V4Design

Visual and textual content re-purposing FOR(4) architecture, Design and virtual reality games

H2020-779962

D3.1 Empirical study of textual data related to visual content

Dissemination level:	Public
Contractual date of delivery:	Month 10, 31/10/2018
Actual date of delivery:	Month 11, 2/11/2018
Work Package:	WP3: Visual and textual content analysis
Task:	T3.1: Compilation of texts relevant to visual data
Type:	Report
Approval Status:	Final version
Version:	3.0
Number of pages:	47
Filename:	D3.1_EmpiricalStudyTextualData_20181102_v3.0.pdf

Abstract

This deliverable reports on the findings of the empirical study of the materials compiled in the initial part of the V4Design project. It outlines the types of textual material that are to be analysed in the framework of the different use cases, along with quantitative and qualitative assessments of the contents for the purposes of the linguistic analysis (T3.2, T3.3, T3.4) and summarization (T5.4) modules. The study will serve as a basis for the definition of the WP3 Language Analysis and WP5 Summarization modules to be reported on month 16.

The information in this document reflects only the author’s views and the European Community is not liable for any use that may be made of the information contained therein. The information in this document is provided as is and no guarantee or warranty is given that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability.

co-funded by the European Union

History

Version	Date	Reason	Revised by
0.1	25/09/2018	ToC creation	Simon Mille
0.2	17/10/2018	Descriptions of experiments	Simon Mille, Gerard Casamayor
0.3	19/10/2018	First results of analyses	Simon Mille
0.4	22/10/2018	First draft circulated to WP3-related partners for comments	Simon Mille
0.5	23/10/2018	Description of texts	Beatriz Fisas
0.6	23/10/2018	Concept extraction	Alexander Shvets
1.0	24/10/2018	Pre-final draft sent for internal review	Gerard Casamayor, Simon Mille
1.1	25/10/2018	Internal review	Jolan Wuyts (EF)
2.0	26/10/2018	Preparation of the final draft	Simon Mille, Alexander Shvets
3.0	01/11/2018	Second round of comments and preparation of an updated final draft	Simon Mille

Author list

Organization	Name	Contact Information
UPF	Simon Mille	simon.mille@upf.edu
UPF	Gerard Casamayor	gerard.casamayor@upf.edu
UPF	Beatriz Fisas	beatriz.fisas@upf.edu
UPF	Alexander Shvets	alexander.shvets@upf.edu

Executive Summary

In this document, we present the empirical study of the textual material compiled in the first phase of the V4Design project. This study consists in a qualitative and quantitative analysis of the different textual genres relevant to the V4Design Pilot Use Cases to be implemented in the project. The objective of the study is to assess the specificities of each genre in terms of linguistic phenomena.

We first briefly describe the six contemplated genres: captions, video descriptions, Wikipedia pages, news and magazine articles, blog posts and tweets. We provide examples and a first analysis of the quality of the texts, in which we pinpoint the possible issues that would cause the Linguistic Analysis pipeline to perform poorly. We then present the features used for the quantitative assessment of the linguistic structures from the perspective of the morpho-syntactic analysis and concept extraction modules: word-based features, grammatical category-based features, and syntactic dependency-based features. The empirical study is performed by extracting these features from the V4Design material compiled up to date. A similar analysis is carried out from the perspective of the semantic analysis and summarization modules, using more abstract features such as average polysemy, sense embedding coverage and meaning frequencies. For the latter, we also evaluate the current coverage of several available tools. Finally, we briefly assess the dynamic aspects of the textual contents and establish a preliminary scenario which consists in looking for opinion trends in professional reviewers. This study shows differences of quality and contents across the six different genres and will serve as a basis for the definition of the WP3 Language Analysis (morpho-syntax, semantics, and concepts) and WP5 Summarization modules to be reported on month 16 and of the WP5 Dynamic 3D Objects Retrieval module to be reported on month 26.

Abbreviations and Acronyms

HTML	Hypertext Markup Language
JSON	JavaScript Object Notation
KB	Knowledge Base
NE	Named Entity
NER	Named Entity Recognition
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
PoS	Part of speech
PTB	Penn Treebank
PUC	Pilot Use Case
SEW	Semantically Enriched Wikipedia

Table of Contents

1 INTRODUCTION

This deliverable summarizes the outcome of the empirical study of the textual material that has been compiled during the first phase of the V4Design project. The texts of the different genres have been analysed with respect to the different phenomena of the linguistic mode (morpho-syntactic, lexical and semantic). The goal is to capture these phenomena in order to have a complete view of the specificities of each genre and serve as a basis for the definition of the WP3 Language Analysis pipeline (T3.2, T3.3, T3.4), the first version of which will be described in D3.3 on month 16. The WP5 summarization component (T5.4, T5.5) is closely linked with the WP3 modules in that it will be fed with the semantically analysed structures. Thus, in this deliverable, we also provide an analysis of the textual material from the perspective of the summarization techniques that UPF plans to use in V4Design (to be described in D5.2 on month 16), together with a preliminary evaluation of the coverage of several off-the-shelf tools. Finally, the WP5 3D object retrieval component (T5.3) will also rely on the structures produced by the Language Analysis pipeline, using them to infer how the opinions of people on, e.g., a building, evolve with time; hence, we briefly report on the initial study of dynamic textual contents. Since the project focused on English during the first year, this empirical study only covers texts in English. We expect the same differences between the genres across the project languages (English, German, Spanish and Greek).

For our analysis, we've drawn upon the initial requirements and use case scenarios descriptions of D7.2, as well as on ongoing discussions with the user partners towards further clarifications and refinements about the desired contents; relevant insights have been afforded also within the ongoing work within WP6 for the mapping of the D7.2 user requirements into respective technical requirements. The agreed Pilot Use Cases (PUCs) are the following:

- PUC1: Architectural design, related to existing or historical buildings and sites and their environments
- PUC2: Architectural design, related to artworks, historic or stylistic elements
- PUC3: Design of virtual environments, related to TV series and VR video games
- PUC4: Design of virtual environments, related to actual news for VR (re) living the date

For each PUC, visual assets will be input in the V4Design platform and the relevant assets will serve as basis for building 3D models to be used in the 3D tools of the architects and game designers. The task of the WP3 Language Analysis component in V4Design is twofold:

1. extract knowledge from the textual material associated to the visual assets in order to identify as precisely as possible the objects, buildings, monuments, etc. present in the images and videos. The textual sources in this case are mainly **image, painting and video captions**, **video descriptions** and **museum tweets**.
2. extract knowledge from associated textual sources (articles, critics, etc.) to discover prominent aesthetic or technical features of the asset. The textual sources for this task are mainly **blog posts**, **specialized magazine articles** and **Wikipedia pages**.

As the project progresses, we will assess the relevance of the different genres for the purposes of the four PUCs and the other tasks; not all the genres may eventually turn out to be equally relevant or to contain information that can be extracted with sufficient quality.

The rest of the document is structured as follows: Section 2 briefly describes the six genres and provides examples, a first analysis of the quality of the texts made available by the content providers, as well as an initial study on the dynamic aspects of the blog and specialized magazine articles. Section 3 presents the features used for the quantitative assessment of the linguistic structures from the perspective of the T3.2 and T3.3 modules (morpho-syntactic analysis and concept extraction) and an analysis of the numbers obtained on the V4Design material. In Section 4, a similar analysis is carried out from the perspective of the T3.4 (semantic analysis) and T5.4/T5.5 (summarization) modules. Finally, Section 5 summarizes the results of the analyses.

2 QUALITATIVE ASSESSMENT OF THE TEXTUAL MATERIAL

In Section 2.1 , we present the six genres agreed to be studied by the consortium: captions, video descriptions, Wikipedia pages, news and magazine articles, blog posts and tweets; for each genre, a short description and examples are provided. Section 2.2 contains a qualitative analysis of the material released by the content providers. Finally, Section 2.3 introduces the opposition between static and dynamic textual data and its implications in V4Design.

2.1 Textual genres

2.1.1 Captions of images, paintings and videos

Description

Image and video captions are the titles associated to each visual asset. The captions describe the contents of an image or summarize those of a video. Captions are very short, and contain mostly nominal groups, instead of full sentences. In other words, the syntactic structure of the captions will play a less important role than the identification of the concepts and named entities they may contain.

Sources

Contents provided by Europeana Foundation, and obtained from scraped Wikipedia images.

Information to be extracted

The main objective for the caption analysis is to identify what object(s)/building(s)/monument(s) can be found in the associated visual asset, and to link it/them with existing database entries (e.g. DBpedia).

Sample cleaned captions

(1) *The gymnasium.*[1]

(2) *The mountain-top stadium at Delphi, far above the temples/theatre below.*[1]

(3) *St John the Baptist Preaching.*[2]

(4) *Josua Defeating Amalek.*[3]

(5) *The castle is one of the many historical buildings that make up the Sintra Cultural Landscape, a UNESCO World Heritage Site, popular with tourists to Portugal.*[4]

2.1.2 Descriptions of videos

Description

Video descriptions are texts with full sentences that describe more into details the contents of a video and possibly some background information about its authors and topics. They can also contain technical data about the filming and editing process. Their size is usually limited to one or a few paragraphs. The styles of the descriptions vary from one author to another.

[1] From the page en.wikipedia.org/wiki/Delphi

[2] http://nationalmuseumse.iiifhosting.com/iiif/ac7716a034230a93c55a34757a1c603516d8ed0a645e015ca2b9fe5cfcfdc2e7/full/full/0/default.jpg

[3] https://www.europeana.eu/api/v2/thumbnail-by-url.json?uri=http%3A%2F%2Fnationalmuseumse.iiifhosting.com%2Fiiif%2F836e5fb47b7ccecfab6d55baecbf9d83ad4ffa4c950c763449916558f201ff4e%2Ffull%2Ffull%2F0%2Fdefault.jpg&size=LARGE&type=IMAGE

[4] From the page http://en.wikipedia.org/wiki/Castle_of_the_Moors

Although most of the texts have well-structured full sentences, some of them are a sequence of short phrases, each one describing a scene in the video. The last sentence may be a rhetorical question, leaving the answer in the air, a poetic ending highlighting the emotions aroused in the video, a summary of the main questions raised by the story or a re-addressing to other similar videos.

Sources

Contents provided by ArtFilms and SLRS MULTIMEDIA AB.

Information to be extracted

The main objective for the video description analysis is to identify what object(s)/building(s)/monument(s) can be found in the associated videos, and to link it/them with existing database entries (e.g. DBpedia), and to a lesser extent some aesthetic features whenever the level of detail of the video description allows for it, and technical details about how the video was shot (frames per second, resolution, etc.).

Sample video descriptions

(1) *Filmed exploration of the importance of calligraphy in Japanese culture, with a survey of the varieties of writing styles, the place of writing as art in everyday life, examples of calligraphy filmed in kabuki theater, a popular restaurant, the shops that sell the inks and papers used in fine writing, monks executing sutras in the old temple of Nara, etc. Writing styles of movie marquis, religious writing, etc. Includes demonstrations by celebrated calligraphers, the teaching of ideograms in grade school, an all-Japan calligraphy contest, writing on designer dresses and on ceramics, odd techniques, ancient styles, the relationship of poetry to the medium that expresses it. With an explanation of how - unlike in the linearity of western writing - an ideogram can combine different meanings and references to present a reader with one new concept in one picture.*[5]

(2) *A film about modern Japanese architecture, its roots in the Japanese tradition and its impact on the Nordic building-tradition. Winding its way through visions of the future, traditions, nature, concrete, gardens and high-tech, Kochuu tells us how contemporary Japanese architects strive to unite the ways of modern man with the old philosophies in astounding constructions. Interviews with and works by Japanese architects Tadad Ando, Kisho Kurokawa, Toyo Ito and Kazuo Shinohara and Scandinavian architects Sverre Fehn, Kristian Gullichsen and Juhani Pallasmaa.*[6]

2.1.3 Wikipedia pages

Description

Wikipedia pages are texts with full sentences that describe a particular place, concept, entity, etc. Their size can vary between one paragraph and several pages, depending on the popularity and the contributions of the concerned topic. Wikipedia articles follow precise guidelines that encourage the contributors to relate facts only, prohibiting subjective opinions and terms, and to be as concise and clear as possible, resulting in a rather homogeneous style across the Wikipedia.

Sources

Contents scraped by CERTH.

Information to be extracted

[5] https://www.artfilms.com.au/item/shodo-japanese-calligraphy-in-daily-life

[6] http://www.solarisfilm.se/portfolio/kochuu/

The main objective for the analysis of Wikipedia articles is to obtain additional information on the object/building/monument identified on the corresponding images or videos, such as prominent aesthetic features, architect, date of construction, etc.

Sample Wikipedia article

(1) *Great Barford Castle, later known as Luton Castle was a 12th-century castle in the town of Luton, in the county of Bedfordshire, England (grid reference TL09062082). 12th Century Castle It was a timber motte-and-bailey structure built in 1139 and demolished in 1154 following a truce. 13th Century Castle. Another castle on a different site was built in 1221 but was destroyed around 1224 or 1225. Earthworks and associated bailey survived but were removed. An excavation was done in 2002, revealing a steep ditch. The site is now home to Matalan, a discount store. Nothing visible remains of either castle.*[7]

2.1.4 News and magazines articles

Description

News and magazines articles are also texts with full sentences, usually written by journalists that have extensive training in writing. The style and level of objectivity of the language is variable, but the size of the articles is usually formatted to be between half a page and two pages.

Sources

For the purposes of this study, we browsed manually some websites recommended by user partners and others found through search engines. We selected a few sources based on the fact that their contents are free of access and the articles contain information that seems particularly relevant for the purposes of the project.

