# V4Design

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

H2020-779962

# D3.3

# Basic version of multilingual semantic text analysis

| | |
|---:|---|
| **Dissemination level:** | Public |
| **Contractual date of delivery:** | Month 18, 30 June 2019 |
| **Actual date of delivery:** | Month 18, 30 June 2019 |
| **Workpackage:** | WP3 Visual and Textual content analysis |
| **Task:** | T3.2 Entity identification and linking, word sense disambiguation and lexical modelling<br>T3.3 Dependency-based semantic parsing<br>T3.4 Conceptual relation extraction |
| **Type:** | Report |
| **Approval Status:** | Approved |
| **Version:** | 1.2 |
| **Number of pages:** | 64 |
| **Filename:** | D3.3_V4Design_BasicAnalysisTechniques_v1.2.pdf |

**Abstract**

In this deliverable, we report the advances on the Language Analysis components achieved during the first half of the V4Design project. The components include in particular a multilingual candidate concept detection tool, multilingual dependency parsers, semantic analysers, lexical resources, and a projection of the extracted dependency-based linguistic representations into ontological ones.

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 | 05/04/2019 | Creation of the ToC | Simon Mille |
| 0.2 | 06/05/2019 | First version of SoA, Basic Techniques and Evaluation for Syntactic and Semantic parsing | Simon Mille, Beatriz Fisas |
| 0.3 | 03/06/2019 | Added contents in SoA (German, Greek) | Simon Mille, Beatriz Fisas |
| 1.0 | 18/06/2019 | Concept extraction, finalised contents | Alexander Shvets, Simon Mille |
| 1.1 | 26/06/2019 | Addressed comments from internal review | Simon Mille, Alexander Shvets |
| 1.2 | 28/06/2019 | Final formatting | Simon Mille |

# Author list

| Organisation | Name | Contact Information |
|---|---|---|
| UPF | Simon Mille | simon.mille@upf.edu |
| UPF | Alexander Shvets | alexander.shvets@upf.edu |
| UPF | Beatriz Fisas | beatriz.fisas@upf.edu |

# Executive Summary

The Language Analysis module addresses the analysis and capture of the natural language textual material into structured, ontological (henceforth *Knowledge Base*, *KB*) representations, so that appropriate system responses can subsequently be inferred by the WP5 reasoning module (CERTH), and that textual summaries can be produced (WP5 TALN-Language Generation). The module combines multilingual dependency parsers and lexical resources, and a projection of the extracted dependency-based linguistic representations into ontologically-oriented ones. This deliverable covers the M1-M18 advances on the following tasks:

- **T3.2: Entity identification and liking, WSD and lexical modelling**
- **T3.3: Dependency-based semantic parsing**
- **T3.4: Conceptual relation extraction**

We thus report the advances on the Language Analysis components achieved during the first half of the V4Design. This includes (i) an extensive overview of the annotated corpora, the lexical resources, and the available open-source tools that can be useful for the analysis pipeline, (ii) the development and evaluation of new statistical algorithms for multilingual concept candidate detection in English, Spanish and German, (iii) advances on the automatic compilation of high-quality lexical resources in Spanish; (iv) the training and evaluation of statistical modules for sentence segmentation, lemmatisation, part-of-speech tagging, morphological tagging and dependency parsing in English, Spanish, German, and Greek; (v) the development of new graph-transduction grammars to be used for multilingual semantic and conceptual relation extraction in English, Spanish and Greek; (vi) the integration of all aforementioned components into the V4Design architecture; (vii) the preliminary definition of the foundations of the conceptual model used as interface with the KB; (viii) preliminary experiments towards the definition of new evaluation metrics for dependency parser evaluation.

# Abbreviations and Acronyms

| | |
|---|---|
| **A1, A2, etc** | First argument, second argument, etc. |
| **AM** | Modifier argument |
| **AUC** | Area under precision-yield curve |
| **CoNLL** | Conference on Natural Language Learning (format standard) |
| **DSynt** | Deep-Syntax |
| **GDT** | Greek Dependency Treebank |
| **IE** | Information Extraction |
| **INV** | Inverted (with A1, A2, etc.) |
| **KB** | Knowledge Base |
| **LA** | Language Analysis |
| **LAS** | Labelled Attachment Score |
| **HLUR** | High-Level User Requirement |
| **LSTM** | Long Short-Term Memory |
| **MTT** | Meaning-Text Theory |
| **MWE** | Multi-Word Expression |
| **N** | Noun |
| **NB** | NomBank |
| **NER** | Named Entity Recognition |
| **NLP** | Natural Language Processing |
| **NMT** | Neural Machine Translation |
| **OOV** | Out-of-vocabulary |
| **PB** | PropBank |
| **PoS** | Part of Speech |
| **PredArg** | Predicate-Argument |
| **Seq2seq** | Sequence-to-sequence |
| **SSynt** | Surface-syntax |
| **TR** | Technical Requirement |
| **UAS** | Unlabelled Attachment Score |
| **UD** | Universal Dependencies |
| **UR** | User Requirement |

| | |
|---|---|
| **VB** | Verb |

# Table of Contents

# 1 INTRODUCTION

The Language Analysis module addresses the analysis and capture of the natural language textual material gathered by the crawling component into structured, ontological (henceforth *Knowledge Base*, *KB*) representations, so that appropriate system responses can subsequently be inferred by the WP5 reasoning module (CERTH), and that textual summaries can be produced (WP5 TALN-Language Generation); see Figure 1. The module combines multilingual dependency parsers and lexical resources, and a projection of the extracted dependency-based linguistic representations into ontological ones. This deliverable covers the M1-M18 advances on the following tasks:

- **T3.2: Entity identification and liking, WSD and lexical modelling**
- **T3.3: Dependency-based semantic parsing**
- **T3.4: Conceptual relation extraction**

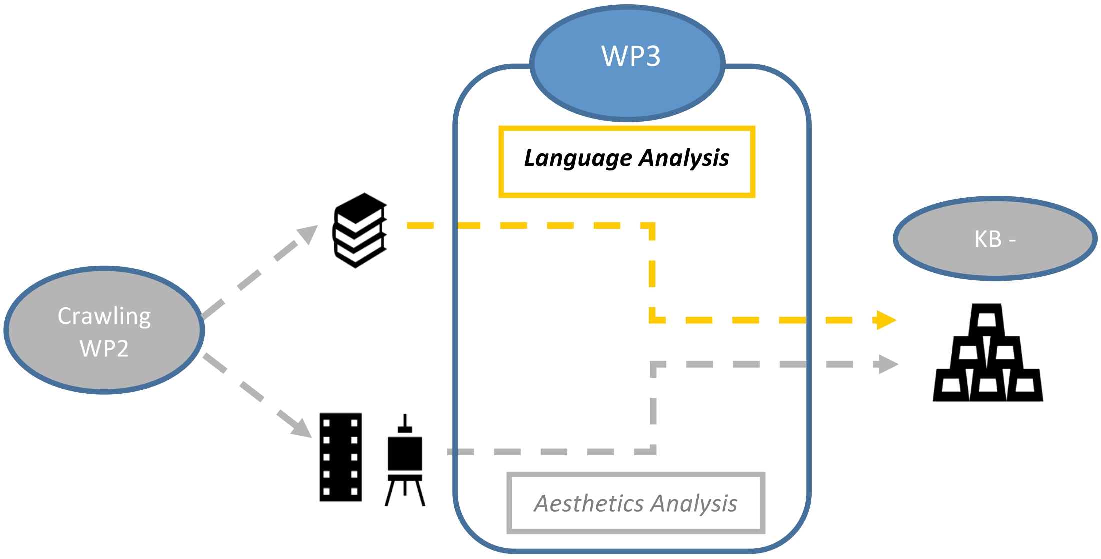


Figure 1: The Language Analysis component in the V4Design architecture

Consider for instance the following sentence: *The Chrysler building was constructed in 1930 by Walter Chrysler, the head of the Chrysler Corporation*, and a conceptual representation in Figure 2, which are respectively input and potential output of the Language Analysis component.

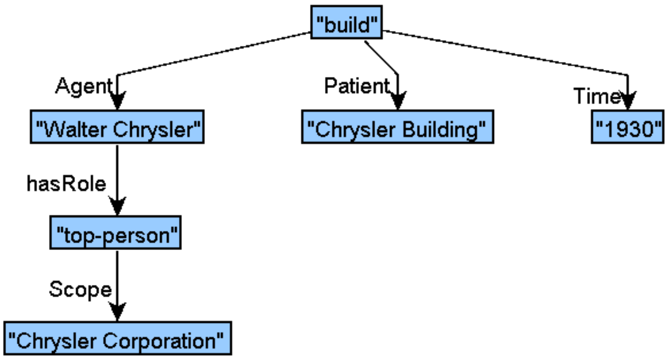


Figure 2: A sample potential conceptual structure

In V4Design, a conceptual structure is seen as an interface between language and knowledge: only content words are relevant (*build*, as opposed to *was* for instance, which only indicates the presence of passive voice), words should be linked to semantically informed knowledge bases (e.g. Walter Chrysler is http://dbpedia.org/page/Walter_Chrysler), words should be generalised (e.g. *head* generalised as *top-person*) and/or clustered (*build* = *construct*), and relations between the words be semantically oriented (Walter Chrysler is the one that performs the action of building –Agent-, he fills the role of head –hasRole-, and this role is defined within a company –Scope).

Since the gap between text and conceptual representations is quite large, the mapping between the two is better addressed as a sequence of steps: (i) morphological analysis, (ii) syntactic analysis, (iii) semantic analysis, (iv) conceptual analysis. The output of each analysis is feeding the next step, and the level of abstractness required in the conceptual structures can be attained gradually.

During the first half of the project, the work on Language Analysis has been focused on:

- making an overview of (i) the annotated corpora, (ii) the lexical resources, and (iii) the available open-source tools that can be useful for the analysis pipeline (all tasks).
- developing new statistical algorithms for multilingual concept candidate detection (T3.2);
- compiling automatically high-quality lexical resources in Spanish (T3.2).
- training statistical modules for the first steps of the analysis (T3.3);
- developing new graph-transduction grammars to be used for multilingual semantic relation extraction between the words of the sentence (T3.3);
- working towards the definition of new evaluation metrics for the dependency parser evaluation (T3.3);
- laying the foundations for the definition of the conceptual model used as interface with the KB (T3.4);
- providing a first evaluation of the modules (all tasks);
- integrating all modules into the V4Design architecture (all tasks).

Due to their tight relation with the summarisation techniques, disambiguation and linking have been addressed mostly in D5.2, submitted on month 16.

The structure of the document is as follows: Section 2 lists the user and technical requirements linked to Language Analysis; Section 3 presents an overview of the existing relevant resources and tools for multilingual analysis; Section 4 contains a description of the tools developed by UPF during the first part of the project and Section 5 presents a first evaluation of these tools; Section 6 shows screenshots of the online demonstrator, and finally Section 7 concludes the document.

# 2 V4DESIGN REQUIREMENTS

In this Section we present the specifications and requirements for the analysis of the textual inputs considered in V4Design. The reported use case and technical requirements for textual analysis are based on D7.2 (*Use cases, requirements and evaluation plan*) and D6.2 (*Technical requirements and architecture*).

The aggregated and detailed user requirements related to Language Analysis are shown in Table 1 and Table 2 respectively, and the three technical requirements that cover the related user requirements are shown in Table 3. A brief discussion of the practical implications of these requirements follows.

Table 1: Aggregated HLUR

| Final High-Level User Requirement (HLUR) | PUC analysis | Final HLUR Title | Final HLUR Description |
|---|---|---|---|
| HLUR_203 | HLUR_1.3<br>HLUR_2.3 | Architectural design tool to form innovative ideas | Architects and designers have a tool that can assist in formulating new, innovative architectural ideas. |
| HLUR_204 | HLUR_1.4<br>HLUR_2.4<br>HLUR_3.1<br>HLUR_4.1 | Multiplicity of assets | Assets can be 3D objects, 2D videos/images, aesthetics, textual information, audio etc. |
| HLUR_207 | HLUR_1.7<br>HLUR_2.7 | Asset accessibility and semantic searching refinement | Architects and designers can have access to a variety of extracted assets and have the ability to filter and refine their search results. |
| HLUR_208 | HLUR_1.8<br>HLUR_2.8<br>HLUR_3.2<br>HLUR_4.4 | Related and suggested assets | Architects/Designers and game developers can have access to a variety of other related or suggested assets to the asset they are working on. |
| HLUR_214 | HLUR_4.6 | Data about the initial asset | Get data about the video that an asset is extracted from. |

Table 2: Analysis aggregation of user requirements

| User Requirement (UR) | Associated HLUR | Detailed description | Functional or Non Functional (FR/N-FR) | Priority based on MoSCoW framework |
|---|---|---|---|---|
| UR_10 | HLUR_203<br>HLUR_208<br>HLUR_214 | As a user I want further details about the acquired footage - image/ video (semantic data/ tags) | FR | MH |

| | | | | |
|---|---|---|---|---|
| UR_11 | HLUR_203<br>HLUR_208<br>HLUR_214 | As a user I want further details about the input footage quality | FR | MH |
| UR_12 | HLUR_203<br>HLUR_208<br>HLUR_214 | As a user I want further details about the extracted data quality | FR | MH |
| UR_13 | HLUR_203<br>HLUR_208<br>HLUR_214 | As a user I want further details about the bounding box of the extracted 3D model (unit independent) | FR | MH |
| UR_14[1] | HLUR_203<br>HLUR_208<br>HLUR_214 | As a user I want further details about the bounding box of the extracted 3D model (unit independent) | FR | MH |
| UR_15 | HLUR_203<br>HLUR_208<br>HLUR_214 | As a user I want further details about geo-location and date/ time of scan | FR | MH |
| UR_16 | HLUR_203<br>HLUR_208<br>HLUR_214 | As a user I want further details about the author and copyrights of the asset | FR | MH |
| UR_17[2] | HLUR_203<br>HLUR_208<br>HLUR_214 | As a user I want further details about the author and copyrights of the asset | FR | MH |
| UR_18 | HLUR_203<br>HLUR_208<br>HLUR_214 | As a user I want further details about visible colours in the asset | FR | CH |
| UR_19 | HLUR_203<br>HLUR_208<br>HLUR_214 | As a user I want further details about related scans of the asset | FR | CH |
| UR_20 | HLUR_204 | As a user I want augmented data of the acquired 3D model (semantic data/ tags) | FR | SH |
| UR_21 | HLUR_204 | As a user I want a description of the acquired 3D model | FR | SH |
| UR_23 | HLUR_204 | As a user I want summarisations of textual content related to the 3D model | FR | SH |
| UR_35 | HLUR_203<br>HLUR_207 | As an architect I want UIX: Search by semantic tags (keywords) | N-FR | MH |
| UR_56 | HLUR_204 | As a game designer I want to get | FR | SH |

[1] This URs is duplicated (same ar UR_13) in the original User requirements.

