% | 7.3% | 1.7% |
| Trento values | 1.3% | 98.7% | 30.3% | 58.1% | 2.5% | 6.3% | 24.0% | 1.0% | 1.3% | 0.3% |
| Milano values | 3.7% | 96.3% | 8.1% | 48.6% | 3.3% | 2.9% | 5.3% | 1.1% | 1.0% | 1.1% |
| Aus values | 1.3% | 98.7% | 7.2% | 58.6% | 1.8% | 4.3% | 2.9% | 0.7% | 0.9% | 0.1% |
| Illinois values | 2.3% | 97.7% | 2.9% | 59.3% | 3.0% | 1.6% | 1.3% | 1.1% | 0.8% | 1.1% |
| UK values | 2.2% | 97.4% | 11.6% | 50.8% | 3.4% | 6.4% | 5.2% | 0.8% | 2.4% | 0.2% |
| Query | 6.8% | 93.8% | 24.4% | 12.8% | 1.4% | 19.6% | 4.8% | 1.4% | 0% | 0% |
| Twitter* | 49.6% | 17.2% | 6.4% | †12.8% | 0% | 0.4% | 5.6% | 0.4% | 0% | 0% |

Table 3: Syntactic features: proportion of labels containing full sentences, disjunctive grammar, block language, named entities, formal labels, concatenation, compression, word order change, interpretational clues, and abstract expressions. *Based on a hand-annotated 500-tweet subset. †Not counting @*handles*.

the context of the dataset, that describes mines, provides the necessary interpretation. Another form of compression, very frequent in English, consists of converting one or more phrases into noun adjuncts, such as *'death country'* instead of *'country of death'*, in order to spare the preposition.

**Concatenation** of nouns and noun phrases without any grammatical marker (e.g. coordinating conjunction or comma) is the second major device of block language. For example, the label *'Cooroy Mary River Road Timber Bridge Rehabilitation'* is a drastically shortened and rearranged form of the sentence *'rehabilitation of the timber bridge on Mary River Road in Cooroy.'* Concatenation is also used for enumeration (*'ristorante bar pizzeria'*). It is an equally frequent form of non-standard syntax in both languages (11–30% of headers, 2–6% of values).

**Non-standard word order** is characteristic of formal or administrative style, and is often motivated by sorting needs. It manifests itself in inverted names (*'Fantozzi, Ugo'*, *'Agriculture, Department of'*) and postpositive adjectives (*'investigation non-specific'*, *'population growth, annual'*). Typically 1–2% of labels contain such constructs.

**Interpretational clues** are frequent in headers (1–10%) but appear also in values (1–2%). They consist of a short instruction at the end of the label, often delimited by punctuation, that serves as a clue for filling in or interpreting the data value. It most often provides a unit of measure (*'Duration (s)'*, *'Expenditure $'*) or scope (*'Total (2016)'*), but it may provide also provide other kinds of clarifications (*'Dog - non std'*).

**Abstract expressions** express their content in a formulaic manner, such as '*Income* $<= 133\%$' or '*H2O_DATE*.' The correct interpretation of such labels usually requires domain knowledge. 0–7% of all labels were found to contain such expressions.

In synthesis, our findings confirm the LoD almost exclusively to follow disjunctive grammar (97–100% of labels). The overwhelming presence of minor or non-finite sentences is further attested by a low frequency of verbs (1–4% in headers, 2–5% in values) and the high proportion of non-finite verbs among these (65–93% of all verbs inside headers, 53–72% inside values). Most disjunctive sentences consist of independent noun phrases, as proven by the large proportion of nouns (see Section 5). Block language (consisting of compression and concatenation) appears in 17–40% of headers and 3–30% of values. The syntax of formal writing is found in 3–13% of headers and 2–3% of values. Finally, as shown by the *Name* column in Table 3, named entities dominate the LoD, with 5–22% of all headers and 49–59% of all data labels containing at least one name.

Once again, among our control corpora, only the query corpus shows similar behaviour, in particular to the *header* corpora. Noun phrases dominate search queries as well. The kind of non-standard syntax employed within queries, however, is mostly limited to concatenation, a phenomenon that may contribute to the common perception of queries to be 'non-grammatical' (Wang et al., 2014). Tweets, in contrast,

show a very different profile: standard grammatical sentences are common (50%) as well as syntactic structures specific to spoken language (28%). Disjunctive grammar is much more rare (17%), and block language devices are mostly limited to compression (6%), typically eliminating sentence-initial pronouns (*'Just came home'*). In summary, while tweets and the LoD do share some common (and computationally challenging) syntactic structures, they belong to two completely different language registers.

## 7 Experiments on NLP Tasks

Our experiments have evaluated the performance of three classic low-level NLP tasks over the LoD: tokenisation, PoS tagging, and NER. These three tasks are often used in knowledge and information extraction, fundamental for semantic data integration and analytics over structured data (Bella et al., 2020). They are typically solved through supervised machine learning trained on annotated 'standard' text. Our goals were (1) to test the performance of state-of-the-art tools, trained on standard text, over LoD header and value datasets; (2) to evaluate the *usefulness* hypothesis: whether the same tools adapted to the LoD (e.g. with models trained on LoD datasets) would perform better, and to what extent; and (3) to obtain further empirical insight into the uniqueness of the LoD, especially with respect to search queries that we have found to be somewhat similar grammatically to structured data labels.