- https://www.architectural-review.com/buildings
- http://www.uncubemagazine.com
- https://www.dezeen.com/
- https://www.architecturaldigest.com/architecture-design/architecture

Information to be extracted

The main objective for the analysis of new and magazine articles is to obtain additional information on the object/building/monument identified on the corresponding images or videos, in particular the prominent aesthetic features.

Sample article

(1) *Alvernia Studios is a bizarre, futuristic wonderland, home to the largest modern film studio in Poland. Set in the countryside 18 kilometres from Krakow airport, it was designed and built in 2002 by media entrepreneur Stanislaw Tyczyński, the founder of Poland's first private radio station, RMF FM. Inspired by the art of H. R. Giger, the distinctly alien-looking 13,000-square-metre facility is comprised of an interconnected web of metallic domes that house sound stages, film scoring studios, and high-tech facilities for visual effects. Described by those in the industry as "a film within a film", every inch of the facility reflects the spirit of a sci-fi movie set: thick tubes of glass corridors stretch between the shiny metal domes – while the interior, decorated from floor to ceiling with futuristic motifs, evokes the feeling of walking through a monumental spaceship. Described by those in the industry as "a film within a film". (Photo: Alexander Belenkiy) Technical features at Alvernia Studios include door handles operated by fingerprint readers and a dome containing the world's largest shade-less spherical blue screen. In the*

[7] http://en.wikipedia.org/wiki/Gannock_Castle

recording studio, a special device opens and closes like a flower, changing the acoustic parameters of the space. The main structure of reinforced concrete, polyurethane foam, and compressed PVC air balloons, is resistant to rain, frost, and wind and probably recycling as well, which means these space-colony style "domes from the future" may outlive us all.[8]

2.1.5 Blogs and forums

Description

Blogs and forums contain contributions ranging from one sentence to full texts. This is where more subjective contents can be found (in general, the first person ("I") is widely used); bloggers usually write well but this is not a constant.

Sources

As for magazine articles, we browsed manually some blogs and forums found through search engines and recommended by users. Many architecture blogs focus on practical aspect: how to be a good designer, how to sell your work, etc. We selected a few sources based on the fact that their contents are free of access and in which the posts focus on the design and architecture aspects.

- https://youngarchitect.com/2016/10/14/six-inspiring-young-architect-bloggers/#unique-identifier1
- https://lukearehart.com/
- http://www.talkitect.com/
- https://www.evolvingarchitect.com/blog/
- http://angryarchi.com/
- http://www.theaspiringarchitect.com/
- https://designobserver.com/
- http://www.bldgblog.com/
- https://unhappyhipsters.com/
- http://supercolossal.ch/
- http://www.bldgblog.com/
- http://www.an-architecture.com/
- http://continuity.msa.ac.uk/
- https://archinect.com/forum/

Information to be extracted

The main objective for the analysis of blog and forum posts is to obtain additional information on the object/building/monument identified on the corresponding images or videos, in particular the prominent aesthetic features and the context in which the object/building/monument has been designed/built.

Sample part of a blog post

(1) *This question was fun for me to think about. I know I am going to leave some out because there is some really interesting work going on right now in the city. The city has changed dramatically since I came here in 1996. We have a whole lot more talent here now. I think there is also a greater awareness of design in the world, which has elevated the Portland discussion as well.*
One of my favorite buildings is the Bank of California building. I feel like it is so completely figured out and understood. All the way from how the exterior detail works and fits into the whole system down to the

[8] http://www.uncubemagazine.com/blog/12594667

> *ceiling detail. There is not a corner in that building that isn't a result of the whole idea. I have a huge amount of respect for that. You can go inside and it feels simple but it's very complex in the way that all of the forms are proportionally gathered together. The material palette is perfectly executed and figured out. In addition, the scale component is carefully thought about and understood. I'm incredibly impressed with that building.*[9]

2.1.6 Tweets

Description

The tweets in the V4Design dataset can be communications issued by museums, or captions of images that are museum objects. They contain a wide range of linguistic constructions and style. Most tweets are written in English, although the dataset also includes tweets in Japanese, Spanish, Catalan, German, Dutch, Flemish and French.

The tweets in the V4Design dataset can be classified into two well differentiated groups:

1. tweets posted by Museum Bots that tweet a random object image with a short text from an art collection once to four times a day (18000 tweets)
2. tweets generated by Museums or other organizations related to art or history (22000 tweets).

Their structure, content and purpose are quite different.

The tweets posted by Museum Bots include a short text, two links and graphical material (generally a photograph). The short text has the title of a piece of art that can be seen in the attached picture and the name of the artist. Some of them also add the date and place where the object was created. The first link addresses to the museum's catalogue where detailed information about the artist, the period to which the object belongs, the material, its dimensions, the technique, etc., can be found. This information is displayed as a text or in many cases in a table format. Sometimes, the museum's webpage offers a summary about contextual information (see the tweets from Victoria and Albert Museum[10]) or more explanations about the techniques used by the artist (see the tweets from Tate Collection[11]). The second address in the tweet links to the tweet itself.

While the first group of tweets is randomly generated by autonomous software, the second group of tweets is completely different regarding the content, structure and purpose of the tweets. They are the Museum's communication channel with their potential visitors and their main purpose is to attract the public towards them. This collection of tweets includes invitations to activities in the museum, such as talks, exhibitions, etc., calls to competitions, information about opening hours, but also retweets, and answers to users asking for specific information. They include a short text, a link to the tweet itself and some of them, but not all, have graphic material (photos or videos).

Source

Contents scraped by CERTH on Twitter.

[9] https://chatterbox.typepad.com/portlandarchitecture/firm_architect_profiles/

[10] https://twitter.com/V_and_A?ref_src=twsrc%5Egoogle%7Ctwcamp%5Eserp%7Ctwgr%5Eauthor

[11] https://twitter.com/Tate?ref_src=twsrc%5Egoogle%7Ctwcamp%5Eserp%7Ctwgr%5Eauthor

Information to be extracted

The main objective for the tweet analysis is to identify what object(s)/building(s)/monument(s) can be found in the corresponding visual asset, and to link it/them with existing database entries (e.g. DBpedia).

Sample tweets

The original tweets can be seen by following the provided links.

(1) *Late 18th century English pocket globe with a star map in the casing, useful for travel as it was just 3 inches* wide http://t.co/NifkIIFHHw

(2) *A Picture Book Mirror of Various Occupations (Wakoku shoshoku ezukushi) https://t.co/InM1O9xpBP https://t.co/pHk8HDTEW7*

(3) *Archaeologists could be about to discover exactly where Henry I is buried at #Reading Abbey: https://t.co/WrJufbrKYt https://t.co/JMpDsqkINC*

(4) *RT @_smartify: Inspiring read by @Culture24 on frugal innovation \u0026 battling engrained organisational culture. https://t.co/sK18ZFnQlu #arts…*

(5) *Blue bird sitting on a plum blossom tree, by Chinese painter Ren Xiong (1823-1857) #FirstDayOfSpring #SpringEquinox… https://t.co/luGqGy3Bq5*

2.2 Analysis of the quality of the collected datasets

The qualitative assessment has only been performed using the “official” V4Design dataset, that is, we did not study qualitatively the texts retrieved on our own since we don’t know under which form they would reach the analysis pipeline at this point.

Subjective ratings on a 10-point scale are provided in order to reflect the level of “cleanliness” of the texts, that is, how easily they can be analysed by the UPF Language Analysis pipeline. Details on how the final rating was obtained are provided: three levels of penalty are applied, corresponding to 1.5-, 1- and 0.5-point penalty.

2.2.1 Europeana Foundation (EF) Dataset

EF provided about 16,000 image and painting captions. The contents are available in the “en” field of a JSON file. The quality of the texts is excellent for the image and some important issues have been identified for the painting captions.

Image captions

The images of the EF dataset cover both scenarios of PUC1, PUC2 and PUC4. The qualitative assessment resulted in a rating of 10/10 for these texts. The captions contain mostly noun keywords (e.g. “statue”), which can sometimes be more complex noun groups (“illuminated manuscripts”, “Black-and-white prints”. These keywords can be processed without major issues, but their informativeness will have to be demonstrated during the project, and it is not certain that relevant information can be extracted from them.

Painting captions

The painting captions have been evaluated at 7/10. The detected issues are the following:

- **Penalty -1**: even though there are language fields, we have no information about the language in which the text is actually written ("en": ["Leben und Werk des Malers Moritz Oppenheim:, Vortrag gehalten am 19. M\u00e4rz 1966 vor dem Hanauer Geschichtsverein /, von Dr. jur. Rudolf M. Heilbrunn."); there can even be several languages in the same field;

- **Penalty -1**: there are some ill-encoded characters ("F\u00fcr");
- **Penalty -0.5**: there are some unusual punctuation combinations (" :, ", ". /,", ". :, ", "--");
- **Penalty -0.5**: there are some unexpected characters such as "[" in the middle of words.

Other comments:

- In many cases, there are keywords only;
- There can be noun phrases or full sentences;
- There can be more than one sentence in the caption;
- Sometimes, only the final punctuation sign is present.

The first issue could be fixed using an off-the-shelf language detection module. The other issues could be fixed either by the content provider or through a pre-processing on the UPF side.

2.2.2 SLRS MULTIMEDIA AB (SLRS) Dataset

Three movie descriptions were provided, all with excellent text quality. The textual contents have been scraped from the "Description" filed of the following HTML pages:

- https://vimeo.com/ondemand/greatexpectations
- https://vimeo.com/ondemand/kochuu
- https://vimeo.com/ondemand/microtopia

Video descriptions

The texts have been assigned 9.5/10. The only detected issue is the unexpected presence of HTML markers (**Penalty -0.5**), which requires an extra processing of the pages. It has been noted that the descriptions can contain information about the architecture-related contents of the video (in particular, building and architect names), but that they also contain an important amount of more generic information not necessarily related with the visual contents of the videos.

2.2.3 ArtFilms (AF) Dataset

ArtFilms provided about 25 movie descriptions stored in the "Description" column of a spreadsheet, which generally have a good quality.

Video descriptions

The descriptions are detailed and can contain a lot of relevant information; they have been assessed at 8/10, with the following issues:

- **Penalty -1.5**: there seems to be a lot of encoding issues ("today&rsquo;s");
- **Penalty -0.5**: In the middle of the text, there can be technical details or metadata about the video that don't follow a particular syntax.

The first issue is very problematic for an analysis pipeline, but only requires additional processing to replace the utf-8 codes by the corresponding characters. The second issue does not happen much, but when it does it may be difficult to fix, since the metadata or technical information has to be identified automatically in order to be removed.

In addition, for each description, more related textual material can be found in the spreadsheet:

- There is a lot of metadata for each movie;
- There is a "summary" field that seems to be a condensed version of the description.

2.2.4 Scraped content

CERTH scraped textual data from Wikipedia and Twitter. The compiled dataset contained about 700 Wikipedia image captions, 300 Wikipedia articles and 40,000 tweets related to architecture and museum objects. All data has been released in the JSON format, and the textual material was found under the “caption”, “text_content” and “text” fields respectively. There is a clear difference of quality between the Wikipedia contents, which are usually very clean, and the tweet material.

Wikipedia image captions

The Wikipedia image captions are of very good quality (9/10); they can be one or more nominal group(s) or full sentence(s). Frequent use of colons, parentheses, non-English characters and unclear abbreviations are reported. The two main issues are the following:

- **Penalty -0.5**: there are escaped characters (\", etc.);
- **Penalty -0.5**: there are possible references in the caption (e.g., "[41]", "(top)", "(bottom)").

One other minor problem which could sometimes affect analysis results is that final punctuations are not always present in the captions. These issues are quite frequent but only would require an additional filtering in order to clean the texts.

Wikipedia articles

The Wikipedia webpages are also of very good quality (8.5/10); the detected issues are the following:

- **Penalty -0.5**: there are explicit line breaks (\n) and other escaped characters (\", etc.);
- **Penalty -0.5**: there are possible references in the body of the text (e.g., "[18]");
- **Penalty -0.5**: there are possible quotes, with escaped chunks ("[...]").

As it is the case for the captions, some relatively simple filtering would be enough to remove the superfluous contents

Twitter posts

Twitter posts are quite different from the rest of the textual material in that they contain a lot of special characters, links, and can cover a wide range of contents in a wide range of styles and syntactic constructions. Frequent non-English characters and unclear abbreviations are reported. These posts received an assessment of 5/10, with the following issues:

- **Penalty -1**: there are frequent chunks that look like cryptic metadata or references ("N45, Type1", "R405", "E100", "\u0026#39");
- **Penalty -1**: some captions are not related with the content ("A brief note: the @metmuseum just made all their open images CC0 licensed. Free to use even w/o attribution!")

- **Penalty -1**: there is no information about the language in which the text is written;
- **Penalty -0.5**: there are a large amount of the "@" character, which indicates mentions to other accounts;
- **Penalty -0.5**: the posts contain many links that are syntactically unrelated to the textual contents;
- **Penalty -0.5**: there are escaped characters (\", etc.);
- **Penalty -0.5**: there are possible references in the body of the text (e.g., "[41]", "(top)", "(bottom)").

Most of the issues have been discussed in the paragraphs above. The main problem with tweets is the in-text metadata and references, and the unrelated captions: solving these issues may require heavy processing, without ensuring a good quality of the final texts.