[2] This URs is duplicated (same ar UR_16) in the original User requirements.

| | | | | |
|---|---|---|---|---|
| | | information about the asset sizes in the video content<br>- To create real life-sized assets for VR 3D environments | | |
| UR_57 | HLUR_204 | As a game designer I want to get information about the assets<br>- Textual and semantic data about the 3D assets<br>- Textual summaries describing the 3D models | FR | SH |
| UR_64 | HLUR_214 | As a game designer I want to be able to get background info about the history of the video content<br>- This will help in the decision of using an asset or not and also give perspective about the video footage used for the asset | N-FR | CH |

Table 3: Technical requirements related to Language Analysis

| TR | Description | Function | Function performed | Related URs |
|---|---|---|---|---|
| TR_LA_1 | Extract knowledge from textual data to be able to map it to the KB | Linguistic Analysis | Tokenisation, Part-of-speech tagging, Lemmatisation, Surface-syntactic parsing | UR_10, UR_11, UR_12, UR_13, UR_14, UR_15, UR_16, UR_17, UR_18, UR_19, UR_20, UR_21, UR_23, UR_35, UR_56, UR_57, UR_64 |
| TR_LA_2 | | Concept extraction | Word Sense Disambiguation, Entity linking | |
| TR_LA_3 | | Relation Extraction | Deep-syntactic parsing, Conceptual relation extraction | |

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. This task consists mainly in linking the entities found in captions and descriptions to the reference DBpedia entry, so that the KB can find additional information on DBpedia and the crawler can retrieve related articles from the web. The textual sources in this case are mainly:
   - **image, painting and video captions:** UR_10, UR_21, UR_23, UR_35, UR_57
   - **video descriptions and museum tweets:** UR_11, UR_12, UR_13, UR_14, UR_15, UR_16, UR_17, UR_18, UR_19, UR_20, UR_21, UR_23, UR_35, UR_56, UR_57, UR_64
2. extract knowledge from associated textual sources (articles, critics, etc.) to discover prominent aesthetic or technical features of the asset. During the first phase of the

project, it has been agreed with the user partners to focus on a few particular aspects: asset name, asset location, asset origin, asset date of construction, asset creator/architect, asset style. This information can usually be found on DBpedia thanks to the linking of the entities; in the event that no DBpedia property can be reached, semantic analysis will be run on the following textual sources to discover the missing information:

- **blog posts:** UR_20, UR_21, UR_23, UR_57, UR_64
- **specialised magazine articles:** UR_20, UR_21, UR_23, UR_57, UR_64
- **Wikipedia pages:** UR_20, UR_21, UR_23, UR_57, UR_64

More details about the different genres can be found in D3.1, submitted on month 12. The three technical requirements are addressed as follows in this deliverable:

- TR_LA1: Section 4.2
- TR_LA2: Section 4.3
- TR_LA3: Sections 4.2 and 4.4

# 3 RELEVANT WORK

In this section, we will compile the available annotated data and lexicons and the state-of-the-art analysis tools in the languages of V4Design (English, Spanish, Greek, German).

## 3.1 Corpora

The following tables compile the state-of-the-art resources in semantic text analysis for the four targeted languages: German, Greek, Spanish and English. Given that UPF has previous experience working on English and Spanish corpora and lexicons, most resources concern German and Greek.

Corpora of annotated sentences are needed in order to train statistical analysers (e.g., part-of-speech taggers, lemmatisers, or syntactic parsers). For all languages, UPF is developing Universal Dependency (UD)-based tools (see Section 4.2.1). However, in case these representations do not allow us to produce fully adequate semantic representations, UPF will resort to alternative resources, as compiled during the first half of the project (UD and best alternative are marked with a grey background in the tables).

Table 4 describes the main characteristics of Greek corpora. Only the GDT is annotated with syntactic and semantic dependencies, and is then our main contingency corpus in Greek.

Table 4: Greek corpora

| GREEK CORPORA | | | | |
|---|---|---|---|---|
| **Name** | **Short description** | **Format** | **Size** | **License** |
| GDT- Greek Dependency Treebank<br>(Prokopidis et al. 2005) | A reference corpus for Modern Greek, annotated at multiple levels: Morphological, syntactic and semantic.<br>The texts include: manually normalised transcripts of European parliamentary sessions, articles from the **Greek Wikipedia** and Web documents pertaining the politics, health, and travel domains. | Dependency-based annotation scheme<br>PDT *fs format | 178,207 tokens<br>7,417 sentences | |
| UD_Greek-GDT<br>(Prokopidis and Papageorgiou, 2017) | The Greek UD treebank is derived from the Greek Dependency Treebank (http://gdt.ilsp.gr), a resource developed and maintained by researchers at the Institute for Language and Speech | CoNLL-X | 61,673 tokens<br>63,441 words<br>2,521 sentences | Creative Commons Attribution-NonCommercial-ShareAlike, CC BY-NC-SA 3.0. |

| | | | | |
|---|---|---|---|---|
| | Processing/Athena R.C. (http://www.ilsp.gr). | | | |
| HNC Hellenic National Corpus<br>(Hatzigeorgiu et al, 2000) | Written texts from several media (books, periodicals, newspapers etc.), which belong to different genres (articles, essays, literary works, reports, biographies etc.) and various topics (economy, medicine, leisure, art, human sciences etc.). | PAROLE format | 47,400,000 words<br>2,462,981 sentences<br>51,179 documents | available over the Internet, for research use only |
| CGT<br>Corpus of Greek Texts<br>(Goutsos 2010) | From radio, television, live, book, telephone, newspaper, magazine, electronic, other.<br>Mixed corpus, including both spoken and written material. | | 30 million words | available and freely accessible online |

Table 5 describes the main characteristics of Spanish corpora. As part of V4Design, UPF develops the AnCora-UPF corpus, which should reach over 10,000 sentences by the end of the project. AnCora-UPF is naturally an annotation resource/corpus which will be used for the V4Design experiments.

Table 5: Spanish corpora

| SPANISH CORPORA | | | | |
|---|---|---|---|---|
| **Name** | **Short description** | **Format** | **Size** | **License** |
| AnCora<br>(Taulé et al, 2008) | Consists mainly of newspaper texts annotated at different levels of linguistic description: morphological (PoS and lemmas), syntactic (constituents and functions), and semantic (argument structures, thematic roles, semantic verb classes, named entities, and WordNet nominal senses). All resulting layers are independent of each other. | CoNLL | 17,680 sentences<br>~500,000 words | freely available from the Web:<br>http://clic.ub.edu/corpus/es/ancora |
| IULA Spanish | A technical corpus of Spanish annotated at surface syntactic level, | Dependency format | 42,000 sentences<br>590,000 tokens | publicly and freely available from the |

| | | | | |
|---|---|---|---|---|
| Treebank (Marimon and Bel, 2015) | following the dependency grammar theory. | | | META-SHARE platform5 with a Creative Commons Attribution 3.0 Unported License |
| AnCora-UPF (Mille and Wanner, 2010) | Following Meaning-Text Theory, MTT (Mel'čuk, 1988), proposes a hierarchical annotation schema that accommodates both fine-grained language-specific dependency structures and a generic picture of abstract dependency relations. | CoNLL | 3,513 sentences<br>~100,000 tokens | freely available from the Web:<br>http://clic.ub.edu/corpus/es/ancora |
| UD-Spanish Ancora (Martínez and Zeman, 2016) | Automatically converted from AnCora. | CoNLL-X<br>uses 17 UPOS tags | 17,680 sentences, 547,681 tokens and 549,570 syntactic words. | GNU GPL 3.0 |
| UD-Spanish-GSD[3] | Automatically converted. | CoNLL-X<br>uses 16 UPOS tags | 16,013 sentences, 423,346 tokens and 431,587 syntactic words. | CC BY-NC-SA 3.0 US |

Table 6 displays the main features of the two reference English corpora we will use:

Table 6: English Corpora

| ENGLISH CORPORA | | | | |
|---|---|---|---|---|
| **Name** | **Short description** | **Format** | **Size** | **License** |
| PennTreebank (Johansson and Nugues, 2007) | Dependency conversion of a constituency treebank, mainly containing Wall Street Journal articles. | CoNLL | ~40,000 sentences<br>~1,000,000 tokens | LDC |
| Universal Dependencies (UD) (Nivre et al, 2016) | Manually revised version of open textual material from electronic journal articles, blogs, etc. | CoNLL-X | ~16,000 sentences<br>~150,000 tokens | GNU GPL 3.0 |

Finally, Table 7 summarises the main resources in German.

[3] http://universaldependencies.org/treebanks/es_gsd/index.html

Table 7: German Corpora

| GERMAN CORPORA | | | | |
|---|---|---|---|---|
| **Name** | **Short description** | **Format** | **Size** | **License** |
| Sense annotated **TüBa-D/Z**<br>The Tübingen Treebank of Written German [4] | Syntactically annotated German newspaper corpus, based on data from the "Die Tageszeitung." All tokens are fully annotated for: inflectional morphology, lemmas, syntactic dependency and constituency, grammatical functions, named entities, anaphora and coreference relations and GERMANET word senses.<br>Partially annotated with discursive connectives. | NEGRA Export format<br>Penn Treebank format v1 and v2<br>Export-XML format (incl. anaphora and coreference relations)<br>CoNLL formats | 3,816 articles:<br>104,787 sentences;<br>1,959,474 tokens | Free of charge for academic purposes |
| TüBa-D/W[5] | A large treebank of modern written German, based on Wikipedia texts.<br>Includes 4 annotation layers with PoS-tags (STTS), lemmas (Tüba-D/Z); morphology (TIGER) and dependency (Tüba-D/Z). | CoNLL-X | 36.1 million sentences<br>615 million tokens | Freely available under a permissive license |
| Sense-annotated **WebCAGe**[6]<br>(Henrich et al, 2018) | A sense-annotated corpus for German, annotated with senses from the German wordnet GermaNet. The corpus is domain-independent. | XML-based format | Number of tagged word tokens 10750<br>Sense-tagged adjectives: 211<br>Sense-tagged nouns: 1499 | Freely available at: http://www.sfs.unituebingen.de/en/webcage.shtml |

[4] https://uni-tuebingen.de/en/faculties/faculty-of-humanities/departments/modern-languages/department-of-linguistics/chairs/general-and-computational-linguistics/resources/corpora/tueba-dz.html

[5] https://uni-tuebingen.de/en/faculties/faculty-of-humanities/departments/modern-languages/department-of-linguistics/chairs/general-and-computational-linguistics/ressources/corpora/tueba-dw/#c616060

[6] https://uni-tuebingen.de/en/faculties/faculty-of-humanities/departments/modern-languages/department-of-linguistics/chairs/general-and-computational-linguistics/resources/corpora/webcage.html

| | | | | |
|---|---|---|---|---|
| | | | Sense-tagged verbs: 897 | |
| TIGER CORPUS v2[7]<br>(v1 still available)<br>(Brants et al, 2004) | A semi-automatically POS-tagged and syntactically annotated corpus from Frankfurter Rundschau german newspaper texts. It also contains morphological and lemma information for terminal nodes. | TIGER XML format<br>NEGRA export text format | 50,000 sentences<br>900,000 tokens | TIGER Corpus license agreement for non-commercial use.<br>Free for research and evaluation purposes |
| SALSA Corpus Release 2.0[8]<br>(Burchardt et al, 2006) | The second version of the German SALSA corpus (based on the TIGER corpus) with role **semantic** annotation in the Berkeley FrameNet paradigm. | XML format | around 24,000 sentences<br>20.000 verbal instances and more than 17.000 nominal instances. | Freely available for academic or educational institution. With license agreement |
| HDT Hamburg Dependency Treebank[9]<br>(Foth et al, 2014) | A largescale corpus of German newscast annotated with dependency relations, morphological information and additional reference specification for relative pronouns. | HDT file format, convertible to the widely used CoNLL-X format, | 261.821 sentences:<br>101,999 manually annotated and checked with DECCA<br>104,795 manually annotated but not checked<br>55,027 automatically parsed sentences<br>~ 4 million hand-annotated tokens | The HDT is free for scientific/academic use.<br>http://hdl.handle.net/11022/0000-0000-7FC7-2 |
| UD German-HDT[10] | An automatic conversion of the HDT, revised manually. | | 173,245 sentences<br>3,055,010 tokens | |
| NEGRA v.2[11] | A German newspaper text corpus from the Frankfurter Rundschau with PoS information and syntactic structures. | a line-oriented export format<br>convertible to<br>Penn Treebank format | 20,602 sentences<br>355,096 tokens | Free license for scientific use |

There are other corpora of plain German texts with possible superficial annotations such as:

[7] https://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/tiger.html

[8] http://www.coli.uni-saarland.de/projects/salsa/

[9] https://corpora.uni-hamburg.de/hzsk/de/islandora/object/treebank:hdt

[10] https://universaldependencies.org/

[11] http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/

- TüPP-D/Z, a partially parsed corpus of written German automatically annotated with clause structure, part-of-speech. It was developed by the University of Tübingen (2004).
- DWDS, a corpus developed in the ZDL (Zentrum für digitale Lexikographie del deutschen Sprache) which aims to be a reference corpus of the 20th century German language. It has a supplementary corpus (DWDS Ergänzungs corpus) consisting essentially of newspaper sources from the last 15 years.
- DEREKO, the Manheim German Reference Corpus of written German, was created in 1964 and it currently contains over 42 billion words (February 2018).
- The Leipziger Korpus-Sammlung (Univ.Leipzig) is a database consisting of 35 million sentences and 500 million running words.