Our approach was to re-train state-of-the-art tools using our LoD corpora. For cross-comparability, we mapped our annotations (the POS and NER tags) to those used by these tools. Our test corpora were the *Australian headers* and *values* datasets, as well as the annotated *TREC* query subcorpus. Our training corpus, in turn, was the fusion of all English-language LoD datasets, obviously excluding the dataset used for testing.

**Tokenisation.** Our analysis in section 4 showed the high frequency of non-standard word separation in data headers, and to a lesser extent in values. We compared three solutions: (1) a classic supervised learning-based approach using the OpenNLP tokeniser tool; (2) a simple rule-based (regular expression) tokeniser that we developed specifically for the LoD; and (3) the OpenNLP tokeniser re-trained on our LoD corpora. The results in Table 4 show that, as expected, standard tokenisation struggles with LoD headers (80.5%) while with 94.5% it performs better on values (which is still not impressive: state-of-the-art performance on standard text is in the 98–99% range). Our simple rule-based tokeniser deals remarkably well with headers but slightly worse on values. The LoD-trained tokeniser provides the best, consistent performance (96.5–96.7%). The query corpus, in turn, proves to be easy to tokenise, with the standard and the LoD-based tool achieving the same score (99.4%). With respect to tokenisation, queries clearly prove to be a very different (much easier) problem than LoD labels.

**PoS tagging.** Over standard English text, PoS tagging is a largely solved problem with state-of-the-art results around 95%. Over the LoD, PoS tagging is expected to be harder: labels are very short, their orthography is different (e.g. inconsistent capitalisation makes the recognition of proper nouns harder), and the distribution of part of speech categories is also different. We compared a classic OpenNLP maximum entropy tagger, the state-of-the-art Stanford POS tagger, and the OpenNLP tagger retrained on the LoD. Table 4 shows that the standard tools perform weakly over the LoD (60–65% of accuracy on headers, 71–73% on values), due to the issues mentioned above. Training over the LoD results in major improvements (20–25% on headers and 8–10% on data), despite our trained model being 10 times smaller (!) than those of OpenNLP and Stanford. Interestingly, the advantage of the LoD-based training fades on the query corpus, where all three tools provide similar performance. If sections 4–6 found the grammar of data and queries to be similar in certain aspects, this does not translate into the transferability of trained models. This empirical result points at the language of queries being distinct from the LoD.

**Named Entity Recognition.** The high frequency of names in the LoD (see Table 3) highlights NER as a major subtask of information extraction from structured data. We experimented with two architectures: a maximum-entropy-based OpenNLP[8] name finder, and the state-of-the-art BERT-NER tool[9]

[8] `http://opennlp.apache.org`

[9] `https://github.com/kamalkraj/BERT-NER`

| Evaluation corpus | Tokeniser (F1) | | | POS Tagger (accuracy) | | | NER (F1) | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **ONLP** | **Rule** | **ON/LoD** | **ONLP** | **Stanfd** | **ON/LoD** | **ONLP** | **ON/LoD** | **BN*** | **BN/LoD** |
| LoD Headers | 80.5% | 95.6% | **96.5%** | 60.5% | 65.0% | **85.9%** | 35.0% | 50.8% | 35.0%* | **67.4%** |
| LoD Values | 94.5% | 92.9% | **96.7%** | 71.0% | 72.7% | **80.6%** | 30.8% | 36.7% | 44.7%* | **60.2%** |
| TREC Queries | **99.4%** | 98.3% | **99.4%** | 73.1% | **74.2%** | 72.1% | 4.3% | 4.0% | 5.3%* | **23.3%** |

Table 4: Performance of three NLP tasks over three evaluation corpora. ONLP: OpenNLP default model (maximum entropy); Rule: rule-based tokenizer developed for the LoD; ON/LoD: OpenNLP model trained on the LoD; Stanfd: Stanford POS tagger; BN: BERT-NER; BN/LoD: BERT-NER finetuned with the LoD corpus. *BN scores were computed excluding date entities due to lack of training data.

that performs supervised finetuning on top of a pretrained BERT language model, trained on the English CoNLL 2003 NER corpus (about 200k tokens, almost four times our LoD training data size). We retrained OpenNLP (with default settings) and finetuned BERT (three training epochs, with the *bert-base-cased* model) using our LoD NER corpus, again reserving *Australia headers* and *values* for testing. For cross-comparability, we only evaluated over person, location, organization, and date tags (we mapped our *GPE* tag to *location*). Table 4 shows overall results: while NER performance was generally weak, its heterogeneity was remarkable. Standard OpenNLP provided the worst scores (35% and 31% of F-measure for headers and values, resp.). The default BERT-NER managed to improve results over data values by 14%. The more impressive improvements were, however, obtained by training over the LoD, with the finetuned BERT-NER model achieving 67.4% over headers and 60.2% on values. While these are still low scores from a practical point of view, they are in a different ballpark with respect to the standard models. Note that BERT-NER was only finetuned to the LoD, while the underlying BERT language model, computed over standard text, was left unchanged. Future work on re-training BERT (or other similar architectures) on unsupervised structured data may raise these scores even further. Finally, let us note the overall bad performance of all models over query strings (even if the result of our finetuned BERT model does stand out). As in the case of the tokeniser and the PoS tagger, NER models trained on the LoD do not transfer well to search queries, once again hinting at major divergences between the two kinds of text.