2.3 Static VS dynamic textual data

As seen with the examples provided in the previous subsections, the knowledge contained in the different textual sources can either be seen as *static* or *dynamic*. On the one hand, captions, descriptions, museum tweets and Wikipedia pages contain *static* knowledge, in the sense that this knowledge corresponds to objective facts, which tend to be more stable in time. On the other hand, blog posts and specialized articles may contain more *dynamic* knowledge, since they may reflect the author's (or a cited source's) opinions, and these can evolve in time or from one individual to another.

From a linguistic perspective, there is no difference between texts that contain static or dynamic knowledge; they will be processed in the same way by the Language Analysis tools. However, in the context of the semantics-aware modules developed in the framework of T5.1, T5.2 and T5.3, the processing will be different: static contents will be added to the Knowledge Base, whereas dynamic contents will possibly update previously added knowledge. Another objective of this empirical study is thus to assess to what extent and under which form(s) dynamic textual data will be encountered in the project. In the following, we briefly present our conclusions.

Before starting, let us state again what we consider as dynamic contents: these are contents for which there is a possibility that they change with time. In this section, we thus focus on opinions at different points in time and consider both contradictory and converging opinions.

The main source of point of views is people's ratings on travel-related websites, but…

… it is not clear to what extent this kind of contents would be usable (reasons detailed below). We examined, e.g., Google and TripAdvisor reviews, which contain a lot of comments for a wide variety of buildings. For instance, the TripAdvisor review page of Barcelona's Sagrada Familia contains more than 65,000 reviews in English. Consider for example the following 5 reviews emitted at different points in time:

1. "When you get there, **the sight is pretty impressive with the different facades** of the church, the most famous facade is the nativity one which Gaudi designed."[12]

[12] https://www.tripadvisor.es/ShowUserReviews-g187497-d190166-r547525648-Basilica_of_the_Sagrada_Familia-Barcelona_Catalonia.html

2. "The lines are long, so **take pictures of the outside of the chapel and just enjoy the view** and take it all in while you wait."[13]
3. "It had a **great view of the building**."[14]
4. "The location is not that great, in a poor designed street and **you can't have good view of the exterior of the building**."[15]
5. "The site is magnificent from what we saw but such a shame that there's **such a limited view from the perimeter**."[16]

It is not usual that one person leaves an opinion twice on such a website, so the dynamism of textual contents could be addressed here as the evolution of trends over time. By analyzing a lot of reviews at different times, it would be possible to detect trends of opinions (e.g., 2 years ago people were complaining about the view around the church, but now it doesn't seem to be an issue anymore) and thus see if a particular aspect of a building is seen differently at different times.

However, there are several problems with general public's opinions. First, they are extremely subjective, and the question of how to assess the value/informativeness of their contents would have to be addressed. Second, the quality of the language can vary tremendously from one review to another; reviews are often written hastily, and the contents of a review are more important than its form. This could be a problem for the analysis tools, which perform poorly when confronted to typos, grammatical mistakes and colloquial writing (e.g. "an hour+ queue" instead of "a queue of more than an hour"). Finally, reviews on these websites are not trivial to scrape; doubts were expressed by the consortium with respect to being able to retrieve such content (T2.1).

In any case, these simple examples show a very important aspect that the Language Analysis pipeline will have to address (for all types of contents and all languages): in order to allow the next modules to detect a contradictory or converging view, it is clear that some *linguistic generalization* is needed, both at the level of the word and of the relations between the words. The parts of the reviews in bold above talk about the same thing (the sight of the exterior of the building), but in very different ways: "impressive sight with the different façades", "take pictures of the outside of the chapel and just enjoy the view ", "great view of the building", "such a limited view from the perimeter", "you can't have a good view of the exterior of the building". Linguistic Analysis cannot infer that a façade or a perimeter are outside, and that "façade", "exterior of the building" "outside the chapel", "building" and "perimeter" mean roughly the same in this context, but it must be able to cope with the lexical variety found in the sentences: in this case, "sight" and "view" must be identified as the same concept, "impressive", "enjoy" and "great" must all be related to a positive evaluation, "limited" and "can't have a good view" to a negative evaluation. On the level of the dependencies between the words (see Section 3.1.3), the negation in "you can't have a

[13] https://www.tripadvisor.es/ShowUserReviews-g187497-d190166-r7591928-Basilica_of_the_Sagrada_Familia-Barcelona_Catalonia.html

[14] https://www.tripadvisor.es/ShowUserReviews-g187497-d190166-r19489715-Basilica_of_the_Sagrada_Familia-Barcelona_Catalonia.html

[15] https://www.tripadvisor.es/ShowUserReviews-g187497-d190166-r547549030-Basilica_of_the_Sagrada_Familia-Barcelona_Catalonia.html

[16] https://www.tripadvisor.es/ShowUserReviews-g187497-d190166-r624422427-Basilica_of_the_Sagrada_Familia-Barcelona_Catalonia.html

good view" must be related with the "good", although it is syntactically related to the modal "can". These few examples show a small sample of the variety of constructions in English, but in V4Design more languages will be addressed (German, Spanish and Greek). That is, this generalization of the concepts and relations between the words will have to be performed both within and across languages (T3.4).

The processing of dynamic textual data may be best addressed as comparisons of different experts' points of views at different moments in time.

In none of the architecture and design blogs or specialized magazine websites we found any case of someone providing explicitly a new opinion to be contrasted with their former opinion. The closest we found are website sections such as the "Revisited" section of architecturenow.co.nz,[17] which compile old reviews of emblematic buildings that have a link with a current event, together with some more recent considerations.[18] This does not mean that such posts or articles do not exist, but it at least indicates that they are scarce, and that we will probably not be able to use such material on a large scale for the purposes of V4Design.

Following the lines of what has been suggested above for social reviews, we thus turned to opinions of different authors at different points in times. Consider for instance three sentences taken from reviews on specialized websites:

1. "… the basilica as an **artificially inflated** space lacking in soul."[19] (Oct. 2014)
2. "The newly completed nave of the Sagrada Família looks **disturbingly fake**."[20] (June 2015)
3. "A view of the nave looking towards the altar. Notice the **organic shaped** columns."[21] (December 2017)

These three sentences, by different authors, talk about the nave of the Sagrada Família. The first two point out some fakeness in its design, whereas the third one emphasizes the organic aspect of the columns that are part of it. The variety of linguistic structures used for depicting comparable aspects shows again the need for concept and relation generalization, as seen with social reviews above. Experts' opinions are not as numerous as non-experts' ones on social websites, but they have the advantages of being generally well-written, containing informed points of view, and being easily retrieved from the web.

Finally, note that what is reported in this section is preliminary and the work carried out in T5.3 may come to different conclusions with respect to what kind of textual material is needed. The modules developed in T3.2, T3.3 and T3.4 will eventually process the textual materials as needed by T5.3.

[17] https://architecturenow.co.nz/search/?q=revisited

[18] https://architecturenow.co.nz/articles/happy-20th-anniversary-sky-tower/

[19] http://www.bbc.com/culture/story/20141014-gaudi-unfinished-business

[20] https://www.nybooks.com/articles/2015/06/25/antoni-gaudis-great-temple/

[21] https://www.sah.org/publications-and-research/fellowship-reports/brooks-fellow-reports/brooks-report-detail/sah-blog/2017/12/13/la-sagrada-fam%C3%ADlia-a-testament-of-architectural-ingenuity

3 QUANTITATIVE ASSESSMENT OF THE TEXTUAL MATERIAL FOR MORPHO-SYNTACTIC ANALYSIS AND CONCEPT EXTRACTION

For the quantitative assessment, we processed the compiled material with off-the-shelf Natural Language Processing (NLP) tools: NLTK,[22] Stanford CoreNLP (Manning *et al.*, 2014), and Bohnet and Nivre's (2012) dependency parser. We studied the linguistic features in prevision of their possible impact on the superficial analysis tools, which target primarily grammatical category assignment (i.e., *part-of-speech tagging*), grammatical function assignment (i.e., dependency parsing), and concept extraction. The tools used for processing the texts described in Section 2.2 are the ones that will be used as baselines for the English pipeline (a full description of the V4Design analysis pipeline will be provided in D3.3).

In the following, Section 3.1 describes the linguistic features used in the empirical study, Section 3.2 gives a brief overview of how the texts were pre-processed, Section 3.3 presents a detailed analysis of the features for morpho-syntactic parsing using off-the-shelf tools, and Section 3.4 presents an analysis of the texts with a different method, using the first version of our V4Design concept extraction tool.

3.1 Linguistic features for morpho-syntactic analysis

For this empirical study, we use word-based features, part-of-speech- and character-based features, and dependency-based features.

3.1.1 Word-based features

Words per sentence

Simply expresses the average number of words per sentence for each genre. The more the words per sentence, the more complex the syntactic analysis can get.

The next features of this category are computed using dictionaries that contain words of specific types; the used dictionaries are publicly available resources, listed as footnotes of the titles of the next paragraphs. The dictionary-based features and the resources used to compute them, are the following:

Acronym usage[23]

This feature computes the ratio of tokens in a text that are acronyms. An acronym is a word or name formed as an abbreviation of the parts of a sentence, multiple word expression, or a word (e.g., Lysergic Acid Diethylamide is widely known as LSD). This feature can give an indication of whether specific Acronym dictionaries may be needed during the analysis.

Stopword usage[22]

This feature computes the ratio of stopwords in a text. Stopwords are common words that usually do not contribute significant meaning to the text. As a result, stopwords are considered to be of lesser importance in analysis processes. On the other side, a low amount of stopwords could mean that the author of a text is using a high amount of infrequent

[22] https://www.nltk.org/

[23] http://onlineslangdictionary.com/thesaurus/words+meaning+acronyms+(list+of).html

words, which could make the analysis more challenging. To implement this feature, the stopwords list provided by Python's Natural Language Toolkit (NLTK) has been used.

Polar word usage[24]

This feature measures the usage of positive and negative sentiment words. To determine which words are positive and which ones are negative, a sentiment analysis lexicon is used. The lexicon contains a list of words that belong to each category. The tendency of using positive and negative polarity words is in some cases directly related to the level of subjectivity with which a text can we written. For reasons that are beyond our control, we were only able to identify Positive words in this study.

Named entity usage[25]

For this feature, the ratio of named entities (names of persons, places, organizations) is computed. For this, the off-the-shelf Stanford Named Entity Recognition (NER) tool was run on the texts (Manning *et al.*, 2014). The NER ratio gives us an idea of the importance of the presence of persons, buildings, etc., in a text.

3.1.2 Part-of-speech- and character-based features

For this group of features, we computed the ratio of the different grammatical categories for each genre, and the ratio of some punctuation signs. The specific part-of-speech tagset used in the experiments is the set of the Penn Treebank Project (Marcus *et al.*, 1993). This choice is motivated by several reasons. Firstly, it is a precise, fine-grained tagset that does not only distinguish between basic part-of-speech tags (such as "noun", "verb", etc.), but also gives information about the specific type of category in question (indicating, for instance, that an adverb is comparative, in which tense a verb is, whether a noun is in singular/plural, or whether it is common or proper). The Penn Treebank tagset provides much more information than the basic tagsets frequently used in the literature. Moreover, this tagset is also used in many widely distributed NLP tools such as CoreNLP, openNLP or the Natural Language Toolkit[22] (NLTK). The tags and their description are displayed in Table 1. For each word in a text, the tagger outputs the corresponding part-of-speech tag. Using this information, we measure the frequency of each of these tags (dividing the number of occurrences of a particular part-of-speech tag by the total number of words in the text). To complement these fine-grained tag frequencies, the frequencies of basic part-of-speech categories (verbs, nouns, adverbs, adjectives, pronouns, determiners and conjunctions) are also computed. In addition, the usage ratios of superlative/comparative adjectives/adverbs as well as verbs in past and present tense (with respect to the total number of verbs) are computed.

Tag	Description
CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential there
FW	Foreign word
IN	Preposition or subordinating conjunction

[24] https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon

[25] https://stanfordnlp.github.io/CoreNLP/ner.html

JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PRP	Personal pronoun
PRP$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	to
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Wh-determiner
WP	Wh-pronoun
WP$	Possessive wh-pronoun
WRB	Wh-adverb

Table 1: Part-of-Speech tags and their meaning

Character-based features simply account for the use of the main punctuation signs: comma, colon, semi-colon, question and interrogation marks, quotes, parentheses.

3.1.3 Dependency-based features

For the dependency-based features, we computed the ratio of the different syntactic dependencies that hold between the words of a sentence, together with tree complexity measures. Syntactic dependency trees are unordered rooted trees that represent the syntactic structure of a sentence according to a specific grammar. Dependency trees are composed of sets of nodes which correspond to the words of the represented sentence and sets of arcs that connect the nodes via binary asymmetrical dependencies. Each word (except the root) can govern or be governed by another word. Robinson (1970) formulates the four basic axioms that a syntactic dependency structure must meet to be considered well-formed. These axioms are the following:

1. One and only one element (the root) is independent.
2. All other elements depend directly on some element.
3. No element depends directly on more than one other element.

4. If A depends directly on B and some element C intervenes between them (in the linear order of the string), then C depends directly on A or B or some other intervening element.

Syntactic dependencies depict syntactic properties of the sentence; for instance, the SBJ (‘subject’) dependency indicates in English that the governing verb takes its number and person from the dependent subject. The syntactic dependencies also indirectly encode some more semantic information, such as for instance the degree of importance of a dependent with respect to its governor. Most dependencies fall into one of the two following categories:

- Core dependency: indicates a strong link between two elements, as it is the case between a verb and its subject, a verb and its object(s) for instance.
- Non-core dependency: indicates a loose link between to elements (but a link anyways), such as in circumstantial groups with their governing verb,

These dependency relations provide useful information about the inner structure of the sentences: we can measure to what extent coordinate or subordinate clauses are being used in a genre, or if appositions, logical subjects in the passive voice, or the verb chains, for example, are prominent or not. Knowing which types of dependencies are present in a genre can be valuable in that it may influence the type of tool or linguistic formalism to use for the text analysis.