## 3.2 Lexical resources

Good quality lexical resources are needed in order to obtain reliable semantic structures. We aim at identifying descriptions of lexical units that include their government patterns (or subcategorisation frames). That is, how many participants does one unit usually have and how they combine with each other. There is a great variety of lexical resources for a great variety of purposes. We focus on the resources that can be used for language analysis, but also in the context of language generation. Lexicons with more generic semantic information can be very useful, and those that include mappings to standard resources (such as BabelNet, PropBank, or VerbNet for instance) are preferred. The following tables compile the lexical resources relevant to our purposes. Table 8 describes the main characteristics of Greek lexicons.

Table 8: Greek lexicons

| GREEK LEXICONS | | | | |
|---|---|---|---|---|
| **Name** | **Short description** | **Format** | **size** | **license** |
| LEXIS[12]<br>GDT-LEXIS<br>(Papageorgiou et al, 2006)<br>LEXIS-EmotionVerbs<br>(Giouli and Fotopoulou, 2012) | A Greek Computational Lexicon of general language based on corpora, language with **morphological, syntactic and semantic** information.<br>GDT-LEXIS: a lexical resource with semantic information for verbal predicates.<br>LEXIS-Emotion Verbs: details the argument structure, distributional properties and possible transformations of greek emotion verbs. | | Comprises ~60,000 entries with morphological information, of which a subset of 30,000 entries also have syntactic information and a further subset of 15,000 with semantic information.<br>In GDT-LEXIS: about 800 verbs | |

[12] http://www.ilsp.gr/en/infoprojects/meta?view=project&task=show&id=140

| | | | | |
|---|---|---|---|---|
| SKEL<br>(Petasis et al, 2001) | Morphological lexicon that was used to develop a lemmatiser and a morphological analyser that were included in a controlled language checker for Greek. | | ~60.000 lemmas that correspond to ~710.000 different word forms. | |
| Conceptual Lexicon<br>(Fotopoulou et al, 2014) | Encodes morphosyntactic and semantic properties of nominal and verbal multi-word expressions (MWEs). | | ~1000 entries | |
| EKFRASI<br>(Tzortzi and Markantonatou , 2014) | Conceptually organised lexicon encoded with Protégé, Includes conceptual and lexical relations as well as their morphosyntactic properties. | | | |

Table 9 summarises the main characteristics of Spanish lexicons:

Table 9: Spanish lexicons

| **SPANISH LEXICONS** | | | | |
|---|---|---|---|---|
| **Name** | **Short description** | **Format** | **size** | **license** |
| ANCORA_VERB_ES<br>(Aparicio et al, 2008) | Semantic info, subcategorisation, Argumental patterns and thematic roles. Pbank id, Verbnet Id, Framenet id, Wordnet id. | XML | 2,820 verbs | Freely available |
| ANCORA_NOM_ES<br>(Peris and Taulé, 2011) | Deverbal nouns: Denotative type, Wordnet synset, argumental pattern and thematic roles. Link to verb. | XML | 1,658 lemmas | Freely available |
| ANCORANET<br>(Taulé et al, 2011) | Contains the AnCora-Verb lexical entries linked to different English knowledge sources: VerbNet, PropBank, FrameNet, WordNet 3.0 and OntoNotes. | XML | | Freely available |
| ADESSE<br>(García-Miguel et al, 2010) | An online database for the empirical study of the interaction between verbs and constructions in Spanish: Subcategorisation frames, diathesis alternations and syntactic semantic schemes. | | ~4,000 verbs | |
| GLiCOm[13] | Computational lexicon of inflected wordforms in Spanish. The lexicon is distributed in two sublexicons: 1. word forms 2. verb-clitic combinations. | | 1,152,242 word forms, and<br>4,283,637 verb-clitic | Freely available |

[13] https://www.upf.edu/documents/107805982/109136461/tec0128_glicom_tbadia.pdf/07632628-f275-425e-b59c-417433c6a327

| | | | | |
|---|---|---|---|---|
| | | | combinations | |

Table 10 describes the main characteristics of English lexicons, while Table 11 contains the German ones and Table 12 gives details about the multilingual resources.

Table 10: English lexicons

| ENGLISH LEXICONS | | | | |
|---|---|---|---|---|
| **Name** | **Short description** | **format** | **size** | **license** |
| PropBank / NomBank<br>(Kingsbury and Palmer, 2002) / (Meyers et al, 2004) | Subcategorisation frames for verbs and nouns, correspondences between syntactic and semantic roles. | XML | 11,781 disambiguated lemmas | CC BY-SA 4.0 |
| VerbNet<br>(Schuler 2005) | Classification of verbs into 270 semantic classes; Subcategorisation frames, diathesis alternations and syntactic semantic schemes. | XML | 2,380 disambiguated verbs | CC BY-SA 4.0 |
| Framenet[14]<br>(Baker et al, 1998) | English resource based on frame semantics, which models “prototypical situations”with participants and their roles. | XML, HTML | 1,224 frames<br>13,640 lexical units<br>10,542 frame elements<br>1,876 frame-to-frame relations<br>20,229 annotated sentences | |
| WordNet[15]<br>(Fellbaum, 2005) | A large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. | 8 files in ASCII format | 117 000 synsets | Freely and publicly available for download |
| ConceptNet<br>(Speer et al, 2017) | ConceptNet is a multilingual knowledge base, representing words and phrases that people use and the common-sense relationships between them. The knowledge in ConceptNet is collected from a variety of resources, including crowd-sourced resources (such as | JSON-LD a linked data format | 34 million edges<br>(statements) | Freely available with Creative Commons Attribution-ShareAlike license |

[14] https://framenet.icsi.berkeley.edu/fndrupal/

[15] https://wordnet.princeton.edu/

| | | | | |
|---|---|---|---|---|
| | Wiktionary and Open Mind Common Sense), games with a purpose (such as Verbosity and nadya.jp), and expert-created resources (such as WordNet and JMDict). | | | |

Table 11: German lexicons

| GERMAN LEXICONS | | | | |
|---|---|---|---|---|
| **Name** | **Short description** | **format** | **size** | **license** |
| IMSLex German Lexicon[16] | A lexical resource comprising morphological and syntactic information that links together previous resources from IMS-Stuttgart, and covers information on inflection, word formation and valence. Follows the LFG theoretical framework. | XML | 11,000 adjectives<br>1,000 adverbs<br>22,500 nouns<br>300 particles<br>10,000 proper nouns<br>6,000 verbs<br>167 derivation suffixes | Academic research license |
| HaGenLex (HAgen GErmaN LEXicon)[17]<br>(Hartrumpf et al, 2003) | A domain independent computational lexicon with **morphosyntactic** and **semantic** information (based on the MultiNet paradigm, which provides a hierarchy of 45 ontological sorts and a more than 100 semantic relations and functions). | The internal representation of HaGenLex entries makes use of a standard typed feature structure forma lism. Some expanded entries also have XML representations | 12986 noun entries<br>6911 verb entries<br>3278 adjective entries<br>579 adverb entries | Contact the author:<br>Rainer Osswald (rainer.osswald@ fernuni-hagen.de) |
| GermaNet v.14.0[18]<br>(Univ.Tübingen)<br>(Hamp and Feldweg, 1997) | A lexical-semantic net that relates German nouns, verbs, and adjectives semantically by grouping lexical units that express the same concept into synsets and by defining semantic relations between these synsets.<br>Related to Wordnet especially in the case of nouns. | Relational database and XML files | Synsets: 136263<br>Lexical units: 175000<br>Literals: 159359<br>conceptual relations: 150003<br>lexical relations: 12203 (synonymy excluded)<br>Wiktionary sense descriptions: 29549 | GermaNet is free for academic users but you have to sign a license |

[16] https://pdfs.semanticscholar.org/1fee/63d8a6114720653c9e2327188491ccf77a92.pdf

[17] http://pi7.fernuni-hagen.de/research/hagenlex/hagenlex-en.html#HHO03

[18] http://www.sfs.uni-tuebingen.de/GermaNet/

| | | | | |
|---|---|---|---|---|
| BilderNetle[19]<br>(Roller and Schulte, 2013) | A Dataset of German Noun-to-ImageNet Mappings<br>ImageNet is a large-scale and widely used image database, built on top of WordNet, which maps words into groups of images, called synsets. Multiple synsets exist for each meaning of a word.<br>This BilderNetle dataset provides mappings from German noun types to images of the nouns via ImageNet. | | 2,022 word-synset mappings for 309 words | Freely available for education, research and other non-commercial purposes |
| German Subcat Database extracted from MATE Dependency Parses[20]<br>(Scheible et al, 2013) | Induced verb subcategorisation information from German MATE dependency parses, based on the SubCat-Extractor tool.<br>The subcategorisation database is represented in a compact but linguistically detailed and flexible format, comprising various aspects of verb information, complement information and sentence information, within a one-line-per-clause style. | | | The SubCat-Extractor is freely available for education, research and other non-commercial purposes |

Table 12: Multilingual lexicons

| MULTILINGUAL LEXICON | | | | |
|---|---|---|---|---|
| **Name** | **Short description** | **format** | **size** | **license** |
| BabelNet<br>(Navigli and Ponzetto, 2010) | Dictionary with fine-grained senses, definitions and mappings to VerbNet among others. | RDF / HTTP API | 284 languages<br>~6,000,000 concepts<br>10,000,000 named entities | CC BY-NC-ND 4.0 |
| UBY[21]<br>(Gurevych et al, 2012) | A large-scale lexical-semantic resource for natural language processing (NLP) based on the ISO standard LMF. UBY combines a wide range of information from expert-constructed and collaboratively constructed resources for English and German. Currently, UBY integrates resources in English and German by linking | UBY database | | Apart from GermaNet and IMSlex which are licensed under an academic research license,<br>all resources in UBY are available under open licenses, requiring |

[19] https://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/bildernetle.en.html

[20] https://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/subcat-database.en.html

[21] https://dkpro.github.io/dkpro-uby/

| | | | | |
|---|---|---|---|---|
| | them pairwise at the word sense level:<br>English WordNet, Wiktionary, Wikipedia, FrameNet and the syntactically rich VerbNet, German Wikipedia, G_ Wiktionary, GermaNet, IMSLex-Subcat and OmegaWiki. | | | either attribution or both, attribution and share alike |
| lemonUBY[22] | A Semantic Web version of UBY. | lemonUBY (an export of UBY data into lemon derived from UBY.) | | |

For the project, we will use primarily lexical resources that have an explicit mapping to standard English resources such as VerbNet.

## 3.3 Tools for syntactic and semantic parsing

A very large amount of NLP tools have been developed in recent years; most tools are language-agnostic and simply need to be trained on the resources of a desired language. One of the most widely used toolkits is Stanford CoreNLP (Manning et al, 2014), which contains all the basic components needed in an NLP analysis pipeline: sentence splitting, tokenisation, lemmatisation, morphological tagging, coreference resolution, dependency parsing. Other popular toolkits are MATE Tools (Bohnet and Nivre, 2012), developed at the university of Stuttgart, Nlp4J[23] and OpenNLP[24]. We currently use components of these off-the-shelf toolkits, which we trained for our purposes, as shown in Table 13:

Table 13: Off-the-shelf tools used in the V4Design analysis pipeline

| | **English** | **Spanish** | **German** | **Greek** |
|---|---|---|---|---|
| **Segmenter** | Stanford CoreNLP v3.8.0 | Stanford CoreNLP v3.8.0 | OpenNlp Tools v1.8.4 | Stanford CoreNLP v3.8.0 |
| **PoS Tagger** | Stanford CoreNLP v3.8.0 | Stanford CoreNLP v3.8.0 | Stanford CoreNLP v3.8.0 | Stanford CoreNLP v3.8.0 |
| **Lemmatiser** | Mate Tools v3.5 | Mate Tools v3.5 | Mate Tools v3.5 | Mate Tools v3.5 |
| **Morph Tagger** | Mate Tools v3.5 | Mate Tools v3.5 | Mate Tools v3.5 | Mate Tools v3.5 |
| **Dependency parser** | Nlp4J v1.1.3 | Nlp4J v1.1.3 | Nlp4J v1.1.3 | Nlp4J v1.1.3 |

UPF has recently been working on English and Spanish, but has less experience with German or Greek. Table 14 and Table 15 compile some alternative NLP tools for these two languages.