## 8 Conclusions and Future Work

Despite the diversity of structured datasets used in our study, we observed *uniform* tendencies in the grammatical phenomena observed. At the same time, significant differences were revealed between the grammar of attribute names and values: among others, the latter tend to be longer, contain more named entities, and are closer in grammar to standard text. In terms of the *uniqueness* of the LoD, major differences were revealed in comparison to standard language on all grammatical levels examined. Only the query corpus showed similarity to the LoD (especially to attribute names), hinting at the transferability of NLP models across the two kinds of corpora: a hypothesis eventually disproved by NLP experiments.

We examined the *usefulness* of our results on a first set of NLP tasks. The results are encouraging as they show that training on LoD corpora in itself leads to major improvements in precision and recall. At the same time, we consider our models as merely the first step in the direction of developing NLP tools dedicated to the LoD. A promising research direction involves the creation and use of unsupervised word embeddings and language models from large LoD corpora, as a basis for subsequent fine-tuning to specific tasks. Another future work concerns the use of constraints imposed by the data structure itself on the interpretation of the labels. The combination of data-structure-based and corpus-based methods for label disambiguation, informed by the grammar of the LoD, is a direction yet to be investigated.

**Acknowledgement.** This paper and the underlying research were supported by the European Union's H2020 Research and Innovation programme under grant agreement no. 826106, project *InteropEHRate*.

## References


Aliaksandr Autayeu, Fausto Giunchiglia, and Pierre Andrews. 2010. Lightweight parsing of classifications into lightweight ontologies. In *International Conference on Theory and Practice of Digital Libraries*, pages 327–339. Springer.

Gábor Bella, Alessio Zamboni, and Fausto Giunchiglia. 2016. Domain-based sense disambiguation in multilingual structured data. In *The Diversity Workshop at the European Conference on Artificial Intelligence (ECAI)*.

Gabor Bella, Fausto Giunchiglia, and Fiona McNeill. 2017. Language and domain aware lightweight ontology matching. *Journal of Web Semantics*, 43:1–17.

Gábor Bella, Liz Elliot, Subhashis Das, Stephen Pavis, Ettore Turra, David Robertson, and Fausto Giunchiglia. 2020. Cross-border medical research using multi-layered and distributed knowledge. In *Proceedings of the 24th European Conference on Artificial Intelligence (ECAI)*, pages 2956–2963.

Cristina Bosco, Vincenzo Lombardo, Leonardo Lesmo, and Vassallo Daniela. 2000. Building a treebank for italian: a data-driven annotation schema. In *LREC 2000*, pages 99–105. ELDA.

Ben Carterette, Virgiliu Pavlu, Hui Fang, and Evangelos Kanoulas. 2009. Million query track 2009 overview. In *TREC*.

Manuel Ciosici and Ira Assent. 2018. Abbreviation expander-a web-based system for easy reading of technical documents. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 1–4.

Vivian Cook. 2013. The language of the street. *Applied Linguistics Review*, 4(1):43–81.

Leon Derczynski, Alan Ritter, Sam Clark, and Kalina Bontcheva. 2013. Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 198–206.

Geoffrey N Leech. 1966. *English in advertising: A linguistic study of advertising in Great Britain*. Longmans.

Xiao Li. 2010. Understanding the semantic structure of noun phrase queries. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1337–1345. Association for Computational Linguistics.

Samuel Reese, Gemma Boleda, Montse Cuadros, Lluís Padró, and German Rigau. 2010. Wikicorpus: A word-sense disambiguated multilingual Wikipedia corpus. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Beatrice Santorini. 1990. Part-of-speech tagging guidelines for the penn treebank project (3rd revision). *Technical Reports (CIS)*, page 570.

Ge Song, Yunming Ye, Xiaolin Du, Xiaohui Huang, and Shifu Bie. 2014. Short text classification: A survey. *Journal of multimedia*, 9(5):635.

Heinrich Straumann. 1935. *Newspaper headlines: A study in linguistic method*. London, Allen.

Andrea Tagarelli, Mario Longo, and Sergio Greco. 2009. Word sense disambiguation for xml structure feature generation. In *European Semantic Web Conference*, pages 143–157. Springer.

Joe Tekli. 2016. An overview on xml semantic disambiguation from unstructured text to semi-structured data: Background, applications, and ongoing challenges. *IEEE Transactions on Knowledge and Data Engineering*, 28(6):1383–1407.

Zhongyuan Wang, Haixun Wang, and Zhirui Hu. 2014. Head, modifier, and constraint detection in short texts. In *2014 IEEE 30th International Conference on Data Engineering*, pages 280–291. IEEE.