The particular set of dependencies that is used for this empirical study is a subset of the Penn Treebank Project’s dependency relation tagset (described in (Surdeanu et al., 2008)); see Table 2 for the complete list. Two types of dependencies have been removed:

- non-atomic dependencies: these are combined dependencies, that can for example indicate one syntactic function and the fact that there is an element missing (e.g. ADV-GAP). Due to their scarcity, they have a low informativeness.
- dependencies already reflected in other features: for instance, the IN dependency holds between a preposition and its complement (e.g. “of V4Design”), and since there is a part-of-speech tag that stands for “preposition”, the IN dependency does not bring any information that is not carried by the ratio of IN tags.

The retained dependencies are distributed in three groups:

1. Arguments, which are core dependents of verbs.
2. Adverbials, which are adverbials are non-core dependents of verbs;
3. Modifiers, which are core and non-core dependents of non-verbal elements;

The parses have been obtained by running Bohnet and Nivre’s (2012) joint parser. For each dependency, a feature is computed. To do so, we divide the number of times that the dependency is used in a text with the total number of sentences. These frequencies correspond to the mean number of the occurrences of each dependency relation per sentence. For example, a high number of “SUB” and “COORD” relations per sentence could indicate that the author tends to use coordinate and subordinate clauses frequently in his/her writings. We then divide this by the number of words per sentences in order to obtain the ratio of each dependency.

Tag	Description
ADV	General adverbial
AMOD	Modifier of adjective or adverbial
APPO	Apposition
BNF	Benefactive
COORD	Coordination
DEP	Unspecified dependency
DIR	Adverbial of direction
DTV	Dative complement in dative shift
LGS	Logical subject below a passive verb
LOC	Locative adverbial or nominal modifier
MNR	Adverbial of manner
NMOD	Modifier of nominal
OBJ	Object
OPRD	Predicative complement of raising/control verb
PRD	Predicative complement
PRP	Adverbial of purpose or reason
SBJ	Subject
TMP	Temporal adverbial or nominal modifier
VC	Verb chain (between auxiliary and verb)
VOC	Vocative

Table 2: Dependency labels and their meaning

To further characterize the writings of the authors, shape-based tree features are computed. Three different metrics are extracted from the trees: depth, width and ramification factor. Depth is defined as the maximum distance (in terms of nodes) between the root and a leaf node. Width is the maximum number of siblings in a level of a tree. Ramification factor is the mean number of children nodes per level. In the tree shown in Figure 1, the maximum width is 6 (below *buildings*), the maximum Depth is 14 (from *lived* to *Morris*) and the ramification factor is 4 (5+4+9+7+5+3+4+4+6+2+1+1+1 = 52 children / 13 levels). Each feature is calculated by dividing width, depth and the ramification factor respectively by the number of sentences in a text (which corresponds to the number of dependency trees). This is a way to see how complex the sentences are in a given text.

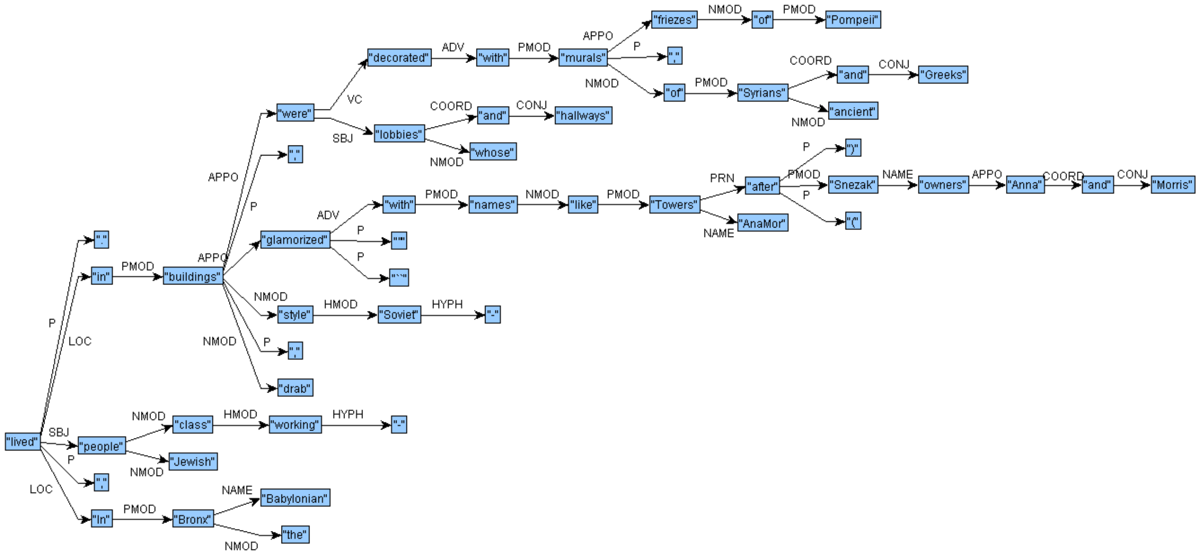


Figure 1: A sample dependency tree for the sentence *In the Babylonian Bronx , Jewish working - class people lived in drab , Soviet - style buildings ``glamorized'' with names like AnaMor Towers (after owners Anna and Morris Snezak), whose lobbies and hallways were decorated with murals of ancient Syrians and Greeks, friezes of Pompeii .*

3.2 Pre-processing of the data for the feature extraction

In order for the tools to be run without any issues the texts as described in Section 2.2 were cleaned with a simple pre-processing before being analysed.

3.2.1 Captions

1. removed captions that contained weird character sequences (\u...).
2. removed captions that are simple dates
3. removed duplicates

The final number of analysed captions is 3,944 (about 24,000 words), out the about 17,000 initially collected ones.

3.2.2 Descriptions

1. removed texts with meta-info that shouldn't be part of the description.
2. replaced manually character codes by the corresponding symbols.

We analysed the 23 video descriptions (about 7,000 words) that were made available at the time this deliverable was written.

3.2.3 Wikipedia pages

1. removed "See also", "References", "External links", "Bibliography", "Location".
2. removed all duplicated lines.

162 Wikipedia articles were analysed (about 103,000 words), that is, all the initially scraped articles.

3.2.4 Tweets

1. removed all links ("http...")
2. removed all incomplete tweets (ending with "... "or a backslash)
3. removed tweets that contained weird character sequences (\u...).
4. removed all tweets that start with a hashtag, an @, or that have just one word before (^[^\s]*\s*[@#].*$).

Almost half of the available tweets were analysed for this study: 18,929 (about 140,000 words) out of the 40,073 collected ones.

3.2.5 News articles, blogs, forums

No processing was applied; we analysed in total 22 articles (about 18,000 words) and 28 posts (about 13,000 words).

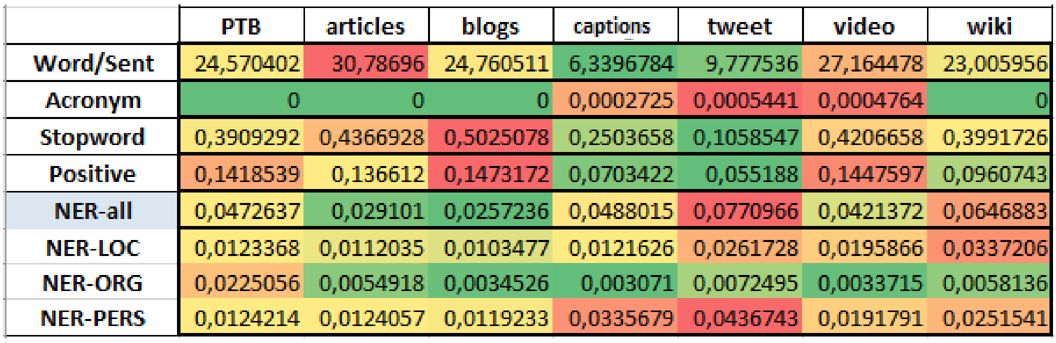
3.3 Quantitative assessment for morpho-syntactic analysis

In this section, we present the results of the ratio calculations for all the features. A three-color conditional formatting has been applied to the tables in order to make the visualization of the numbers easier: a red cell indicates a high value, whereas a green cell indicates a lower value and yellow intermediate ones. These "heatmaps" are to be interpreted within the thick boxes: for instance, in Table 3, the first five rows are independent heatmaps, and the last three correspond to one single heatmap.

Numbers for texts from the six different genres are reported, and contrasted to the Penn Treebank (PTB) numbers, a corpus of about 40,000 sentences (1,000,000 words) on which the dependency parser used in the project has been trained. The PTB corpus is the reference corpus used in English nowadays; it contains texts from the Wall Street journal, with articles about the stock market, theatre, books and movie reviews, and generic articles. Contrasting the linguistic features in all the genres with those of the PTB allows for determining which types of phenomena may or may not be an issue for the analysis of texts of the different genre.

3.3.1 Word-based features

Table 3 shows the ratio of words per sentence, acronyms, stopwords, positive words and Named Entities across all genres.



	PTB	articles	blogs	captions	tweet	video	wiki
Word/Sent	24,570402	30,78696	24,760511	6,3396784	9,777536	27,164478	23,005956
Acronym	0	0	0	0,0002725	0,0005441	0,0004764	0
Stopword	0,3909292	0,4366928	0,5025078	0,2503658	0,1058547	0,4206658	0,3991726
Positive	0,1418539	0,136612	0,1473172	0,0703422	0,055188	0,1447597	0,0960743
NER-all	0,0472637	0,029101	0,0257236	0,0488015	0,0770966	0,0421372	0,0646883
NER-LOC	0,0123368	0,0112035	0,0103477	0,0121626	0,0261728	0,0195866	0,0337206
NER-ORG	0,0225056	0,0054918	0,0034526	0,003071	0,0072495	0,0033715	0,0058136
NER-PERS	0,0124214	0,0124057	0,0119233	0,0335679	0,0436743	0,0191791	0,0251541

Table 3: Word-based features

Words per sentence

The average ratio of word per sentence is an indicator of the possible complexity of the sentences to analyse. It is about 25 in the PTB; it is similar in blogs, video descriptions and Wikipedia articles, and as expected, much lower in captions and tweets (6 and 10 respectively). Analysing short sentences is less challenging than analysing long ones, so from the perspective of the size, the existing PTB-based parser has no issues parsing these genres. In specialized articles, the ratio is slightly about 30 words per sentence that is on average 6 more words per sentence. The PTB contains about 12,000 sentences with 30 words or more, that is, one third of the total number of sentences in the corpus. Thus, the number of words in the specialized articles is not an issue for our tool, which is able to return well-formed trees in a reasonable time for much larger sentences.

Acronyms

Acronyms have not been detected in the PTB, but were identified in captions, tweets and video descriptions. This means that acronym identification will be needed at least for the analysis of these three genres, and that the dependency parsing tool will probably not help for this task. We thus plan to cover acronym identification as part of the concept extraction module, which will be applied before the parsing step.

Stopwords

Stopwords are words that do not contain meaning, as opposed to meaning-bearing words, which are the ones that will be mapped on the Knowledge Base (KB). The stopword ratio is around 0.40 in the PTB, and significantly lower for captions and tweets, which thus contain a higher proportion of meaningful units. In other words, captions and tweets tend to be denser in terms of meaning, even though they are generally rather short. In other words, being able to analyse them properly will be crucial in the project.