Table 14: NLP tools for Greek

[22] https://www.lemon-model.net/lexica/uby/

[23] https://emorynlp.github.io/nlp4j/

[24] https://opennlp.apache.org/

| GREEK NLP TOOLS | |
|---|---|
| **Name** | **Short description** |
| ILSP | Natural Language Processing services developed by the NLP group of the Institute for Language and Speech Processing: chunker, dependency parser, FBT PoS-tagger, lemmatiser, named-entity recogniser, sentence splitter and tokeniser, transliterator and Wikipedia multilingual domain-related terma and URL lists extractor<br>http://nlp.ilsp.gr/ws/ |
| AUEB | NLP software developed by the Natural Language Processing Research group in the Dept.Informatics of Athens University: PoS-tagger, named entity recogniser and NaturalOWL generator for Greek and English.<br>**http://nlp.cs.aueb.gr/software.html** |
| ELTL | NLP tools: lemmatiser, PoS-tagger, grammatical tagger, VerbTagGr and link to WordNet<br>http://hermes.di.uoa.gr/glosseng.htm |
| LEXISCOPE | A compound language tool that provides information about a Modern Greek word or phrase, combining the functionality of Neurolingo's Hyphenator, Speller, Lemmatiser, Morphological Lexicon and Thesaurus.<br>http://www.neurolingo.gr/en/online_tools/lexiscope.htm |

Table 15: NLP tools for German

| GERMAN NLP TOOLS | |
|---|---|
| **Name** | **Short description** |
| WEBLICHT | An execution environment for automatic annotation of text corpora. Linguistic tools are encapsulated as Web services, which can be combined by the user into custom processing chains. The resulting annotations can then be visualised in a table or tree format.<br>Includes:<br>Sentence splitters<br>Tokenisers<br>Pos-taggers<br>Morphological Analysers and Lemmatisers<br>Syntax-Parsers and Chunkers<br>Word Sense Disambiguation<br>Coreference Resolution and Anaphora<br>Named Entity Recognition<br>Geovisualisations<br>Sentence and Word Aligners<br>https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Main_Page |
| SALT-ATOMIC | SALT is an easily understandable meta model for linguistic data and an open source API for storing, manipulating and representing data. It is an abstract model, poor in linguistic semantics and independent of any linguistic schools or theories. The core model is graph-based, thereby keeping the structural restrictions very low and allowing for a wide range of possible linguistic annotations, such as syntactic, morphological, coreferential annotations and many more. |

| | |
|---|---|
| | http://corpus-tools.org/salt/index.html<br>https://github.com/korpling/salt<br>ATOMIC is a cross-platform multi-layer corpus annotation tool – and extensible platform – for the desktop. ATOMIC is an implementation of SALT, and embeds its complementary conversion framework PEPPER, allowing the mapping between data formats.<br>http://corpus-tools.org/atomic/ |
| IMS-STUTTGART TOOLS | The Institute for Natural Language Processing in the Univ.Stuttgart offers a set of NLP tools for:<br>Coreference (HotCorefDE and CorefAnnotator v 1.9.2)<br>Semantic Parsing (Zubr_2017)<br>An interactive platform (ICARUS) for visualising and querying corpora with dependency syntax, and coreference structures<br>Verb subcategorisation (SubCat-Extractor)<br>Morphological analysis (SMOR, TMV annotator)<br>Deep syntactic parsing (MATE tools)<br>Name Entity Recognition (German NER 2018)<br>Parsing and Treebank Tools (MATE tools)<br>Pos-tagging (MATE, RFT Tagger)<br>Speech tools (IMS-Speech)<br>https://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/index.en.html |
| U.HAMBURG TOOLS | Language Technology Group (LT) has developed NLP tools such as a text segmenter and splitter, a named entity recogniser, a corpus processor and visualiser, a multi-word detector as well as annotation tools and visualisations for entity networks.<br>Some of them are:<br>GermaNer and microNer (German Named Entity Recognition Tool and a micro-service for German Named Entity Recognition)<br>Secos (an unsurpervised compound splitter)<br>LT-ABSA (Aspect-based Sentiment Analysis)<br>JobimText (for Distributional Semantics using lexicalised features)<br>WebAnno (a multi-user tool, fully web-based, for supporting annotators, curators, and a project manager)<br>Lexical chains (for annotation of German texts)<br>Topic Tiling (for text segmentation)<br>https://www.inf.uni-hamburg.de/en/inst/ab/lt/resources/software.html |
| U.SAARLANDES NLP TOOLS | SHALMANESER: a shallow semantic parser<br>Automatic assignment of semantic classes and roles to text. Although the system was developed for Frame Semantics it should be usable for other paradigms (e.g., PropBank roles) with some adaptations. The system is extensively configurable and extendable.<br>SALTO: the SALSA annotation tool<br>FRAMENET TRANSFORMER<br>The FrameNet Transformer is a software tool for deriving customised versions of the FrameNet database based on frame and frame element relations.<br>MAJO – WORD SENSE DISAMBIGUATION and ACTIVE LEARNING TOOLKIT |

| | |
|---|---|
| | http://www.coli.uni-saarland.de/projects/salsa/page.php?id=software |
| DKPRO<br>(UNIV. DARMSTADT & DUISBERG-ESSEN) | The Dkpro-Core is a collection of software components for NLP based on Apache UIMa framework. It provides Apache UIMA components wrapping these tools (and some original tools) so they can be used interchangeably in UIMA processing pipelines.<br>It includes:<br>Spellchecker and Grammar Checker<br>Chunkers, Coreference Resolver<br>Gazeteer<br>Language Identifier<br>Lemmatisers<br>Morphological analysers<br>Named Entity recognisers<br>Parsers<br>Pos-taggers<br>Phonetic transcriptor<br>Semantic role labeller<br>Stemmer<br>Topic model<br>https://dkpro.github.io/dkpro-core/releases/1.10.0/docs/user-guide.html<br>https://dkpro.github.io/ |

## 3.4 Tools for concept extraction, entity linking and word sense disambiguation

There is a vast variety of different methods for information extraction and named entity recognition that are suitable for the concept extraction task solved within the project. The main state-of-the-art models, algorithms, and tools are listed in Table 16. Their peculiarities and drawbacks are highlighted.

Table 16: Concept extraction tools

| | |
|---|---|
| OLLIE | OLLIE – Open Language Learning for Information Extraction (Schmitz et al., 2012) - is an open Information Extraction (IE) system for extracting relational tuples from text, without requiring a pre-specified vocabulary, by identifying relation phrases and associated arguments in arbitrary sentences. It outperforms its strong predecessors REVERB (Fader et al., 2011) and WOEparse (Wu and Weld, 2010) by addressing their limitations – (1) they extract only relations that are mediated by verbs, and (2) they ignore context, thus extracting tuples that are not asserted as factual. First, OLLIE achieves high yield by extracting relations mediated by nouns, adjectives, and more. Second, a context-analysis step increases precision by including contextual information from the sentence in the extractions. OLLIE obtains 2.7 times the area under precision-yield curve (AUC) compared to REVERB and 1.9 times the AUC of WOEparse.<br>Concept extraction is not the primary goal of OLLIE since it aims at detecting relations and not all concepts participate in relations. Therefore some concepts are missed causing a low recall. |
| AutoPhrase | AutoPhrase (Shang et al., 2018) is a framework for automated phrase mining which leverages this large amount of high-quality phrases in an effective way and |

| | |
|---|---|
| | achieves better performance compared to limited human labelled phrases. It is based on positive-only distant training with random forest (Geurts et al., 2006) for phrase classification. It also provides a PoS-guided phrasal segmentation model, which incorporates the shallow syntactic information in part-of-speech (PoS) tags to enhance the performance.<br>The drawback of the approach is that it checks an exhaustive set of n-grams without straight restrictions on possible combinations of PoS-tags and therefore leaves a chance to outcome low-quality phrases in case they add a higher value to the overall score calculated with dynamic programming algorithms. |
| SpaCy NER | SpaCy (Honnibal and Montani, 2017) features a fast statistical entity recognition system, that assigns labels to contiguous spans of tokens. The default model identifies a variety of named and numeric entities, including companies, locations, organisations, and products.<br>As it was trained exclusively on named entities it might miss some real-world concepts that are not names. |
| AIDA | AIDA (Shang et al., 2018) is a framework and an online tool for entity detection and disambiguation. Given a natural-language text or a Web table, it maps mentions of ambiguous names onto canonical entities (e.g., individual people or places) registered in the YAGO2 knowledge base.<br>It mostly focuses on capitalised named entities resulting at high precision with a comparably low recall. |
| FRED | FRED (Gangemi et al., 2017) is a machine reader for the semantic web: its output is an RDF/OWL graph, whose design is based on frame semantics. Nevertheless, FRED’s graphs are domain- and task-independent making the tool suitable to be used as a semantic middleware for domain- or task-specific applications. To serve this purpose, it is available both as a REST service and as a Python library.<br>It detects a hierarchical set of relations between entities that sometimes leads to problems with the processing of long sentences with a large number of entities. |
| DBpedia Spotlight | DBpedia Spotlight (Mendes et al., 2011) is a system for automatically annotating text documents with DBpedia URIs. DBpedia Spotlight allows users to configure the annotations to their specific needs through the DBpedia Ontology and quality measures such as prominence, topical pertinence, contextual ambiguity, and disambiguation confidence. DBpedia Spotlight is shared as open source and deployed as a Web Service freely available for public use. It heavily relies on large gazetteers built on top of entire Wikipedia for interconnecting the Web of Documents with the Web of Data.<br>Integration of an exhaustive number of real-world entities makes this system one of the most competitive on the market. |
| Lample et al., 2016 | Lample (Lample et al., 2016) provides a state-of-the-art named entity recognition model that avoids heavily relying on traditional hand-crafted features and domain-specific knowledge. The model is a neural architecture based on bidirectional LSTMs and conditional random fields that relies on two sources of information about words: character-based word representations learned from the supervised corpus and unsupervised word representations learned from unannotated corpora. It obtains state-of-the-art performance in NER in four languages without resorting to any language-specific knowledge or resources such as gazetteers.<br>As other named entity recognition tools it mostly focuses on capitalised words sometimes missing entities written in lower-case. |
| BERT NER | BERT (Devlin et al., 2018) is a state-of-the-art language representation model, which stands for Bidirectional Encoder Representations from Transformers. Unlike recent language representation models, BERT is designed to pre-train deep |

| | |
|---|---|
| | bidirectional representations from unlabelled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks including named entity recognition without substantial task-specific architecture modifications. At the same time, such universality of the model might lead to weaker results in comparison to models specially designed for a particular task. |
| Flair NER | Flair (Akbik et al., 2019) is a library that provides state-of-the-art natural language processing models including a named entity recognition model. The core idea of the framework is to present a simple, unified interface for conceptually very different types of word and document embeddings. This effectively hides all embedding-specific engineering complexity and allows researchers to "mix and match" various embeddings with little effort. The framework also implements standard model training and hyperparameter selection routines, as well as a data fetching module that can download publicly available NLP datasets and convert them into data structures for quick set up of experiments. The NER model shows an improvement of the results over the best models for many standard datasets. |

Preliminary related work analysis shows the necessity in developing new methods that would address drawbacks of the state-of-the-art tools.

Recent advances in neural machine translation (NMT) encourage solving the task in a supervised manner. Firstly, the copying mechanism in pointer generator networks (See et al., 2017) is one of the most promising techniques naturally applicable to transferring the subsequence of tokens from the analysed text, i.e., a potential concept, into a joint sequence of sequences, i.e., the resulting list of extracted concepts. To make the copied concepts isolated from each other it is enough to train the model to add a termination token between them. Secondly, as is common for NMT bidirectional LSTM models (Luong et al., 2015), they take into account context that is crucial for the defined task since the same combination of words may or may not be a concept depending on a topic of a text and some linguistic peculiarities, e.g., control predicates, that might be caught by «remembering» the surrounding tokens.

The advantage of supervision, in particular, is a possibility to tune the model to a given domain by conditioning the training data. However, in order to obtain a domain-independent model in the context of concept extraction, the labelled data need to cover a vast variety of topics. Taking into account the significant number of examples required to make the concept extraction generic, we opt for distant supervision (Mintz et al., 2009) to obtain a sufficiently large and diverse dataset. This technique consists in automatically labelling the potentially useful data via some handy, not necessarily accurate, algorithm to result in annotation, which is expected to be noisy but, at the same time, contain enough information to train a robust model.

# 4 BASIC TECHNIQUES FOR TEXT ANALYSIS

In this section, we describe the modules for which UPF dedicated most efforts during the first half of the project. These include semantic parsing (Section 4.2), concept candidate detection (Section 4.3) and conceptual-level analysis (Section 4.4). Section 4.1 goes quickly over the previous steps needed for the completion of the processing.

## 4.1 Text pre-processing

Input texts, whatever their source (captions, descriptions, articles, etc.) need to be pre-processed the same way:

- Segmentation: detection of sentence boundaries (if more than one sentence in the input);
- Lemmatisation: prediction of the base form of the words (*building* VS *buildings*);
- PoS tagging: assignment of grammatical categories (*building = NOUN*, *build = VERB*)
- Morphological analysis: detection of morphological features (*builds = build+PRESENT+SINGULAR+3rdPERSON+INDICATIVE*).

The tools used for the pre-processing are listed in Table 13.

## 4.2 Syntactic and semantic parsing

The current NLP analysis pipeline outputs two different types of structures, which correspond to three different levels of abstraction of the linguistic description:

- SSynt: surface-syntactic structures (SSyntSs), i.e., language-specific syntactic trees with fine-grained relations over all the words of a sentence;
- DSynt: deep-syntactic, or *shallow-semantic*, structures (DSyntSs), i.e., language-independent syntactic trees with coarse-grained relations over the meaning-bearing units of a sentence;

This stratified view is strongly influenced by the Meaning-Text Theory -MTT (Mel'čuk, 1988). The MTT model supports fine-grained annotation at the three main levels of the linguistic description of written language: semantics, syntax and morphology, while facilitating a coherent transition between them via intermediate levels of deep-syntax and deep-morphology. At each level, a clearly defined type of linguistic phenomena is described in terms of distinct dependency structures.

In the framework of V4Design, UPF is using primarily a Universal Dependency-based pipeline, which uses similar approaches and tagsets across languages. In order to circumvent possible issues due to the unequal annotation quality of the UD structures, we develop in parallel tools that target a language in particular. During the first half of the project, a Penn Treebank-based pipeline has been setup in English. If the approach shows significantly more efficient than the UD-based pipeline, we will research similar approaches for other languages, using the alternative resources described in Section 3.1.

### 4.2.1 Towards a uniform UD-based pipeline

Universal Dependencies is a generic framework for cross-lingual syntactico-semantic annotation that has been applied to over 80 languages so far, for a total of over 140 different treebanks.[25] Most treebanks have been obtained through automatic conversions of other treebanks, themselves in general obtained via automatic annotation. The resulting annotations are known to lack consistency and quality, but they have the advantage to provide a framework that reduces the differences across different languages. In V4Design, we intend to test the usability of Universal Dependencies as intermediate representations for multilingual relation extraction.

For surface-syntactic parsing, we train the off-the-shelf Nlp4J parser on the freely available UD corpora of the V4Design languages (English, Spanish, German, Greek); see Section 5.1. The resulting surface structures are syntactic trees with lemmas, part-of-speech tags, morphological and dependency information under the form of grammatical functions such as *subject*, *object*, *adverbial*, etc.

The deep structures in this configuration consist of predicate-argument structures obtained through the application of graph-transduction grammars to the UD surface-syntactic structures. The deep and surface structures are aligned node to node. In the deep structures, we aim at removing all the information that is language-specific and oriented towards syntax:

- determiners and auxiliaries are replaced (when needed) by attribute/value pairs, as, e.g., Definiteness, Aspect, and Mood:
    - auxiliaries: *was built-> build*;
    - determiners: *the building-> building*;
- functional prepositions and conjunctions that can be inferred from other lexical units or from the syntactic structure are removed;
    - *built by X-> built X*
- edge labels are generalised into predicate argument (semantics-oriented) labels in the PropBank/NomBank fashion:
    - *subject(built, by X)-> FirstArgument(build, X)*

The UD-based pipeline doesn't make any use of lexical resources at this point; the predicate-argument relations are derived using syntactic cues only. The deep input is a compromise between (i) correctness and (ii) adequacy in a generation setup. Indeed, the conversion of the UD structures into predicate-argument structures depends not only on the mapping process, but also on the availability of the information in the original annotation. Table 17 shows that different labels that the UD-based graph-transduction grammars currently produce.

Table 17: Semantic labels in the output of the UD-based pipeline.

| **Semantic label** | **Type** | **Description** | **Example** |
|---|---|---|---|
| A1/A1INV | Core | 1st argument of a predicate | build-> an architect |
| A2/A2INV | Core | 2nd argument of a predicate | build-> a building |
| A3/A3INV | Core | 3rd argument of a predicate | inaugurate-> on March 15 |
| A4, A5, A6 | Core | 4th to 6th arguments | *Very uncommon* |
| AM | Non-Core | None of governor or dependent are argument of the other | build-> next to the museum |
| LIST | Coordinative | List of elements | built-> and-> inaugurated |
| NAME | Lexical | Part of a name | Chrysler-> Building |
| DEP | UKN | Undefined dependent | N/A |

[25] http://universaldependencies.org/

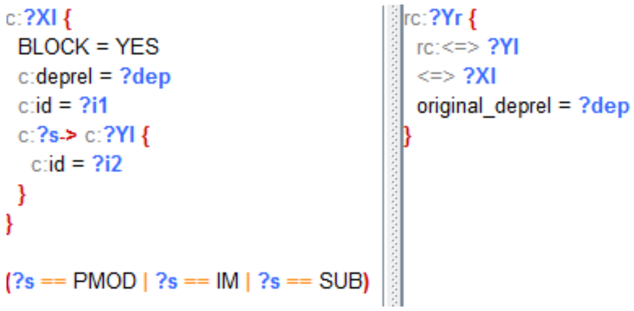
The following phenomena should be highlighted:

- **Alignment between surface and deep nodes**
  On the deep nodes, we use one or more feature ids with attributed as suffix the line number of the corresponding surface nodes: on a deep node, id1=4|id2=15 means that this deep node is aligned with the surface nodes on the lines 4 and 15 of the corresponding surface structure. Only elements triggered by other elements (as opposed to be triggered by the structure of the sentence) are aligned with deep nodes. That is, a subcategorised preposition is aligned with a deep node, while a void copula or an expletive subject are not.
- **Core relations**
  Each defined core relation is unique for each predicate: there cannot be two arguments with the same slot for one predicate. If a predicate has an A2 dependent, it cannot have another A2 dependent, and it cannot be A2INV of another predicate.
- **Auxiliaries**
  Auxiliaries are mapped to the universal feature "Aspect".[26]
- **Conjunctions/prepositions**
  The prepositions and conjunctions maintained in the deep representation can be found under a A2INV dependency. A dependency path Gov-AM-> Dep-A2INV-> Prep is equivalent to a predicate (the conjunction/preposition) with 2 arguments: Gov <-A1-Prep-A2-> Dep.
- **Modals**
  They are mapped to the universal feature "Mood".
- **Pronouns**
  - Relative: only subject and object relative pronouns directly linked to the main relative verb are removed from the deep structure.
  - Subject: a dummy pronoun node for subject is added if an originally finite verb has no first argument and no available argument to build a passive; for a pro-drop language such as Spanish, a dummy pronoun is added if the first argument is missing.
- **Punctuations**
  Only the final punctuations are encoded in the deep representations: the main node of a sentence indicates if the latter is declarative, interrogative, exclamative, suspensive, or if it is involved in a parataxis, with the feature "clause_type".

Our graph-transduction grammars are rules that apply to a subgraph of the input structure and produce a part of the output structure. During the application of the rules, both the input structure (covered by the left side of the rule) and the current state of the output structure at the moment of application of a rule (i.e., the right side of the rule) are available as context. The output structure in one transduction is built incrementally: the rules are all evaluated, the ones that match a part of the input graph are applied, and a first piece of the output graph is built; then the rules are evaluated again, this time with the right-side context as well, and another part of the output graph is built; and so on. The transduction is over when no rule is left that matches the combination of the left-side and the right-side. Consider, for illustration, a sample rule from the SSynt-DSynt mapping in Figure 3. This rule, in which we can see the left-side and the right-side fields, collapses the functional

[26] http://universaldependencies.org/u/feat/index.html

prepositions (*?Xl*, identified during the pre-processing stage with the *BLOCK=YES* attribute/value pair) with their dependent (*?Yl*). That is, a functional preposition such as *by* in *built by Y* is removed from the output structure and made to correspond with the right-side node *Y* (i.e., the dependent).[27] The right-side context is indicated by the prefix *rc:* before a variable or a correspondence; in practice, it means that the rule looks for the *rc:*-marked elements in the current state of the output structure, and builds the elements that are not *rc:*-marked, in this case the correspondence between the right-side *Y* and the left-side *by*, and the new feature *original_deprel*, which stores the left-side incoming dependency relation. A similar rule would apply to *building* and *in* in Figure 4, *in* being the dependent in this configuration; as a result of the application of this rule, only *building* is left in Figure 5, which has a correspondence with both *building* and *in* from Figure 4.



```
c:?Xl {
  BLOCK = YES
  c:deprel = ?dep
  c:id = ?i1
  c:?s-> c:?Yl {
    c:id = ?i2
  }
}

(?s == PMOD | ?s == IM | ?s == SUB)
```

```
rc:?Yr {
  rc:<=> ?Yl
  <=> ?Xl
  original_deprel = ?dep
}
```

Figure 3: A sample graph-transduction rule; *?* indicates a variable; *?Xl{}* is a node, ?s-> is a relation, *a=?b* is an attribute/value pair.

Table 18 sums up the current state of the graph-transduction grammars and rules for the mapping between surface-syntactic structures and UD-based semantic structures.

Table 18: Graph-transduction rules for UD-based deep parsing. *Includes rules that simply copy node features (~40 per grammar)

| Grammars | # rules* | Description |
|---|---|---|
| Pre-processing | 76 | Identify nodes to be removed<br>Identify verbal finiteness and tense |
| SSynt-Sem | 120 | Remove idiosyncratic nodes<br>Establish correspondences with surface nodes<br>Predict predicate-argument dependency labels<br>Replace determiners, modality and aspect markers by attribute-value feature structures<br>Identify duplicated core dependency labels below one predicate |
| Post-processing | 60 | Replace duplicated argument relations by best educated guess<br>Identify remaining duplicated core dependency labels (for posterior debugging) |

[27] Correspondences are meta-information used during the transduction; they are not explicit as such in the output structure. In order to maintain the alignments between surface and deep nodes, attribute/value pairs can be used: e.g. if *by* has a surface identifier “id=2”, and *Y* id = “3”, the deep *Y* node could have to identifiers “id=2,3” to mark the correspondence.

Figure 4 and Figure 5 respectively show a syntactic structure as parsed by the integrated parser and the semantic structure produced by the graph-transduction grammars for the sentence *It is located in the huge HONKA Log Homes building, by Walmart*. *It* is correctly identified as the second argument of *locate* (A2), and the relation between *locate* and *Walmart* is correctly identified as non-core (AM), but no more information is provided at this point (in particular, that *Walmart* is a location in this case); the fact that *HONKA Log Homes* is a named entity is recognised by the pipeline. The relations with the suffix *INV* (e.g. between *huge* and *building*) indicate an inverted core relation between the two elements; their purpose is to maintain a tree format (in which every node has at most one governor), easier to process, as opposed to a graph format (in which a node can have several governors).

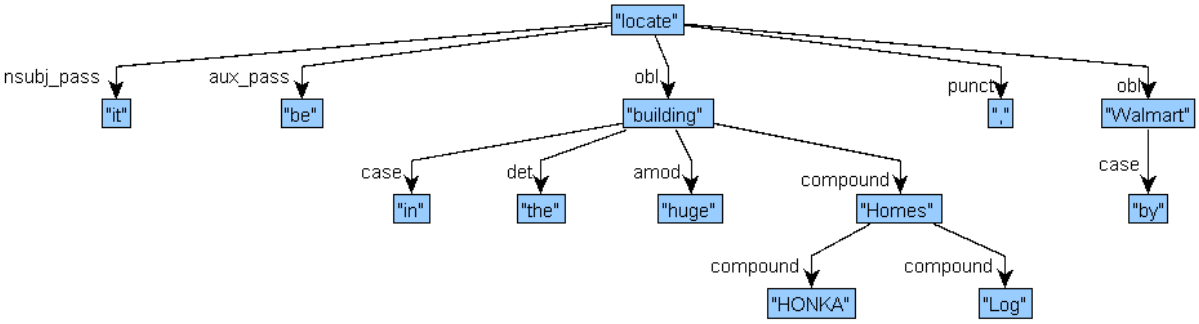


Figure 4: Surface-syntactic UD-Structure: *It is located in the huge HONKA Log Homes building, by Walmart.*

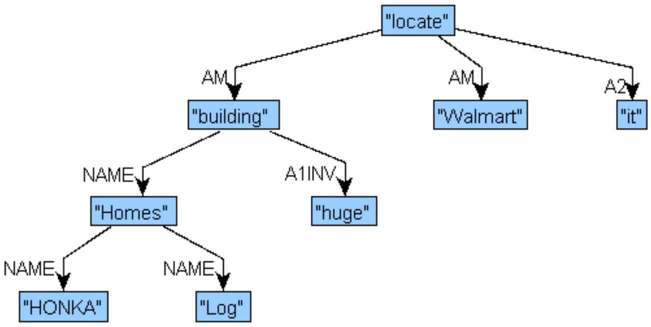


Figure 5: UD-based predicate-argument structure: *It is located in the huge HONKA Log Homes building, by Walmart.*

When the semantic analysis pipeline has to discover the location of a building, simple rules such as the following can apply:

- Find predicate *locate;*
- Find second argument (*A2*): this is the asset;
- Find an AM dependent which had a locative preposition in surface syntax (*building + in*): this is the location of the asset.