Positive words

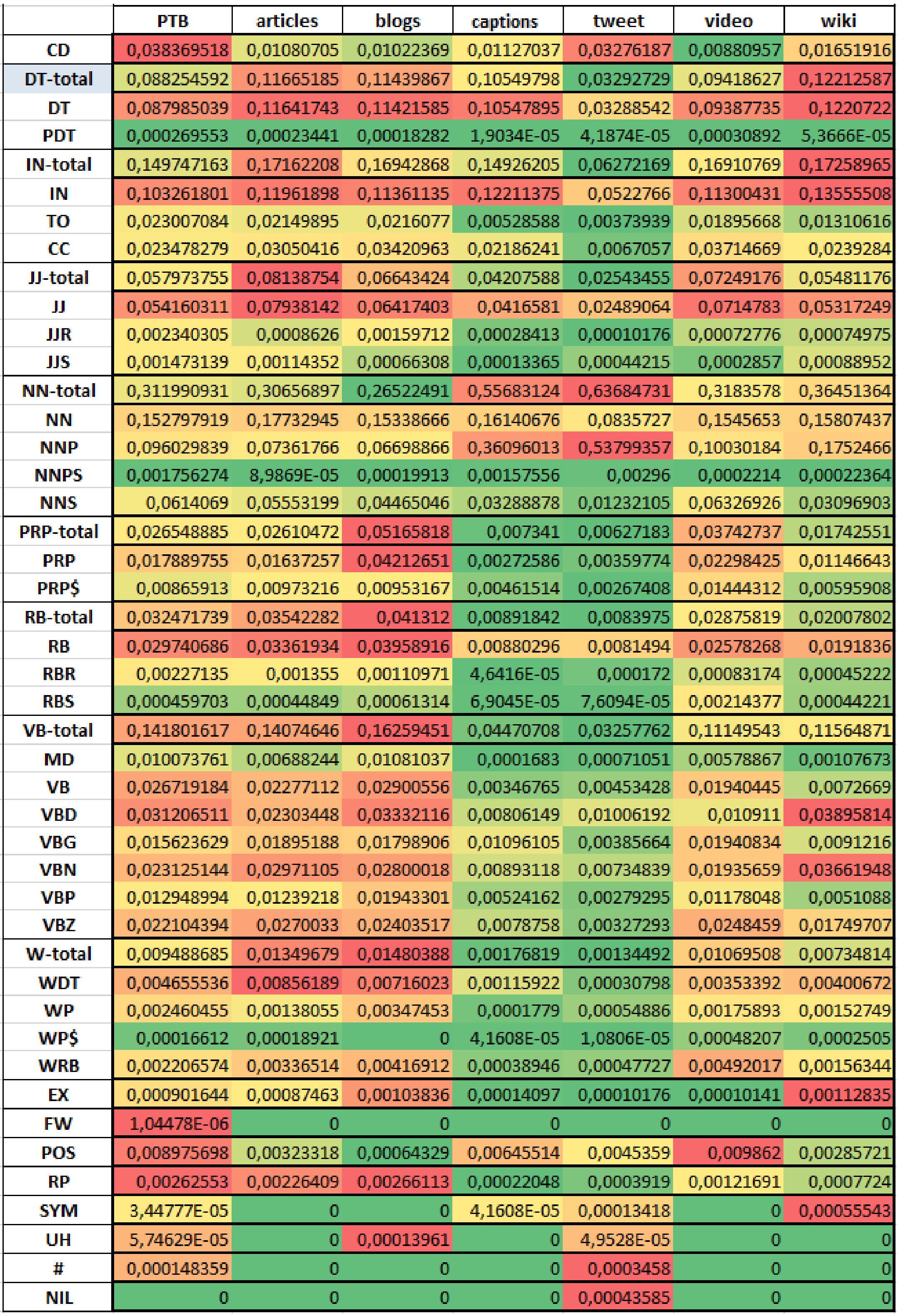
The ratio of positive words is an indicator of the subjectivity with which the text is written. Captions and Wikipedia pages are very descriptive and supposed to be objective; they exhibit rather low ratios of positive words (about half of what is found in PTB). At first sight, the fact that tweets have an even lower ratio may seem surprising, but this is due to the fact that the tweets in the V4Design dataset are mostly museum tweets, instead of people’s opinions. The features examined so far show that the tweet material is actually similar to captions in terms of linguistic contents; the features examined in the following sections confirm this statement.

Named Entities

The Named Entity (NE) ratio shows in which genres Named Entity Recognition is particularly important, in this case tweets and Wikipedia articles. Specialized articles and blog posts seem to contain less mentions of NE in general. Wikipedia articles contain a higher proportion of locations, while captions and tweets contain more mentions of persons. Mentions of organizations seem to be less relevant across all genres. As a result, the NER tools used in V4Design will have to focus particularly on persons and locations.

3.3.2 Part-of-speech- and character-based features

Table 4 and Table 5 respectively show the ratio of the different parts of speech and punctuation signs across all genres. Section 3.1.2 contains short descriptions of each tag.



	PTB	articles	blogs	captions	tweet	video	wiki
CD	0,038369518	0,01080705	0,01022369	0,01127037	0,03276187	0,00880957	0,01651916
DT-total	0,088254592	0,11665185	0,11439867	0,10549798	0,03292729	0,09418627	0,12212587
DT	0,087985039	0,11641743	0,11421585	0,10547895	0,03288542	0,09387735	0,1220722
PDT	0,000269553	0,00023441	0,00018282	1,9034E-05	4,1874E-05	0,00030892	5,3666E-05
IN-total	0,149747163	0,17162208	0,16942868	0,14926205	0,06272169	0,16910769	0,17258965
IN	0,103261801	0,11961898	0,11361135	0,12211375	0,0522766	0,11300431	0,13555508
TO	0,023007084	0,02149895	0,0216077	0,00528588	0,00373939	0,01895668	0,01310616
CC	0,023478279	0,03050416	0,03420963	0,02186241	0,0067057	0,03714669	0,0239284
JJ-total	0,057973755	0,08138754	0,06643424	0,04207588	0,02543455	0,07249176	0,05481176
JJ	0,054160311	0,07938142	0,06417403	0,0416581	0,02489064	0,0714783	0,05317249
JJR	0,002340305	0,0008626	0,00159712	0,00028413	0,00010176	0,00072776	0,00074975
JJS	0,001473139	0,00114352	0,00066308	0,00013365	0,00044215	0,0002857	0,00088952
NN-total	0,311990931	0,30656897	0,26522491	0,55683124	0,63684731	0,3183578	0,36451364
NN	0,152797919	0,17732945	0,15338666	0,16140676	0,0835727	0,1545653	0,15807437
NNP	0,096029839	0,07361766	0,06698866	0,36096013	0,53799357	0,10030184	0,1752466
NNPS	0,001756274	8,9869E-05	0,00019913	0,00157556	0,00296	0,0002214	0,00022364
NNS	0,0614069	0,05553199	0,04465046	0,03288878	0,01232105	0,06326926	0,03096903
PRP-total	0,026548885	0,02610472	0,05165818	0,007341	0,00627183	0,03742737	0,01742551
PRP	0,017889755	0,01637257	0,04212651	0,00272586	0,00359774	0,02298425	0,01146643
PRP$	0,00865913	0,00973216	0,00953167	0,00461514	0,00267408	0,01444312	0,00595908
RB-total	0,032471739	0,03542282	0,041312	0,00891842	0,0083975	0,02875819	0,02007802
RB	0,029740686	0,03361934	0,03958916	0,00880296	0,0081494	0,02578268	0,0191836
RBR	0,00227135	0,001355	0,00110971	4,6416E-05	0,000172	0,00083174	0,00045222
RBS	0,000459703	0,00044849	0,00061314	6,9045E-05	7,6094E-05	0,00214377	0,00044221
VB-total	0,141801617	0,14074646	0,16259451	0,04470708	0,03257762	0,11149543	0,11564871
MD	0,010073761	0,00688244	0,01081037	0,0001683	0,00071051	0,00578867	0,00107673
VB	0,026719184	0,02277112	0,02900556	0,00346765	0,00453428	0,01940445	0,0072669
VBD	0,031206511	0,02303448	0,03332116	0,00806149	0,01006192	0,010911	0,03895814
VBG	0,015623629	0,01895188	0,01798906	0,01096105	0,00385664	0,01940834	0,0091216
VBN	0,023125144	0,02971105	0,02800018	0,00893118	0,00734839	0,01935659	0,03661948
VBP	0,012948994	0,01239218	0,01943301	0,00524162	0,00279295	0,01178048	0,0051088
VBZ	0,022104394	0,0270033	0,02403517	0,0078758	0,00327293	0,0248459	0,01749707
W-total	0,009488685	0,01349679	0,01480388	0,00176819	0,00134492	0,01069508	0,00734814
WDT	0,004655536	0,00856189	0,00716023	0,00115922	0,00030798	0,00353392	0,00400672
WP	0,002460455	0,00138055	0,00347453	0,0001779	0,00054886	0,00175893	0,00152749
WP$	0,00016612	0,00018921	0	4,1608E-05	1,0806E-05	0,00048207	0,0002505
WRB	0,002206574	0,00336514	0,00416912	0,00038946	0,00047727	0,00492017	0,00156344
EX	0,000901644	0,00087463	0,00103836	0,00014097	0,00010176	0,00010141	0,00112835
FW	1,04478E-06	0	0	0	0	0	0
POS	0,008975698	0,00323318	0,00064329	0,00645514	0,0045359	0,009862	0,00285721
RP	0,00262553	0,00226409	0,00266113	0,00022048	0,0003919	0,00121691	0,0007724
SYM	3,44777E-05	0	0	4,1608E-05	0,00013418	0	0,00055543
UH	5,74629E-05	0	0,00013961	0	4,9528E-05	0	0
#	0,000148359	0	0	0	0,0003458	0	0
NIL	0	0	0	0	0,00043585	0	0

Table 4: PoS-based features

Cardinal Numbers (CD)

The PTB contains a lot of numbers, in particular due to the large number of articles related to stock market (share values, dates, etc.). Only in the tweet material the proportion of numbers is similar. This is due to the widespread use of dates associated to paintings and images. In general, across the genres, numeric dates are used a lot and will need to be given special attention during the analysis phase of the V4Design pipeline. Dates can usually be identified because they are between parentheses or after a preposition (e.g., *in 1815*).

Determiners (DT, PDT)

Determiners are part of the stopword list mentions in the previous section. They usually bring little information in terms of contents and tend to be used more in more formally written material, such as specialized articles, blog posts and Wikipedia articles. Note that the regular determiner ration (DT) is quite high in captions, even though the corresponding stopword ratio was low. This is an indicator that captions contain mostly nominal groups (see confirmation in Nouns below), and that nouns are generally not used in their bare form. Tweets are the genre with the lowest ratio of determiners, indicating a more concise way of writing, a consequence of the Twitter 140-character limit.

Prepositions and conjunctions (IN, TO, CC)

Prepositions and conjunctions are the other main contributors to the stopword count. They are an indicator of the syntactic complexity of the sentences: the more the prepositions and conjunctions, the more embedded groups in the sentence structure. This feature shows that again tweets seem to have a simpler structure than other genres, and that on the contrary, specialized and Wikipedia articles, blog posts and video descriptions exhibit more embeddedness below nouns or verbs. The ratio of infinitive markers *TO* indicates a lower proportion of infinitive verbs, a prominent feature in captions and tweets, which are mostly nominal groups. Interestingly, Wikipedia articles seem to also contain fewer infinitive verbs, which could be due to the limited use of bare forms and their inherent ambiguity (the verb arguments should be expressed explicitly in order to avoid interpretation errors).

Adjectives (JJ, JJR, JJS)

The frequent use of adjectives is an indicator of (i) the richness of the writing (professional writers tend to use more adjectives), and (ii) the subjectivity of the writing (adjectives are often used to refer to personal judgements). Specialized articles and blog posts are both written by trained writers, and often convey personal considerations. Video descriptions are also rich because they need to make the video attractive to the potential viewers. The identification of adjective-nous pairs will be of primary importance in V4Design, since most aesthetic features are conveyed through the use of adjectives. From this perspective, captions and tweets do not seem to contain a lot of aesthetic features.

Nouns (NN, NNP, NNPS, NNS)

In the PTB, the ratio of nouns is roughly one every three words, which can be taken as a reference for a typical text. Captions and tweets contain a much higher proportion of nouns (twice as much), which indicates a wide use of nominal groups, and in particular of singular proper nouns *NNP*. As seen in the previous section, many of these nouns refer to persons and locations. The Wikipedia article ratio is slightly above the PTB ratio, while the blog post ratio is slightly lower. In Wikipedia, it seems like this compensates a lower adverb and verb usage. For blog posts, it is compensated by a wider use of pronouns (see next paragraph).

Another interesting phenomenon is the difference between the use of common nouns *NN(S)* and proper nouns *NNP(S)*. In our reference corpus, the PTB, the proportion of common nouns is significantly higher than that of proper nouns (about twice as high). In specialized articles, blog posts and video descriptions, the same proportion is found, but in Wikipedia articles there are as many common and proper nouns, while in captions and tweets it is the other way around, with a larger number of proper nouns compared to common nouns (6 times as many in the tweets). Not all these proper nouns are Named Entities, and in addition to the need of powerful NER tools mentioned in the previous section, more generally proper noun identification and disambiguation will need to be addressed with particular care in order to extract meaningful information from these texts.

Pronouns (PRP, PRP$)

The use of pronouns is quite different according to the genre. In the PTB corpus and the specialized articles the proportion is similar, but it is higher in video descriptions (especially third person pronouns) and in blogs (especially first-person pronouns). On the contrary, there are significantly less pronouns in Wikipedia articles, again possibly due to the efforts in making the texts non-ambiguous, and the ratio is as expected very low in the nominal captions and tweets.

Adverbs (RB, RBR, RBS)

Adverb usage is related to adjective usage in that it depicts a certain richness and/or subjectivity in the writing. Blog posts exhibit a higher ratio of adverbs, while Wikipedia articles have a lower ratio. Adverbs are most of the time associated to verbs, hence the very low ratios for tweets and captions.

Verbs (MD, VB, VBD, VBG, VBN, VBP, VBZ)

Verbs link entities (mostly common and proper nouns) with one another, and are very important in the analysis of the meaning of texts. In the reference corpus, verbs are the third most frequent part of speech, and their ratio is about half of the noun ratio (0.14 VS 0.31). Again, the ratio of verbs in tweets and captions is very low, as expected in material that contains mostly nominal groups; however, it is not null because nouns groups often contain embedded verb groups as relative clauses. The ratio of verbs is quite similar in PTB, specialized articles and blog posts, but is significantly lower in video descriptions and Wikipedia articles. In these two genres, the ratio of verbs is even down to about a third of the ratio of nouns. Wikipedia articles favour nominal and prepositional constructions, whereas for video descriptions, adjective and pronouns are more prominent. In Wikipedia articles, past verbs *VBD* and past participles *VBN* are used in higher proportions, but the use of modal verbs *MD* -which are often used to convey a probability according to the speaker- and non-finite infinitive (*VB*) and gerund (*VBG*) are particularly low.

Interrogative words (WDT, WP, WP$, WRB)

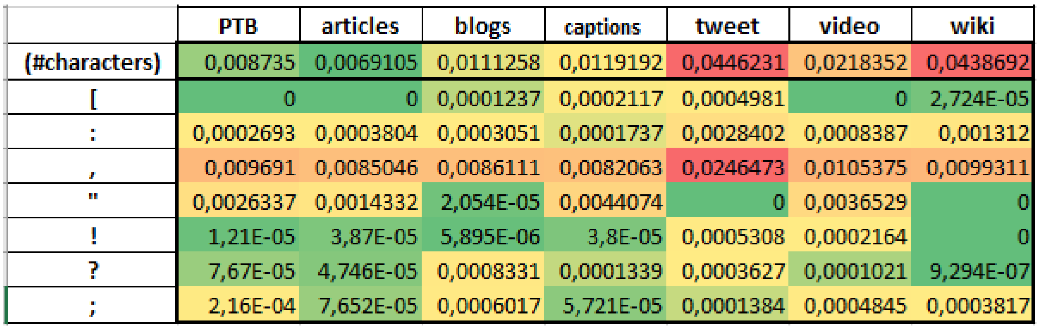
Interrogative words are quite infrequent across the genres, and are mostly found in articles that contain personal views, such as specialized articles and blog posts.

Others

Among the other PoS in Table 4, the use of possessive *POS* seems to be more frequent video descriptions and captions; interjections *UH* are found only in blog posts and tweets; symbols *SYM* are found only in captions, tweets and Wikipedia articles; hashtags are only present in tweets, which are the only genre in which the tagger sometimes cannot assign a tag (*NIL* is used when no part of speech is identified).

Punctuation signs

Table 5 shows the ratio of different punctuation signs, and the average number of characters between parentheses per sentence. The ratio of character between parentheses is particularly high in tweets and Wikipedia articles; the colon and comma are used quite uniformly across the genres; no exclamation sign is found in Wikipedia articles, and very few question marks, unlike in blog posts, captions and tweets. The question mark does not have the same functions in captions and tweets, in which it indicates uncertainty, than in blogs and video descriptions, in which it usually stands for real questions.



	PTB	articles	blogs	captions	tweet	video	wiki
(#characters)	0,008735	0,0069105	0,0111258	0,0119192	0,0446231	0,0218352	0,0438692
[	0	0	0,0001237	0,0002117	0,0004981	0	2,724E-05
:	0,0002693	0,0003804	0,0003051	0,0001737	0,0028402	0,0008387	0,001312
,	0,009691	0,0085046	0,0086111	0,0082063	0,0246473	0,0105375	0,0099311
"	0,0026337	0,0014332	2,054E-05	0,0044074	0	0,0036529	0
!	1,21E-05	3,87E-05	5,895E-06	3,8E-05	0,0005308	0,0002164	0
?	7,67E-05	4,746E-05	0,0008331	0,0001339	0,0003627	0,0001021	9,294E-07
;	2,16E-04	7,652E-05	0,0006017	5,721E-05	0,0001384	0,0004845	0,0003817

Table 5: Character-based features

3.3.3 Dependency-based features

Table 6 and Table 7 show the dependency-based features: the first table focuses on the individual dependencies, while the second table focuses on the complexity of the structures. The individual dependencies are split in four groups for analysis: (i) arguments; (ii) adverbials; (iii) modifiers; (iv) others.

Arguments

The argumental relations in the PTB tagset are the dependencies that hold below verbal elements and indicate the presence of an important participant. The ratio of argumental relations directly reflects the ratio of verbs seen in the PoS-based features. This means that in most cases, when a verb is present, its participants are present too. For all genres, as expected, the most frequently present argument is the subject of the verb *SBJ*. The second most important argument is also the second in terms of ratio, the object *OBJ*. The ratio of arguments in Wikipedia articles is about 2/3 of the ratio for PTB material, but the ratio of objects is less than half of PTB's. This tends to show that impersonal constructions using the passive voice (that is, with no syntactic object: *X was constructed in 1912*) are used more often, as the ratio of past participles seems to confirm (see *VBN* in Table 4). In the Wikipedia articles, more use if made of other impersonal constructions, such as constructions with logical subjects (*it*) *LGS*. The proportion of third nominal arguments of ditransitive verbs *DTV* is particularly low in specialized articles, and of third adjectival arguments *OPRD* is particularly low in Wikipedia articles.

Adverbials

Adverbial relations being found under verbs, it is again expected that their ratio is lower in captions and tweets. In general, temporal *TMP* and locative *LOC* adverbials are the most frequent among the typed adverbials (as opposed to ADV, which is not typed in the sense that it does not indicate a particular circumstantial label, as the other do). The Word-based features in Section 3.3.1 showed that Wikipedia pages contained a higher ratio of locative named entities; they also have a higher ratio of locative (and temporal) adverbials compared to the other genres, although they have less adverbs, which implies that the ratio of adverbial clauses that are not adverbs but rather prepositional groups is prominent in this genre (as is confirmed by the higher ration of prepositions noted in the previous section). On the contrary, purpose *PRP* and manner *MNR* adverbials exhibit lower ratios in Wikipedia pages compared to other genres.

Modifiers

Under the nouns, the ratios of modifiers across the genres are comparable. In articles and captions however, a higher ratio of modifiers emerges, which is directly related to the higher ratio of noun phrase components discussed in Section 3.3.2 (*DT*, *NN*, *JJ*, etc.). A significantly higher ratio of appositions *APPO* is seen in tweets.

Others

Complex verb groups are identified through the "verb chain" *VC* dependency; they comprise a base verb and one or more auxiliaries (e.g., *have been built*). A slightly higher ratio of auxiliaries is noted for blog posts, whereas it is rather low in video descriptions. Coordinations *COORD* are user in higher proportions in video descriptions, blog posts and specialized articles, and very little in tweets. Finally, a rather low ratio of unidentified dependencies is exhibited in all genres.

	PTB	articles	blogs	captions	tweet	video	wiki
ARG-all	0,157983158	0,14727925	0,17870483	0,06600529	0,04378554	0,11903348	0,11164715
DTV	0,000363583	3,3625E-05	0,00023211	8,3216E-05	2,4314E-05	0,00024464	0,00048205
LGS	0,002723739	0,0057889	0,00232935	0,00194127	0,00118103	0,0021829	0,00805448
OBJ	0,056523602	0,05328255	0,0610576	0,02673585	0,01658454	0,03940424	0,0253382
OPRD	0,009953612	0,00825612	0,00949037	0,00157393	0,00154213	0,00851258	0,0034154
PRD	0,013883027	0,0158493	0,02303298	0,00174247	0,00213873	0,01379353	0,01645222
SBJ	0,074535596	0,06406875	0,08256242	0,03392857	0,0223148	0,05489559	0,05790479
ADV-all	0,087664292	0,08672839	0,08679132	0,04356217	0,02808198	0,07938829	0,09064317
ADV	0,040427733	0,04488768	0,04700184	0,0110739	0,00724168	0,0361041	0,03347088
BNF	1,25E-05	1,17E-04	0,00E+00	0,00E+00	0,00E+00	0,00E+00	2,34E-05
DIR	0,004670163	0,00263792	0,00193901	0,00046117	0,00020036	0,00091354	0,00143949
LOC	0,016486616	0,01857066	0,01704373	0,02606217	0,01503223	0,02177793	0,0309629
MNR	0,003315085	0,00368651	0,0029731	0,00030166	0,00045836	0,00282895	0,00196198
PRP	0,002657918	0,00191008	0,00423005	0,00013083	0,00015039	0,00166229	0,00069939
TMP	0,020076478	0,01491842	0,01360358	0,00540762	0,00495302	0,01610148	0,02208228
VOC	1,77612E-05	0	0	0,00012482	4,5926E-05	0	2,8134E-06
MOD-all	0,301838811	0,35754159	0,30124882	0,34568026	0,243081	0,32788746	0,31375733
AMOD	0,014612282	0,01396221	0,0178639	0,00397901	0,00360522	0,01364523	0,00725454
APPO	0,01415049	0,01857874	0,0136948	0,02309527	0,04647115	0,02194624	0,02026258
NMOD	0,273076039	0,32500064	0,26969012	0,31860599	0,19300463	0,29229599	0,28624022
VC	0,02939173	0,02677436	0,03592981	0,00272907	0,00282852	0,01961944	0,02618247
COORD	0,025387091	0,03750531	0,04349271	0,0291787	0,01401294	0,04845221	0,02994928
DEP	0,022892158	0,00700058	0,00846326	0,00452239	0,0070542	0,00686598	0,00541491

Table 6: Dependency tag-based features

Syntactic complexity results are shown in Table 7. A larger number for the depth corresponds to a higher level of embeddedness in the structures, and combined with a high number for the mean maximal width of the trees accounts for the general complexity of the trees. From this perspective, specialized articles exhibit the highest complexity, whereas for blog posts, video descriptions and Wikipedia articles the complexity is comparable. Captions and tweets are expectedly the least complex in terms of tree shape. Note that the ramification factor of images is below that of tweets, which indicates a general lesser amount of children per node.

	PTB	articles	blogs	captions	tweet	video	wiki
Depth	7,2682839	9,4726668	7,9397046	2,6580117	3,1140173	8,0685725	6,9634627
Width	4,8972976	4,9178601	4,7065864	2,5959184	3,9699738	4,7691807	4,6305367
Ram	1,7130795	1,7382382	1,7226964	1,1529788	1,794088	1,7463553	1,7118452

Table 7: Syntactic complexity-based features

3.4 Preliminary concept extraction

The preliminary concept extraction method developed in order to assess the textual material in terms of the number of concepts of different types is based on a statistical measure that shows the importance of collocations of words taking into account their frequencies in topic-independent literature. The types of concepts we distinguish are the following: single-word concepts, multiword concepts, named entities, and numbers.

For concept extraction, we use the Google N-gram dataset for English[26] to obtain the statistics on usage of word combinations, i.e. the frequencies mentioned above. The dataset was originally inferred from the large collection of books dated from the beginning of the 19th century. It includes n-grams that occur in at least 40 books with automatically predicted part-of-speech tags and dependency relations.

The method comprises two general steps: candidate detection and candidate selection.

For the detection of candidates, we use the results of the text processing stage. First, we store recognized named entities as concepts of a particular type and eliminate them from further consideration. Second, we do the same with numbers selecting them as tokens having a part-of-speech tag "CD". Then we look for predefined templates corresponding to candidates or their parts, such as "noun", "adjective + noun", "noun + noun", "noun + *of* + noun", "verb + noun" (only verbs with the tags "VBD", "VBG", "VBN" or with a tag "VB" but not ending on a letter "s"), "adjective + adjective", "noun + adjective", using punctuation, stopwords, named entities, and numbers as the borders of candidates.

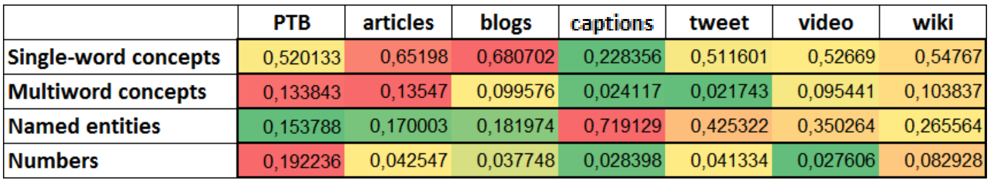
To select the candidates to be treated as concepts, we check the conditioned frequency of each found two-word candidate in the Google bigram dataset using its part-of-speech tags assigned to single words taking into account possible confusions of a PoS-tagger: we find a position of a bigram among other bigrams starting with the same word and having the same part-of-speech tag for the second word, which are sorted by their frequencies, and evaluate the slope to the curve drawn through the normalized frequencies in a point corresponding to the surveyed bigram. We use at most 50 neighbours from each side of the probable concept to calculate the slope. We find a slope for the bigram being placed among the bigrams ending with the same word and having the same part-of-speech tag for the first word the same way and select the highest value of two slopes. We use a threshold to define if the bigram is highly frequent in comparison to its neighbours. In case the slope is less than the threshold we select nouns in the bigram for the collection of single-word concepts.

[26] https://books.google.com/ngrams/

Otherwise, we combine all adjacent highly frequent two-word combinations which share the same tokens selecting the one most to the right finishing with the noun as the end of a multiword concept. All the rest of nouns that were not included in any group of concepts are treated as single-word concepts. To define the value of the threshold we used an annotated dataset from Codina-Filbà and Wanner (2016) that was split into train and test sets. We varied the value on the train set to achieve the highest F_1-score and selected the one equal to 85 degrees. It also gave high F_1-score equal to 0.63 on the test set.

For the evaluation, we selected 200 paragraphs for each type of content, where a paragraph is a set of sentences starting from a new line. For some cases, a paragraph is a whole text corresponding to the object, for some other cases, it is just a part of a text which includes several of them.

The results of the assessment are shown in Table 8. Several peculiarities of the data might be highlighted. Single-word concepts are the main component of the paragraphs of all the types of content except images with extreme values for articles and blogs which are 25-30% higher than the average number of single concepts in a general text. Multiword concepts are more typical in articles and could be rarely found in captions and tweets. In opposite, named entities are the main component of captions and one of the most common components of tweets and videos that is only a bit less frequent than single-word concepts. Numbers appear more often in Wikipedia pages rather than in other types of content.



	PTB	articles	blogs	captions	tweet	video	wiki
Single-word concepts	0,520133	0,65198	0,680702	0,228356	0,511601	0,52669	0,54767
Multiword concepts	0,133843	0,13547	0,099576	0,024117	0,021743	0,095441	0,103837
Named entities	0,153788	0,170003	0,181974	0,719129	0,425322	0,350264	0,265564
Numbers	0,192236	0,042547	0,037748	0,028398	0,041334	0,027606	0,082928

Table 8: Different types of concepts

Note that the numbers reported in Table 8 are not comparable to the ones reported in Table 4 for instance, which also mentions Named Entities. Here, the values represent the ratio of named entities over the detected concepts (i.e., a subset of the words), whereas in Table 4 it was the ratio of Named Entities over all the words.