The main issue with these deep structures is that they are underspecified, that is, some information is missing; for instance, *building* is not recognised as an argument of *locate,* even though it is one (to be located *somewhere*, in which *somewhere* is the third argument). The rules are thus adapted (relaxed) in order to allow for matching underspecified cases too.

### 4.2.2 Towards language-specific pipelines

For alternative English surface-syntactic (SSynt) annotation, many annotation schemes are available. We chose to use the Penn Treebank representation followed in the CoNLL'09 shared task on dependency parsing, because we believe it is one of the most syntactically sound representations that are available; in particular:

i. Its dependency tagset is fine-grained enough to take into account the most basic syntactic properties of English; unlike the UD-based tagset that is a hybrid syntax/semantics tagset, which does not reach the same level of syntactic fine-grainedness.

ii. One lexeme corresponds to one and only one node in the tree. For instance, in a relative clause, the relative pronoun is viewed from the perspective of its function in the relative clause and not from the perspective of its conjunctive properties.

iii. Unlike in UDs, the subject is a dependent of the inflected top verb, not of the non-finite verb, which might also occur in the sentence. This accounts for the syntactic agreement that holds between the auxiliary and the subject; the relation between the non-finite verb and the subject is more of a "semantic" one, and thus made explicit at a higher level of abstraction. The finite verb in an auxiliated construction is a dependent of the closest auxiliary.

iv. Again unlike UDs, subordinating and coordinating conjunctions depends on the governor of the first group, and governs the one of the second group. This hierarchical approach accounts for the linking properties of conjunctions. The only exception to this are the relative pronouns, as mentioned above.

For the Penn Treebank-based analysis pipeline, we use Bohnet and Nivre's 2012 joint parser and tagger, to which we plugged in another set of graph-transduction grammars. The pipeline currently outputs deep structures at two different levels of representation:

- **DSynt**: deep-syntactic structures (DSyntSs), i.e., syntactic trees with coarse-grained relations over the meaning-bearing units of a sentence;
- **PredArg**: predicate-argument structures (PerdArgSs), i.e., directed acyclic graphs with predicate-argument relations over the meaning-bearing units of a sentence.

**Deep syntactic (DSynt) structures** are dependency structures that capture the argumentative, attributive and coordinative relations between full words (lexemes) of a sentence. Compared to SSynt structures, in DSynt structures, functional prepositions and conjunctions, auxiliaries, modals, and determiners are removed, as in the deep UD structures. Each lexeme is associated with attribute/value pairs that encode such information as part of speech, verbal finiteness, modality, aspect, tense, nominal definiteness, etc. The nodes are labelled with lemmas; in addition, they are aligned with the surface nodes through attribute/ value pairs (each DSynt node points to one or more SSynt node, using the surface IDs). All nodes have a PoS feature, which is copied from the SSynt output. The abstraction degree of the DSynt structures is in between the output of a syntactic dependency parser and the output of a semantic role labeller as the PredArg structures presented below: on the one hand, they maintain the information about the syntactic structure and relations, but, on the other hand, dependency labels are oriented towards predicate/argument relations, and the dependencies directly connect meaning-bearing units, that is, meaning/void/functional elements are not available anymore. Predicate-argument relations include I, II, III, IV, V, VI; modifier relations include ATTR and APPEND (the latter is used for modifiers that generally correspond to peripheral adjuncts); the other two relations are COORD (for coordinations) and NAME (connecting parts of proper nouns). Table 19 summarises the different labels used at this level.

Table 19: Deep-syntactic labels

| | | | |
|---|---|---|---|
| I | Core | 1st argument of a predicate | build-> an architect |
| II | Core | 2nd argument of a predicate | build-> a building |
| III | Core | 3rd argument of a predicate | inaugurate-> on March 15 |
| IV, V, VI | Core | 4th to 6th arguments | *Very uncommon* |
| ATTR | Non- Core | Adjunct | build-> next to the museum |
| COORD | Coordinative | List of elements | built-> and-> inaugurated |
| NAME | Lexical | Part of a name | Chrysler-> Building |
| APPEND | Non- Core | Peripheral adjunct | N/A |

In order to obtain DSynt structures, as for the UD-based pipeline, we run a sequence of rule-based graph transducers on the output of the SSynt parser. But unlike the UD-based grammars, the SSynt-DSynt mapping is based on the notion of hypernode. A hypernode, known as syntagma in linguistics, is any surface-syntactic configuration with a cardinality equal or superior to 1 that corresponds to a single deep-syntactic node. For example, *to report* or *a citizen* constitute hypernodes that correspond to the DSynt nodes *report* and *citizen* respectively. Hypernodes can also contain more than two nodes, as in the case of more complex analytical verb forms, e.g., *would have been reported*. In this way, the SSyntS–DSyntS correspondence boils down to a correspondence between individual hypernodes and between individual arcs, such that the transduction embraces the following three subtasks: (i) hypernode identification, (ii) DSynt tree reconstruction, and (iii) DSynt arc labelling.

Table 20 shows the different steps of the SSynt–DSynt mapping. During a two-step preprocessing, specific constructions and hypernodes are marked. Auxiliaries, meaning-void conjunctions and determiners are easy to identify, but to know which prepositions belong to the valency pattern (subcategorisation frame) of their governor, we need to consult a lexicon extracted from PropBank and NomBank. The output of these preprocessing steps is still a SSynt structure. The third transduction (SSynt-DSynt) is the core of this module: it "wraps" the hypernodes into a single node and manages the labelling of the edges, again looking at the PropBank-based lexicon (i.e., at the valency pattern of the predicates), together with the surface dependencies. For instance, an object of a passive verb is mapped to a first argument (I), while the subject of a passive verb is mapped to a second argument (II). An object introduced by the functional preposition *to* is mapped to a second argument in the case of the predicate *want*, but to the third in the case of *give*, etc. The SSynt-DSynt mapping inevitably produces duplications of core relations, which need to be fixed. The post-processing grammar evaluates the different argument duplications and modifies some edge labels in order to get closer to a correct structure.

Table 20: Graph-transduction rules for deep-syntactic parsing. *Includes rules that simply copy node features (~30% of the rules in each grammar)

| **Grammars** | **# rules*** | **Description** |
|---|---|---|
| Pre-processing 1 | 15 | Assign default PB/NB IDs.<br>Mark passive, genitive, possessive constructions. |
| Pre-processing 2 | 17 | Mark hypernodes. |

| | | |
|---|---|---|
| SSynt-DSynt | 55 | Wrap hypernodes.<br>Assign DSynt dependencies.<br>Transfer aspect/modality as attr.<br>Mark duplicate relations.<br>Mark relative clauses. |
| Post-processing | 78 | Relabel duplicate relations.<br>Re-establish gapped elements.<br>Mark coord. constructions. |

For other languages than English, obtaining good quality lexical resources is also very important. Based on the available lexical resources (see Section 3.2), we started selecting the most relevant ones for each language and converting them into a usable format for the UPF tools; in the process, the data is cleaned and enriched whenever possible. We performed the first iteration of the conversion of the AnCora-Verb (Aparicio et al, 2008) and AnCora-Nom (Peris et al, 2011) lexicons. UPF linguists are still currently curating the outcome of the conversion, and the resource is not finalised yet, nor has it been used in the pipeline so far. We will report on the full conversion in the final WP3 deliverable (D3.5).

## 4.3 Candidate concept detection

We propose generic open-domain out-of-vocabulary (OOV)-oriented concept extraction based on distant supervision with neural sequence-to-sequence learning that takes advantages of the state-of-the-art techniques as described above.

The proposed approach to concept detection consists of the following steps: extraction of a preliminary set of concept candidates from a large collection of texts using a token-cooccurrence frequency-based method developed on earlier stages of the project (cf. its description below) or some large-vocabulary dictionary-based approach; compilation of an input-output dataset and training the neural model to generate lists of concepts for a given fragment of a text; application of the model to unseen texts and determining the positions of concepts in a plain text. The remainder of this section describes details of the method, data used for training neural models, details of training, and the final models chosen for further evaluation on the specific data relevant to the project.

### 4.3.1 Identification of candidate concepts

We propose the method of extracting candidate concepts that is based on the analysis of statistical and linguistic features of sequences of tokens. The algorithm consists of the following steps: 1) determining part-of-speech tags for a given text and selecting potential parts of noun phrases comprising exactly two terms (i.e., “meaningful” words); 2) assessing the distinctiveness of each selected part depending on its position within a list of similar term-cooccurrences (differ by one term) ordered by frequencies; 3) combining intersected highly distinctive parts into concepts and leaving the remainder as separate concepts if they form noun phrases by themselves; 4) applying a statistical NER model to detect additional out-of-vocabulary multi-word expressions; 5) eliminating parts of detected concepts that have an overlap with named entities; 6) compiling an output list of non-overlapped and non-nested concepts including named entities as the result to be used as a target sequence in seq2seq learning.

For the first step, possible part-of-speech patterns for matching complex noun phrases as candidates for parts of concepts have been designed. They include, but are not limited to, the patterns introduced in (Cordeiro et al., 2016). The whole set of patterns $\bigcup c_i$ is the following: N-N, J-N, V-N, N-J, J-J, V-J, N-of-N, N-of-DT-N, N-of-J, N-of-DT-J, N-of-V, N-of-DT-V, CD-N, CD-A, where V is limited only to verbs of types VBD|VBG|VBN. Each pattern is to be used for extracting *n*-grams with two terms and at most two auxiliary tokens in between them that are parts of potential concepts. The concepts are to be formed using combinations of these parts.

The distinctiveness of selected *n*-grams is assessed on word co-occurrences from the Google Books dataset[28]. Given an n-gram $T_1A_1A_2T_2 \in c_k$, where $T_1$ and $T_2$ are terms ("meaningful" words out of stop-word list) and $A_1$ and $A_2$ are optional auxiliary tokens, and $c_k$ is a particular kind of pattern, we use it as a point on a continuous function passing through frequencies of a set of similar n-grams $\bigcup_{c_k} T_1A_1A_2T_j$, taken in a descending order of their frequencies, to find the gradient of decrease of the function. In other words, we check how strong the prominence of n-grams differs from the prominence of their neighbours. In case an *n*-gram is located within a long range of equally prominent *n*-grams we do not consider it as a potential part of a concept as it does not possess the notable distinctiveness inherent in concepts especially in ones with not a direct meaning. The thresholds $Q_{min1}$ and $Q_{min2}$ for a minimum allowed angles of a slope among the sets $\bigcup_{c_k} T_1A_1A_2T_j$ and $\bigcup_{c_k} T_hA_1A_2T_2$ are to be predefined. The $Q_{min1}$ concerns the maximum value of two angles and $Q_{min2}$ – the minimum value of them.

Once the potential parts of concepts are detected we join those that share common tokens and iteratively drop the last token in each grouped sequence of tokens if it is not a noun, in order to end up with the complete noun phrase candidate's concept. Afterwards, we take all nouns and numbers in a text as single-word candidate concepts and drop those that have already been included in the compound candidate concepts.

The described frequency-based criterion for selecting the parts of concepts allows the detection of prominent commonly used compound terms and named entities. At the same time, OOV concepts appear to be out of focus. Some of these concepts, such as novel terms written in lower case, are skipped at this point. They are detected by a trained neural model afterwards. However, the capitalised OOVs might often be caught by a statistical named entity recogniser (Lample et al., 2016, Honnibal and Montani, 2017). Therefore, a state-of-the-art NER model is to be applied with a successive elimination of parts of previously found candidate concepts included in extracted named entities to leave non-nested and non-overlapped n-grams for the outcome.

The above method for labelling large dataset within distant supervision might be substituted by a dictionary-based approach, e.g., DBpedia Spotlight (Mendes et al., 2011). However, dictionary-based approaches appear to be worse in detecting lower-case multi-word concepts. This negatively influences the outcome of training but improves results if used as a compliment as it will be shown in the section devoted to evaluation.

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

### 4.3.2 Training and applying a seq2seq model

Provided necessary characteristics of neural architectures discussed in the section devoted to state-of-the-art, we choose a bidirectional LSTM (Luong et al., 2015) with a copy mechanism (Gu et al., 2016; See et al., 2017) as a model for sequence generation. In particular, we use the biLSTM realisation of the OpenNMT toolkit (Klein et al., 2018) that enables a so-called pointer, which allows copying tokens from the reference text.

The input parts of training examples are subsequent pairs of tokens and their PoS tags, separated by a white space (e.g., '*conceptA NN is VBZ followed VBD by IN the DT second JJ concept NN*'). The target sequence is a list of concepts separated by a special token (e.g., '*conceptA * second concept*'). The sequences are taken from sentences with a sliding overlapping window of a fixed length which are prolonged in case of incomplete candidate concepts at the end.

The trained model is applied to unseen sentences, which are also split into sequences of tokens with an overlapping window of the same size. Finally, determining the positions in a raw text is performed since the output format does not imply including offsets. In particular, we find all possible matches for all detected concepts and then iteratively select non-nested concepts from the beginning to the end of the sentence, giving priority to the longest in case more than one concepts start with the same token.

### 4.3.3 Training and selection the best models on generic data

We used a snapshot of Wikipedia provided by Schenkel et al. (2007) for training several models. The provided semantic annotation was not taken into account. Rather, we used only raw texts of pages and texts of the pointers to other pages as ground truth concepts.

Several subsets were selected from the collection of Wikipedia pages: 220K pages as a dataset to be annotated and used for training, 30K pages for internal deep learning validation steps, 7K pages as a validation set for choosing the best model among several trained with different parameters, and 7K pages as a test set.