4 QUANTITATIVE ASSESSMENT OF THE TEXTUAL MATERIAL FOR SEMANTIC ANALYSIS AND AUTOMATIC SUMMARIZATION

WP3 foresees several deep analysis tasks with the overall goal of extracting relevant information that can be integrated as linked data (T5.3) in a semantic repository. This deep analysis involves, amongst other tasks, identifying references in the text to real-world entities and concepts, a task that involves addressing various linguistic analysis problems such as named entity recognition and linking, word sense disambiguation and coreference resolution –all of them part of T3.2.

We present the results of a quantitative analysis of textual materials relevant to the project use cases that, unlike the linguistic analysis presented in the previous sections of this document, focuses on aspects related to the meaning of the texts. More precisely, we evaluate the coverage and performance of various linguistic semantic resources and tools that are key to the extraction of information from text to be integrated in a semantic repository with information coming from other sources and aligned with the project ontologies. These resources will also be important for other tasks such as the planning and generation of explanatory texts (T5.4 and T5.5). In this section, we assess the coverage of BabelNet meanings (4.1), the average polysemy of the words in the texts (4.2), the coverage of sense embeddings (4.3), the frequencies of the different meaning (4.4) and the distribution and coverage of Named Entities and coreference chains (4.5).

4.1 Coverage of BabelNet meanings

BabelNet is a multilingual lexical database and knowledge base that contains entries for a very large number of meanings and real-world entities, both accompanied with lexicalizations in multiple languages. This database is a mapping between language-specific versions of Wikipedia and WordNet, resulting in a large coverage of both entities, word meanings and their associated names and lexicalizations. We aim to evaluate the suitability of BabelNet as a target repository of word meanings and named entities for entity linking and word sense disambiguation tasks, and we do so by measuring the coverage provided by this resource of the meanings conveyed by selected texts relevant to V4Design.

A precise evaluation would require the manual analysis of the texts in order to assess to what extent the entries are available in BabelNet able to correctly express the meaning of the texts. The inherent ambiguities in natural language and its interpretation by human readers make this type of analysis a very difficult and time-consuming task. For this reason, we conduct an automated quantitative analysis where the coverage of the meaning of a text by BabelNet is approximated by calculating the share of content words -words carrying meaning- that produce at least one meaning when looked up in BabelNet.

The results of this analysis are shown in Table 9, broken down by grammatical category (rows) and document genres (columns). For each combination of category and genre, the table lists the total number of tokens belonging to the category in each of the genre-specific collection of documents (columns labelled ‘NT’), the number of tokens with at least a BabelNet meaning (columns labelled ‘CT’), and the proportion of tokens covered by BabelNet (columns labelled ‘C’). This proportion is obtained by averaging the results of dividing for each document and category the number of tokens with a BabelNet meaning by the total number of tokens. Lookups in BabelNet are conducted using both word forms and their corresponding lemmas.

		Articles	Blogs and Forums	Captions	Tweets	Video descriptions	Wikipedia articles	All
Nouns	NT	5335	3666	14985	82582	2402	36395	145365
	CT	4998	3513	9940	58122	2266	34443	113282
	C	0.94	0.96	0.66	0.70	0.94	0.95	**0.80**
Verbs	NT	2615	2050	1607	8263	916	13597	29048
	CT	989	914	404	2702	367	3323	8699
	C	0.38	0.45	0.25	0.33	0.41	0.24	**0.30**
Adjectives	NT	1860	1103	1882	6389	655	8069	19958
	CT	1725	1057	1471	4097	624	7610	16584
	C	0.93	0.96	0.78	0.64	0.95	0.94	**0.83**
Cardinals	NT	250	141	856	17904	102	4139	23392
	CT	226	135	430	5977	96	3768	10632
	C	0.90	0.96	0.50	0.33	0.94	0.91	**0.46**
Adverbs	NT	842	582	321	1372	251	2426	5794
	CT	822	538	305	978	248	2409	5300
	C	0.98	0.92	0.95	0.71	0.99	0.99	**0.92**
Content words	NT	10902	7542	19651	116510	4326	64626	223557
	CT	8760	6157	12550	71876	3610	51553	154506
	C	**0.80**	**0.82**	**0.64**	**0.62**	**0.83**	**0.80**	**0.69**

Table 9: Number of tokens indexed in BabelNet

Table 9 only reflects the coverage of single-word meanings. Multiple words can also communicate meanings, which in some cases bear little or no relation to the meanings of the individual words. The number of such multiword expressions indexed in BabelNet serves as an approximation of the coverage of BabelNet of meanings conveyed in texts by more than one word. We consider candidate multiwords all sequences of up to 5 consecutive tokens with at least one noun and no punctuation signs. Table 10 shows the results of looking up the collected multiwords in BabelNet, broken down by grammatical category and genre. As in Table 9, we report the total number of tokens in each category and the share of them that are part of a multiword indexed in BabelNet.

		Articles	Blogs and Forums	Captions	Tweets	Video descriptions	Wikipedia articles	All
Nouns	NT	5335	3666	14985	82582	2402	36395	145365

	CT	1692	1223	3818	21801	871	16199	45604
	C	0.32	0.33	0.25	0.26	0.36	0.45	**0.31**
Verbs	NT	2615	2050	1607	8263	916	13597	29048
	CT	31	14	25	261	15	131	477
	C	0.01	0.01	0.02	0.03	0.02	0.01	**0.02**
Adjectives	NT	1860	1103	1882	6389	655	8069	19958
	CT	240	117	463	836	133	1961	3750
	C	0.13	0.11	0.25	0.13	0.20	0.24	**0.19**
Cardinals	NT	250	141	856	17904	102	4139	23392
	CT	31	135	38	387	21	490	1102
	C	0.12	0.13	0.04	0.02	0.21	0.12	**0.05**
Adverbs	NT	842	582	321	1372	251	2426	5794
	CT	6	7	79	44	3	10	149
	C	0.01	0.01	0.25	0.03	0.01	0.01	**0.02**
Content words	NT	10902	7542	19651	116510	4326	64626	223557
	CT	2000	1380	4423	23329	1043	18791	50966
	C	**0.18**	**0.18**	**0.23**	**0.20**	**0.24**	**0.29**	**0.23**

Table 10: Number of tokens part of multiwords indexed in BabelNet

While BabelNet has a large coverage of nouns, adjectives and adverbs (coverage is 0.80, 0.83 and 0.92 respectively), only a third of the verbs in our corpora produce at least one meaning when looked up in this resource. The differences between grammatical categories remain constant across genres, but the overall coverage of content words is significantly lower for captions and tweets. Unsurprisingly, nouns and adjectives are more often part of multiword expressions indexed in BabelNet than words of other grammatical categories. Multiwords seem particularly abundant in Wikipedia, perhaps due to the presence of named entities. A high number of multiwords in video descriptions may be explained by the fact that many of these descriptions contain fragments of Wikipedia pages.

4.2 Average polysemy

Word or multiwords removed from context may correspond to multiple meanings. Solving this semantic ambiguity is a difficult task the complexity of which depends on the degree of polysemy -the number of potential candidate meanings of a given expression. Not all words possess the same degree of polysemy. Common words are often more ambiguous than terms used in specific domains and textual genres. In order to gain a sense of the scale of the disambiguation problem, we measure the average number of meanings in BabelNet indexed for each word. The resulting numbers, listed in Table 10, are broken down by POS and genre and account both for meanings associated to individual words and meanings associated to multiword expressions they are part of.

	Articles	Blogs and Forums	Captions	Tweets	Video descriptions	Wikipedia articles	All
Nouns	11,40	11,49	7,33	10,88	8,30	10,61	**10,00**
Verbs	8,52	6,92	0,01	0,05	3,72	9,75	**4,83**
Adjectives	3,56	4,36	1,55	0,43	3,44	3,21	**2,76**
Cardinals	12,10	25,03	0,06	3,99	0,88	5,40	**7,91**
Adverbs	3,12	2,50	0,08	0,01	3,55	3,66	**2,15**
Content words	**9,48**	**8,38**	**6,85**	**9,30**	**6,84**	**8,47**	**8,22**

Table 11: Average polysemy

Nouns and cardinals have the largest overall degree of polysemy, while adjectives and adverbs have the lowest. These observed differences, however, are not consistent across genres. Verbs, for instance, are highly polysemous in articles and Wikipedia compared to captions and tweets. The figures are likely to be distorted in some genres due to the low number of tokens belonging to certain categories in some genres, e.g. there are only 141 cardinal tokens in the blogs and forums corpus. Nevertheless, captions and video descriptions exhibit a lower average degree of polysemy when compared to other corpora, indicating important differences between the lexicon of the genres.

4.3 Coverage of sense embeddings

Deep analysis of text up to a semantic level implies several advantages for downstream applications, ranging from abstraction from language-specific idiosyncrasies to easier integration in knowledge repositories and with non-linguistic ontologies. Depending on the text genre and goals of the overall system, not all the semantic content of a text may be relevant. Parts of the text may be redundant or serve rhetorical purposes rather than purely informative ones. In addition, when processing long documents or large volumes of text it may be necessary to determine what the central topics are so that they can be presented to the user of the system in a concise way.

Research in automatic summarization has produced multiple strategies for detecting relevant contents. Most of them are based on measuring the similarity between parts of the text, and on the frequency of content items in the text or in reference corpora. We conduct a study of how existing BabelNet-based resources can be used to obtain similarity and frequency values for BabelNet meanings in this section and the following one respectively.

Our quantitative assessment of resources to obtain similarity values involves comparing four different sets of distributional vectors for BabelNet senses in terms of their coverage of the meanings that can be associated with words and multiwords in texts belonging to each of the genres. A large coverage would imply that, a priori, the meanings in the V4Design domain can be effectively compared in terms of their similarity. Table 12 gives coverage figures for the following state-of-the-art distributional vectors: SenseEmbed (column labelled 'SE'), Nasari (column labelled 'NA'), SEWEmbed based on Nasari (column labelled

'SEW-NA') and SEWEmbed based on Word2Vec (column labelled 'SEW-WV'). As in previous tables, the results are broken down by grammatical category and genre.

		Articles	Blogs and Forums	Captions	Tweets	Video descriptions	Wikipedia articles	All
Nouns	SE	0,50	0,51	0,45	0,45	0,45	0,52	**0,48**
	NA	0,90	0,91	0,92	0,92	0,92	0,91	**0,91**
	SEW-NA	0,87	0,87	0,88	0,88	0,88	0,88	**0,88**
	SEW-WV	0,87	0,86	0,88	0,88	0,88	0,88	**0,88**
Verbs	SE	0,77	0,74	0,78	0,79	0,75	0,73	**0,76**
	NA	0,00	0,00	0,00	0,00	0,00	0,00	**0,00**
	SEW-NA	0,25	0,25	0,19	0,27	0,27	0,24	**0,25**
	SEW-WV	0,25	0,25	0,19	0,27	0,27	0,24	**0,25**
Adjectives	SE	0,92	0,91	0,93	0,93	0,93	0,91	**0,92**
	NA	0,00	0,00	0,00	0,00	0,00	0,00	**0,00**
	SEW-NA	0,28	0,29	0,34	0,28	0,31	0,30	**0,30**
	SEW-WV	0,28	0,29	0,34	0,28	0,31	0,30	**0,30**
Cardinals	SE	0,32	0,35	0,31	0,32	0,33	0,33	**0,33**
	NA	0,94	0,95	0,95	0,91	0,95	0,90	**0,93**
	SEW-NA	0,77	0,78	0,78	0,76	0,78	0,72	**0,77**
	SEW-WV	0,77	0,78	0,77	0,76	0,78	0,72	**0,76**
Adverbs	SE	0,58	0,57	0,53	0,52	0,56	0,59	**0,56**
	NA	0,00	0,00	0,00	0,00	0,00	0,00	**0,00**
	SEW-NA	0,32	0,32	0,36	0,34	0,32	0,35	**0,34**
	SEW-WV	0,32	0,32	0,36	0,34	0,32	0,35	**0,34**
Content words	SE	0,55	0,55	0,47	0,48	0,50	0,53	**0,51**
	NA	0,70	0,70	0,87	0,87	0,76	0,80	**0,78**
	SEW-NA	0,73	0,72	0,84	0,84	0,77	0,79	**0,78**
	SEW-WV	0,73	0,72	0,84	0,84	0,77	0,79	**0,78**
All		**0,53**	**0,53**	**0,54**	**0,54**	**0,53**	**0,53**	**0,53**

Table 12: Coverage of distributional vectors for meanings

Coverage of content words varies greatly across resources due to differences in their coverage of grammatical categories. Thus, SensEmbed ('SE') has the best coverage of verb, adverb and adjective meanings (0,76, 0,56 and 0,92 respectively), but falls behind other vector sets in defining vectors for meanings of nouns and cardinals (0,48 and 0,33). Nasari vectors ('NA') have a very good coverage of nominal and cardinal meanings (0,91 and 0,93) but do not cover meanings for words of any other grammatical category. Both versions of the SEW vectors (‘SEW-NA’ and ‘SEW-WV’) offer the most balanced coverage across categories. As indicated by the average across vector sets in the bottom row of Table 12, the results do not vary significantly with the genre.

4.4 Meaning frequencies in corpora

Frequency-based metrics are another important mechanism for evaluating contents in automatic summarization literature. Frequencies can be obtained from the input documents and compared with reference corpora and may be pondered according to other criteria such as the relative position in the document of each occurrence of a content item. In our quantitative assessment of BabelNet meanings we look at the frequencies of meanings both in the documents belonging to the V4Design corpora and in Semantically Enriched Wikipedia (SEW) corpus, a dump of Wikipedia annotated with BabelNet senses.

Table 13 shows results of collecting frequencies for each candidate BabelNet meaning of a word or multiword in the corpora. As usual, the results are shown separately for each grammatical category and genre. The values shown correspond to the median of:

1. the number of mentions of a meaning in a document (column labelled ‘F’),
2. the number of documents mentioning a meaning (column labelled ‘DF’),
3. the number of mentions of a meaning in SEW (column labelled ‘SEW-F’), and
4. the number of documents mentioning a meaning in SEW (column labelled ‘SEW-DF’).

		Articles	**Blogs and Forums**	**Captions**	**Tweets**	**Video descriptions**	**Wikipedia articles**	**All**
Nouns	F	1	1	2	2	1	2	**1,50**
	DF	1	1	1	2	1	2	**1,33**
	SEW-F	15	14	21	24	15	27	**19,33**
	SEW-DF	8	7	11	14	8	15	**10,50**
Verbs	F	2	2	1	2	2	2	**1,83**
	DF	2	2	1	2	2	2	**1,83**
	SEW-F	2	2	2	3	2	3	**2,33**
	SEW-DF	2	2	1	3	2	3	**2,17**
Adjectives	F	2	1	1	2	2	3	**1,83**
	DF	2	1	1	2	1	3	**1,67**
	SEW-F	3	2	2	5	2	6	**3,33**
	SEW-DF	3	2	2	5	2	6	**3,33**
Adverbs	F	0	1	2	5	2	3	**2,17**
	DF	1	1	2	5	2	3	**2,33**
	SEW-F	2	8	12	15	10	6	**8,83**

	SEW-DF	2	6	9	14	8	6	**7,50**
Content words	F	10	3	3	2	2	4	**4,00**
	DF	9	2	1	2	2	4	**3,33**
	SEW-F	2	6	3	23	4	12	**8,33**
	SEW-DF	2	6	3	12	4	12	**6,50**

Table 13: Frequencies of meanings

The average median of a content word meaning in SEW is 8,33, and the average median of the number of documents in SEW containing that meaning is 6,50. These values vary with the genre, being much higher in the tweet and Wikipedia corpora (23/12 and 12/12 respectively) than in other corpora. In the case of Tweets, the higher values are caused by the prevalence of nouns which are the grammatical category with larger number of annotations in SEW. In Wikipedia articles, the higher frequencies are likely to be caused by fact that SEW is an annotated version of the same texts in the Wikipedia corpus, thus leading to higher frequencies.

4.5 Coverage of NEs and coreference chains

While the coverage of BabelNet for common concepts and word meaning is likely to be reasonably extensive, references to real-world entities cannot be expected to be fully indexed. This lack of coverage may be caused by references to entities -people, locations, etc.- that do not have a Wikipedia page and therefore do not appear in BabelNet. In other cases, entities may be referred to in the text using lexicalizations not being indexed in BabelNet, e.g. due to being uncommon or because of spelling errors. Anaphoric expressions that require interpretation using the context cannot be linked to entries in BabelNet without specialized coreference resolution methods.

We use state-of-the-art Named Entity Recognition (NER) and coreference resolution tools to gauge the amount of such entities and referring expressions in the V4Design corpora -we use tools included in the Stanford CoreNLP package for this purpose. Table 14 lists the number of named entities (column labelled '#NE') broken down by type, and the portion of its tokens that have at least one BabelNet meaning associated to them (column labelled 'C'). Table 15 lists for each genre the number of coreference chains detected by the tool (row labelled '#Chains'), the overall fraction of tokens in the corpora that are part of a chain ('Coverage') and the share of these tokens that have at least one BabelNet meaning (row labelled 'Meaning coverage').

	Articles		**Blogs and Forums**		**Captions**		**Tweets**		**Video descriptions**		**Wikipedia articles**		**Total**	
	#NE	C	#NE	C	#NE	C	#NE	C	#NE	C	#NE	C	#NE	C
CAUSE OF DEATH	0	1,00	5	1,00	77	0,75	186	0,89	20	1,00	212	0,93	500	0,89
CITY	95	0,91	75	1,00	295	0,55	1521	0,92	56	1,00	815	1,00	2857	0,91

COUNTRY	51	1,00	10	0,91	262	0,53	2195	0,96	47	1,00	535	0,99	3100	0,93
CRIMINAL CHARGE	0	1,00	2	1,00	10	0,82	25	0,88	0	1,00	23	1,00	60	0,92
DATE	178	0,99	86	0,97	668	0,74	11699	0,30	70	0,98	3355	0,98	16056	0,50
DURATION	32	0,91	19	0,97	336	0,25	2132	0,37	22	1,00	233	0,82	2774	0,42
EMAIL	0	1,00	0	1,00	0	1,00	1	0,00	0	1,00	0	1,00	1	0,00
IDEOLOGY	20	1,00	1	1,00	8	1,00	15	0,89	4	1,00	35	1,00	83	0,98
LOCATION	86	0,80	36	0,98	451	0,53	1843	0,84	67	0,88	1767	0,91	4250	0,83
MONEY	5	0,91	2	1,00	0	1,00	72	0,82	0	1,00	21	0,69	100	0,81
NATIONALITY	63	1,00	12	1,00	194	0,73	345	0,94	62	0,87	702	1,00	1378	0,94
NUMBER	158	0,87	94	0,96	285	0,85	3800	0,53	61	0,91	1790	0,83	6188	0,63
ORDINAL	24	1,00	22	1,00	94	0,95	213	0,96	7	1,00	437	1,00	797	0,98
ORGANIZATION	99	0,90	60	0,93	177	0,80	1573	0,90	30	0,96	614	0,90	2553	0,89
PERCENT	1	1,00	5	1,00	0	1,00	7	0,86	1	1,00	0	1,00	14	0,93
PERSON	309	0,74	250	0,90	2292	0,53	11047	0,68	292	0,92	3291	0,88	17481	0,70
RELIGION	0	1,00	3	1,00	70	0,93	33	1,00	8	1,00	78	1,00	192	0,97
SET	5	1,00	5	1,00	4	0,75	58	0,23	4	1,00	28	0,97	104	0,57
STATE OR PROVINCE	15	1,00	17	0,90	36	0,54	580	0,83	22	1,00	25	1,00	695	0,84
TIME	10	1,00	6	0,83	47	0,65	223	0,78	3	1,00	18	0,92	307	0,78
TITLE	91	1,00	34	1,00	519	0,77	903	0,95	72	1,00	1026	1,00	2645	0,94
URL	0	1,00	12	0,00	0	1,00	30	0,00	0	1,00	1	0,00	43	0,00
TOTAL	**1242**	**0,96**	**756**	**0,93**	**5825**	**0,76**	**38501**	**0,71**	**848**	**0,98**	**15006**	**0,90**	**62178**	**0,74**

Table 14: Coverage of named entities

	Articles	Blogs and Forums	Captions	Tweets	Video descriptions	Wikipedia articles	Total
#Chains	409	315	267	426	145	2703	**4265**
Coverage	0,10	0,07	0,03	0,00	0,08	0,12	**0,07**
Meaning coverage	0,84	0,92	0,91	0,91	0,94	0,91	**0,91**

Table 15: Coverage of coreferent expressions

The distribution of entity types is similar across genres, with entities related to places, time, people, organizations and numbers being the most frequent. Numbers and temporal entities have a lower coverage in BabelNet compared to other types (0,42 for duration, 0,50 for date, 0,57 for set. 0,63 for number and 0,78 for time). Entities of type=person also have a lower average (0,70) compared to geographical entities and organizations (0,93 for country, 0,91 for city, 0,89 for organizations, etc.). The tweets and captions datasets have the lowest average coverage for the NEs detected in them (0,71 and 0,76 compares to coverage values over 0,90 for the rest). The preference for shorter names or abbreviated forms in these genres may be causing this.

BabelNet has good coverage of the content words that are part of coreferent expressions in the text. Our counts exclude demonstratives and pronouns used in anaphorical expressions, thus reflecting mostly the coverage of first and representative mentions that tend to use fully qualified names. While coreferent expressions cover less than 10% of the texts in most genres (and much less in shorter texts like captions and social media messages), detecting them in larger texts contributes toward finding many mentions to entities that wouldn't be detected otherwise.

5 CONCLUSIONS

In this deliverable, we reported on the empirical study carried out on the V4Design made available by the consortium during the first months of the project. This empirical study showed that the textual sources have different linguistic shapes, and the specificities of each genre have been pointed out. For this, we gave an account of the linguistic features of 6 textual genres relevant to V4Design (specialized articles, blog posts, captions, tweets, video descriptions and Wikipedia articles) in the context of two tasks: morpho-syntactic analysis on the one hand, and semantic analysis and summarization on the other hand, for which we also evaluated the coverage of several tools. We also conducted an initial study on the nature of the dynamic textual contents, concluding that the most promising way to address these contents seems to be through the detection of opinions trends among professional reviewers at different points in time.

The assessment of the texts for the morpho-syntactic analysis was carried out in terms of **word-based**, **character-based**, **part-of-speech-based** and **dependency-based** features, and **concept distribution**. The following table provides an overview of the main features observed for each genre.

	General comments	**Low ratio of…**	**High ratio of…**	**Similar to…**
Articles (Specialized)	- Long sentences - High syntactic complexity	- Di-transitive nominal arguments - Named Entities	- Adjectives - Determiners - Interrogative words - Prepositions - Single word concepts	- Penn Treebank - Blog posts
Blog posts	- Long sentences - Mild syntactic Complexity	- Named Entities	- Adjectives - Auxiliaries - Coordinations - Determiners - Interrogative words - Prepositions - Pronouns (1st pers.) - Verbs - Single word concepts	- Penn Treebank - Specialized articles
Captions	- Short sentences - Very low syntactic Complexity - Language and encoding issues in EF painting captions	- Adverbs - Adjectives - Interrogative words - Pronouns - Verbs - Single word concepts - Multi-word concepts	- Common nouns - Named Entities (persons) - Proper nouns	- Tweets
Tweets	- Short sentences - Low syntactic Complexity - Important qualitative issues in text	- Determiners - Adjectives - Multi-word concepts	- Common nouns - Hashtags - Named Entities (persons, locations) - Proper nouns - Unidentified PoS	- Captions
Video descriptions	- Long sentences - Mild syntactic Complexity - Encoding issues in the AF dataset	- Auxiliaries - Verbs	- Coordinations - Prepositions - Pronouns (3rd pers.)	- Wikipedia articles
Wikipedia articles	- Long Sentences - Mild syntactic Complexity - Minor character	- Adverbs - Di-transitive adjectival arguments - Infinitive verbs	- Impersonal constructions - Locative adverbials - Named Entities	- Video Descriptions - Penn Treebank

	Issues in text	- Modal verbs - Objects - Verbs	(persons, locations) - Prepositional adverbial clauses - Proper nouns	

The assessment of the texts for the semantic analysis and automatic summarization was carried out using more semantic features. The analysis of the **BabelNet coverage** showed that while BabelNet has a large coverage of nouns, adjectives and adverbs, only a third of the verbs in our corpora produce at least one meaning when looked up in this resource. The differences between grammatical categories remain constant across genres. As far as **average polysemy** is concerned, nouns and cardinals have the largest overall degree of polysemy, while adjectives and adverbs have the lowest. These observed differences, however, are not consistent across genres. The **sense embedding** coverage of content words varies greatly across resources due to differences in their coverage of grammatical categories. SensEmbed has the best coverage of verb, adverb and adjective meanings, but falls behind other vector sets in defining vectors for meanings of nouns and cardinals. Nasari vectors have a very good coverage of nominal and cardinal meanings but do not cover meanings for words of any other grammatical category. Both versions of the SEW vectors offer the most balanced coverage across categories. In terms of **meaning frequencies,** the average median of a content word meaning values vary with the genre, being much higher in the tweet and Wikipedia corpora than in other corpora. For **fine-grained named entities**, the distribution of entity types is comparable across genres, with entities related to places, time, people, organizations and numbers being the most frequent. Finally, for **coreference** coverage, we noted that BabelNet has good coverage of the content words that are part of coreferent expressions in the text.

The qualitative study showed that most of the textual contents contemplated by the consortium can be processed by the Natural Language processing tool with some basic cleaning and filtering rules. For tweets however, the analysis may be limited to the extraction of concepts, since some major issues are foreseen that could impact the quality of a deeper analysis. The quantitative study highlighted some similarities and differences across the six different genres, which will be taken into account for the development of the WP3 and WP5 linguistic modules, and lead us to some initial conclusions with respect to which tools to use in the analysis pipeline.

6 REFERENCES

Bohnet, B. and Nivre, J. (2012). A transition-based system for joint part-of-speech tagging and labeled non-projective dependency parsing. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 1455–1465. Association for Computational Linguistics.

Codina-Filbà, J. and Wanner, L., 2016. Combining Dictionary-and Corpus-Based Concept Extraction. In MMDA@ ECAI (pp. 39-44).

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. and McClosky, D., 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations* (pp. 55-60).

Marcus, M.P., Marcinkiewicz, M.A. and Santorini, B., 1993. Building a large annotated corpus of English: The Penn Treebank. Computational linguistics, 19(2), pp.313-330.

Robinson, J. J. (1970). Dependency structures and transformational rules. Language, pages 259–285.

Surdeanu, M., Johansson, R., Meyers, A., Màrquez, L., and Nivre, J. (2008). The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies. In Proceedings of the Twelfth Conference on Computational Natural Language Learning, pages 159–177. Association for Computational Linguistics.