The grid search was applied in order to find the best combination of parameters $Q_{min1}$ and $Q_{min2}$ from the three possible angles of a slope corresponding to the different levels of the distinctiveness of a concept described above: $85°, 60°, 0°$. SpaCy NER (Honnibal and Montani, 2017) was used to expand the set of detected candidate concepts with named entities. A separate annotation with DBpedia Spotlight was conducted in order to check if models trained on different annotations might complement each other to improve the outcome.

The training was performed several times with the different amount of resources: networks of two layers with 10K and 20K steps of training and of three layers with 10K and 100K steps with their different checkpoints were tested.

Table 21 presents the best models chosen using the validation set and reached measures of precision, recall, and $F_1$-score on the test set. "DS" stands for distant supervision annotation made either with the dictionary-based approach or with the proposed token-cooccurrence frequency-based method, "$S2S_{(3L,80K)}$" and "$S2S_{(2L,18K)}$" – parameters of the seq2seq models (3 layers, 80K/100K training steps and 2 layers, 18K/20K training steps respectively).

Table 21: Selected models. Precision, recall, and F1-score on generic data

| # | Model | P | R | $F_1$ |
|---|---|---|---|---|
| 1 | $DS_{DICTIONARY}+S2S_{(3L,80K)}$ | 0.7 | 0.72 | 0.71 |
| 2 | $DS_{SLOPE(60,0)}+S2S_{(2L,18K)}$ | 0.67 | 0.77 | 0.72 |
| 3 | (1) + (2) | 0.73 | 0.79 | 0.76 |

The tests on generic data showed that two models trained on differently annotated data applied together improved the *$F_1$-score* by about 10% in comparison to the values obtained by individual models (cf. the third row in Table 21). For fusion the outcomes of the models, we append one list of detected concepts to another and determine the positions of all the concepts in a raw text the same way as it is described in Section 4.3.2.

## 4.4 Towards concept generalisation and conceptual relation extraction

The main advantage of using conceptual structures is to reduce the gap between linguistic structures as provided by language analysis tools and the Knowledge Base (KB) representations, and to make the mapping between the two more generic and reusable. Conceptual structures are by nature abstract and loosely defined; the objective of UPF in V4Design is to achieve the definition of a representation layer that can prove useful for not only language analysis, but also language generation tasks. During the first half of the project, UPF focused on providing a full operational analysis pipeline, and at this point, PredArg structures, similar to the shallow semantic structures described in Section 4.2 are used for the interface with the KB; that is, a mapping to the KB is already in place, but it is not the definitive one.

The work carried out in the framework of Task 3.4 has been centred on four aspects:

1. Definition of the requirements for the conceptual structure.
2. Examination of available resources for cross-lingual concept and conceptual relation generalisation.
3. Setting up of an annotation environment to store reference structures for a set of sentences that contain challenging phenomena.
4. Development of an alternative structure to the targeted conceptual structure.

For (3), an annotation environment has been defined in Brat, and the guidelines for annotating conceptual structures have been started, together with the annotation of a few sentences. For (1) and (2), the outcome is the following:

- A “concept” is seen as single- or multi-word unit that corresponds to an atomic entity or widely used class in the real world. We thus address concepts as “quality phrases” in terms of Liu et al (2015). For instance, “building” is a concept, “tall building” may be a concept, but “blue building with narrow windows” would not be, as it is too specific. The division between what is a concept and what is not a concept is rather blurry, and there are no formal definitions. For this reason, we used large amounts of data in order to predict if a group of words is a candidate concept or not (see Section 4.3).

- A concept needs to be linked to an available multilingual resource which will allow for obtaining the same identifier independently of the language. Two resources are currently usable for this purpose, although none has all the required features: BabelNet and ConceptNet. The concepts in ConceptNet are generic enough, but the main limitation is that they are not disambiguated nor linked to existing lexico-semantic resources. BabelNet however is linked to WordNet and VerbNet for instance, but the entries tend to be very specific, and the same "quality phrase" can end up being tagged with different identifiers depending on the context the original word is used. We intend to solve this topic by the end of the project.
- A concept needs to be generalised, so as to reach the same label independently of the input language or wording within a language. We examined the class hierarchies of WordNet and concluded that they cannot be used in a straightforward way for the generalisation: the hierarchy is actually not a tree, but rather a graph with cycles, and the graph is not balanced in the sense that words have varying numbers of hypernyms. This issue has to be further explored.
- Relations between the concepts need to be generic enough so as to be (i) as language- and linguistics-independent as possible, but (ii) precise enough so as to model the knowledge with an appropriate amount of detail. For instance, simple predicate-argument relations are too generic and linguistic (less than 10 relations, see Section 4.2). FrameNet relations are less linguistic but not generic, since they amount to about 300 different labels (called *frame elements*). On the other hand, VerbNet proposes a hierarchy of about 60 relations which seems to provide a good balance between the two parameters.

For (4), we currently use Predicate-argument (PredArg) structures, which are representations with abstract semantic role labels which also capture the underlying argument structure of predicative elements (which is not made explicit in syntax). This layer of representation is very similar to the deep-syntactic / shallow semantic presented in Section 4.2. The main difference is that lexical units are tagged according to several existing lexico-semantic resources, namely PropBank, NomBank, and VerbNet. The current system is limited to choose the first meaning for each word. During this transition, we also aim at removing support verbs. For the time being, this is restricted to light be-constructions, that is, constructions in which the second argument of *be* in the DSyntS is a predicate P that can have a first argument and that does not have a first argument in the structure. In this case, the first argument of the light *be* becomes the first argument of P in the PredArg representation; for instance, a structure like *building <-I be II-> tall* is annotated as *tall A1-> building*.

The predicate-argument relations are sorted in two subtypes: on the one hand, the argumental, or "core" relations: Argument1, Argument2, Argument3, Argument4, Argument5, Argument6; and, on the other hand, the "non-core" relations: Benefactive, Direction, Extent, Location, Manner, Purpose, Time, NonCore (which is the only underspecified relation). The non-core labels come mainly from the corresponding labels in the Penn Treebank, that is, they are provided by the surface syntactic parser. Table 22 lists the relations used at the PredArg level, and Figure 6 and Figure 7 show a sample sentence representation at the DSynt and PredArg levels respectively.

Table 22: Predicate-argument labels

| **Semantic label** | **Type** | **Description** | **Example** |
|---|---|---|---|
| Argument 1 | Core | 1st argument of a predicate | build-> an architect |
| Argument 2 | Core | 2nd argument of a predicate | build-> a building |
| Argument 3 | Core | 3rd argument of a predicate | inaugurate-> on March 15 |
| Argument4,5,6 | Core | 4th to 6th arguments | *Very uncommon* |
| Benefactive, Direction, Extent, Location, Manner, Purpose, Time | Non-Core | Circumstancials | build-> next to the museum |
| NonCore | Non-Core | None of governor or dependent are argument of the other | the building,-> a hotel |
| NAME | Lexical | Part of a name | Chrysler-> Building |
| Set | coordinative | List of elements | built-> and-> inaugurated |
| Elaboration | Non- Core | Underspecified | N/A |

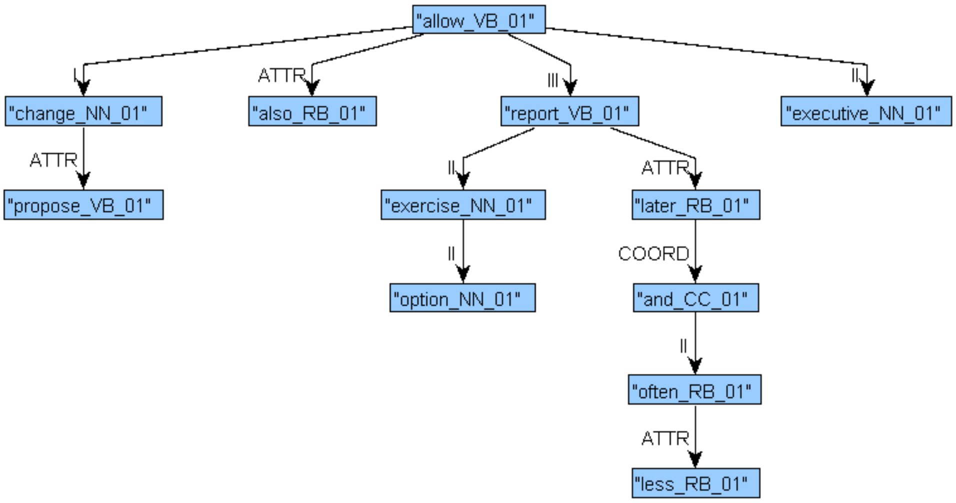


Figure 6: DSynt structure corresponding to the sentence *The proposed changes also would allow executives to report exercises of options later and less often*.

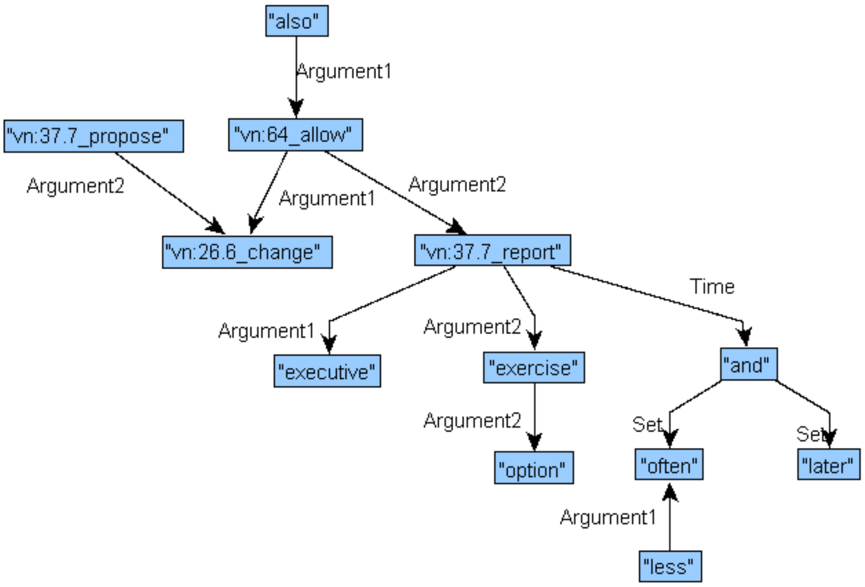


Figure 7: PredArg structure corresponding to the sentence *The proposed changes also would allow executives to report exercises of options later and less often*.

In order to obtain the PredArg structures, we run another sequence of graph-transducers on the output of the DSynt parser. The first grammar in this module creates a pure predicate-argument graph, with the mapping of DSynt relations onto PredArg relations according to PropBank/NomBank. Coordinating conjunctions are linking elements in the Penn Treebank and DSynt representations; in a predicate-argument graph, they are represented as predicates, which have all the conjuncts as arguments and which receive all incoming edges to the coordinated group; see e.g. the representation of *later and less often* in Figure 7, in which *and* receives the incoming edge, instead of *later* in Figure 6. Lexical units are assigned a VerbNet class, as, e.g., *propose* in Figure 7, which is assigned the VerbNet class 37.7. Once this is done, a few post-processing grammars are applied; they recover the shared arguments in coordinated constructions and remove light verbs. The distinction between external and non-external arguments is also removed during this step: for all predicates that have an Argument0 according to PropBank (as it is the case for *allow* and *report* in Figure 6), we push all the arguments one rank up: Argument0 becomes Argument1, Argument1 becomes Argument2, etc. (see the Argument1 relations in Figure 7 for *allow* and *report*). PropBank, NomBank, VerbNet classes are assigned through a simple dictionary lookup. For this purpose, we built dictionaries that can be consulted by the graph-transduction environment and that contain the classes and their members, together with the mappings between them. Table 23 summarises the different steps of this module.

Table 23: Graph-transduction rules for mapping to PredArg structures. *Includes rules that simply copy node features (~30% of the rules in each grammar)

| **Grammars** | **# rules*** | **Description** |
|---|---|---|
| DSynt-Sem | 59 | Assign core dependencies.<br>Recover shared arguments.<br>Establish coord. conj. as predicates.<br>Assign VerbNet classes. |
| Post-processing 1 | 11 | Recover shared arguments in coordinated constructions. |

| | | |
|---|---|---|
| | | Mark light verbs. |
| Post-processing 2 | 23 | Remove light verbs.<br>Assign frames (FrameNet). |
| Post-processing 3 | 30 | Normalise argument numberings. |
| Post-processing 4 | 31 | Introduce non-core dependencies |

# 5 PRELIMINARY EVALUATIONS

There is no annotated reference dataset pertinent to the V4Design domain, so the developed tools were evaluated on an available generic dataset, which contains sources relevant to the PUC (Wikipedia pages, online articles and blogs, etc.).

## 5.1 PoS tagging, syntactic and semantic parsing

In this section, we provide an evaluation of the UD-based and the Penn-Treebank-based analysis pipelines.

For the UD-based pipeline, we report on the evaluation of the Part-of-Speech tagging (Table 24, Table 26, Table 28, Table 30), on the syntactic dependency parsing (Table 25, Table 27, Table 29, Table 31), and on the deep analysis (Table 32). For this, we use the official UD test sets as provided in the CoNLL 2017 shared task.[29] For the evaluation of the deep analysis, we annotated manually about 900 deep tokens (~75 sentences) in English and Spanish at this point. For the dependency relation assignment, we provide both the results of the parser only, using gold standard features, and of the whole pipeline, that is, using the features predicted by the previous modules, in order to reflect the accuracy in a real-life setting.

Table 24: Results of the evaluation of the UD-based PoS tagging (Greek)

| PoS tag | Recall | Precision |
|---|---|---|
| NOUN | 0.9321739 | 0.8355417 |
| PUNCT | 0.9313815 | 1.0 |
| DET | 1.0 | 0.96156085 |
| ADJ | 0.8653396 | 0.6798528 |
| AUX | 0.91150445 | 0.9809524 |
| ADV | 0.9111111 | 0.8523908 |
| PART | 0.9897698 | 1.0 |
| SCONJ | 0.962963 | 0.9017341 |
| VERB | 0.9043977 | 0.87108654 |
| CCONJ | 0.9498681 | 1.0 |
| ADP | 0.7048611 | 0.9902439 |
| PRON | 0.8898129 | 0.9861751 |
| PROPN | 0.48916408 | 0.8586956 |
| X | 0.33070865 | 0.56 |
| NUM | 0.67021275 | 0.9402985 |

Table 25: Results of the evaluation of the UD-based dependency parsing (Greek)

| | UAS | LAS |
|---|---|---|
| Gold PoS, Lemma and Feats | 84.03 | 77.61 |
| Predicted PoS, Lemma and Feats | 74.05 | 64.90 |

Table 26: Results of the evaluation of the UD-based PoS tagging (English)

| PoS tag | Recall | Precision |
|---|---|---|
| PRON | 0.9791474 | 0.97869384 |
| SCONJ | 0.7209302 | 0.8857143 |
| PROPN | 0.7817919 | 0.7143486 |

[29] http://universaldependencies.org/conll17/

| | | |
|---|---|---|
| VERB | 0.9197438 | 0.92217606 |
| ADP | 0.96432114 | 0.9284351 |
| NOUN | 0.9070668 | 0.8336299 |
| PUNCT | 0.9059884 | 0.9918929 |
| CCONJ | 0.9878214 | 0.9945504 |
| ADV | 0.87510204 | 0.9061707 |
| ADJ | 0.86296517 | 0.880651 |
| DET | 0.9836498 | 0.98468846 |
| AUX | 0.9839465 | 0.9558155 |
| PART | 0.9873016 | 0.9242199 |
| NUM | 0.6604478 | 0.8962025 |
| X | 0.23021583 | 0.74418604 |
| SYM | 0.59782606 | 0.9166667 |
| INTJ | 0.75 | 0.9574468 |

Table 27: Results of the evaluation of the UD-based dependency parsing (English)

| | UAS | LAS |
|---|---|---|
| Gold PoS, Lemma and Feats | 83.79 | 79.65 |
| Predicted PoS, Lemma and Feats | 77.63 | 71.51 |

Table 28: Results of the evaluation of the UD-based PoS tagging (Spanish)

| PoS tag | Recall | Precision |
|---|---|---|
| ADJ | 0.908776 | 0.8142783 |
| ADP | 0.9974796 | 0.9975993 |
| DET | 0.9940178 | 0.9811865 |
| PUNCT | 0.8169148 | 0.99980617 |
| NOUN | 0.96965563 | 0.9229462 |
| PROPN | 0.8998785 | 0.8678228 |
| VERB | 0.94282407 | 0.8844734 |
| NUM | 0.8659044 | 0.9731308 |
| CCONJ | 0.99652535 | 0.9944522 |
| PRON | 0.9314602 | 0.94432455 |
| ADV | 0.9456776 | 0.9648391 |
| AUX | 0.91425043 | 0.94095236 |
| SCONJ | 0.91645986 | 0.93977946 |
| SYM | 0.8378378 | 1.0 |
| PART | 0.7222222 | 0.8125 |
| INTJ | 0.53846157 | 1.0 |

Table 29: Results of the evaluation of the UD-based dependency parsing (Spanish)

| | UAS | LAS |
|---|---|---|
| Gold PoS, Lemma and Feats | 85.62 | 80.69 |
| Predicted PoS, Lemma and Feats | 79.65 | 73.73 |

Table 30: Results of the evaluation of the UD-based PoS tagging (German)

| PoS tag | Recall | Precision |
|---|---|---|
| DET | 0.9757384 | 0.9231537 |
| NOUN | 0.8900675 | 0.9233078 |
| AUX | 0.9247911 | 0.8936743 |
| ADP | 0.98506534 | 0.9747537 |
| PUNCT | 0.953508 | 0.99121267 |
| CCONJ | 0.91201717 | 0.9770115 |

| | | |
|---|---|---|
| PRON | 0.8334928 | 0.9385776 |
| ADV | 0.80375296 | 0.9097345 |
| ADJ | 0.84607846 | 0.74589455 |
| NUM | 0.81007755 | 0.9766355 |
| PROPN | 0.8432327 | 0.6339678 |
| VERB | 0.84757507 | 0.87311655 |
| SCONJ | 0.8154762 | 0.9513889 |
| PART | 0.9395349 | 0.9665072 |
| X | 0.04347826 | 0.125 |

Table 31: Results of the evaluation of the UD-based dependency parsing (German)

| | **UAS** | **LAS** |
|---|---|---|
| Gold PoS, Lemma and Feats | 0.84 | 0.78 |
| Predicted PoS, Lemma and Feats | 0.73 | 0.67 |

Table 32: Results of the evaluation of the UD-based deep graph-transduction grammars

| | **LAS** |
|---|---|
| English | 79.83 |
| Spanish | 67.28 |

For the English-specific pipeline, we report the numbers of the MATE tools parser, which assigns jointly lemmas, parts of speech, morphological features, and dependencies (Table 33). For the analysis of the deep analysis, we annotated manually a gold standard of about 300 sentences (5,000 deep tokens); the precision and recall of the hypernode identification (Table 34) and the labelled and unlabelled attachment scores are provided (Table 35). A formal evaluation of the PredArg structures has not been carried out at this point.

Table 33: Results of the evaluation of the PTB-based joint parsing

| | **UAS** | **LAS** |
|---|---|---|
| English | 93.67 | 92.68 |

Table 34: Results of the evaluation of hypernode identification

| | **Precision** | **Recall** |
|---|---|---|
| English | 97.00 | 99.96 |

Table 35: Results of the evaluation of the deep-syntactic graph-transduction grammars

| | **UAS** | **LAS** |
|---|---|---|
| English | 96.74 | 91.24 |

The results of the evaluation of the English-specific pipeline are better than that of the generic UD-based pipeline, but it remains to be seen if this has an actual impact on the overall performance of the analysis component in V4Design.

## 5.2 Candidate concept detection

The goal of the evaluation was to test if the models trained on a generic dataset might be applicable to domain-specific texts relevant to the project.

### 5.2.1 Dataset

We used ten Wikipedia pages with a broad coverage of architectural solutions of different periods and various styles: “Islamic Architecture”, “Romanesque Architecture”, “Neoclassical Architecture”, “Byzantine Architecture”, “Contemporary Architecture”, “Gothic Architecture”, “Modern Architecture”, “Modernism”, “Postmodern Architecture”, and “Functionalism”. We considered texts of pointers to other pages as ground truth concepts. These pointers often share the headings of anchor pages which are in most cases some real-world entities, e.g. “Arthur Heurtley House”, “Price Tower”, etc. They are also sometimes lexical variations of terms behind the link with a correspondingly selected span, e.g., the highlighted pointer in the fragment “the *TWA Terminal* at JFK airport in New York (1956–1962)” leads to the page named *“TWA Flight Center”.*

Most of the texts of pointers are capitalised named entities such as names of buildings, famous people, organisations, historical events, etc. However, there are still some non-capitalised pointers, e.g. “chippendale furniture”, “prefabricated building”, “industrial design” and other that make the task more complex and position it beyond pure named entity recognition. The total number of sentences in the dataset is equal to 3480; the total number of ground truth concepts is equal to 4305.

Similarly, datasets for German and Spanish versions of above-mentioned pages were formed with 500 and 2380 ground truth concepts respectively.

### 5.2.2 Setup of the Experiments

We evaluated values of precision, recall, and $F_1$-score, aiming at high recall first of all. Since the positive ground truth examples are sparse, and there is no negative example, we treated the detected concepts that partially overlapped the ground truth concepts as false positives. This perfectly meets our goal of detecting the correct spans rather than just something close to them. It also allows for penalising brute force high-recall algorithms producing a large number of nested concepts that give no chance for correct text understanding.

In order to compare the obtained models with existing techniques, several efficient, different in nature, entity extraction algorithms were chosen: OLLIE (Schmitz et al., 2012), AIDA (Yosef et al., 2011), AutoPhrase+ (Shang et al., 2018), DBpedia Spotlight (Mendes et al., 2011), SpaCy NER (Honnibal and Montani, 2017) and two deep learning based models (Lample et al., 2016; Delvin et al., 2018). FRED (Gangemi et al., 2017) was not tested as it is not scalable enough for the task: its REST service has a strong limitation on a number of possible requests per day and it fails on processing long sentences (approximately more than 40 tokens). Flair (Akbik et al., 2019) requires a specific environment to be run therefore it is left for further consideration. AutoPhrase+ was combined with StanfordCoreNLP PoS tagger[30] (as it shows better performance with PoS tags) and was trained separately on its default DBLP dataset, and on the above-mentioned Wikipedia training set formed within the project. Its output was slightly modified by removing a defined set of auxiliary tokens from the beginning and the end of the phrase to make it more competitive with the rest of the algorithms. OLLIE’s and SpaCy’s outcomes were also modified the same way that improved

[30] https://github.com/hcl14/AutoPhrase

their performance. DBpedia Spotlight was applied with different values of confidence coefficient equal to 0.5 (default value) and 0.1 to increase the recall.

### 5.2.3 Results of the Evaluation

Table 36 presents reached measures. The sign '*' stands for the modifications made on cutting some first and last words of detected concepts in order to present them as "canonic" noun phrases. The rows in bold correspond to algorithms that include models developed within the project.

Table 36: Concept detection evaluation on domain-specific texts in English

| # | **Model** | **P** | **R** | **F$_1$** |
|---|---|---|---|---|
| 1 | OLLIE* | 0.46 | 0.2 | 0.28 |
| 2 | AutoPhrase$_{wiki}$+* | 0.42 | 0.52 | 0.46 |
| 3 | SpaCy NER | 0.59 | 0.51 | 0.55 |
| 4 | AIDA | 0.76 | 0.57 | 0.65 |
| 5 | SpaCy NER* | 0.71 | 0.61 | 0.66 |
| **6** | **DS$_{SLOPE(60,0)}$** | **0.63** | **0.74** | **0.68** |
| 7 | **DS$_{DICTIONARY}$+S2S$_{(3L,80K)}$** | **0.67** | **0.77** | **0.72** |
| 8 | BERT NER | 0.77 | 0.73 | 0.74 |
| 9 | Spotlight$_{0.1}$ | 0.7 | 0.79 | 0.74 |
| 10 | Lample et al., 2016 | 0.78 | 0.71 | 0.75 |
| **11** | **DS$_{SLOPE(60,0)}$+S2S$_{(2L,18K)}$** | **0.7** | **0.8** | **0.75** |
| 12 | Spotlight$_{0.5}$ | 0.85 | 0.74 | 0.79 |
| **13** | **(7) + (11)** | **0.75** | **0.83** | **0.79** |
| **14** | **(7) +(11)+(9)** | **0.78** | **0.85** | **0.81** |
| **15** | **(7) +(11)+(12)** | **0.78** | **0.85** | **0.81** |
| **16** | **(11) + (12)** | **0.79** | **0.86** | **0.82** |

DBpedia Spotlight applied with its confidence coefficient equal to 0.1 showed significantly better recall than with the default value of 0.5, although F1-score was lower.

Table 36 shows that the proposed models learnt stable templates denoting the positions of concepts and diminished the errors provided by noisy automatic annotation (e.g., compare the 6th row with the 11th).

A combination of two proposed sequential models trained on different annotated datasets that does not rely on external knowledge after training (the 13th row) gives comparable performance (even higher by recall) to the DBpedia Spotlight dictionary-based approach (rows 9 and 12) which is hard to beat as it was applied to the "known" data.

The best seq2seq model, fully developed within the project, used on top of DBpedia Spotlight improves the F$_1$-score by approximately 5% resulting at 0.82 with high recall equal to 0.86 that shows that the developed method is sufficient enough for the purposes of the project. As the experiment was done on detecting of entities already linked to Wikipedia pages, it implies that extracted concepts allow for disambiguating and linking without necessity in correcting their spans.

For German and Spanish texts, candidate concept identification was conducted using the algorithm proposed for annotation within distant supervision. SpaCy NER was run for comparison. Results are provided in Table 37 and Table 38.

Table 37: Concept detection evaluation on domain-specific texts in German

| # | Model | P | R | $F_1$ |
|---|---|---|---|---|
| 1 | SpaCy NER | 0.65 | 0.48 | 0.56 |
| **2** | $\mathbf{DS}_{\mathbf{SLOPE(60,0)}}$ | **0.53** | **0.73** | **0.61** |

Table 38: Concept detection evaluation on domain-specific texts in Spanish

| # | Model | P | R | $F_1$ |
|---|---|---|---|---|
| 1 | SpaCy NER | 047 | 0.36 | 0.41 |
| **2** | $\mathbf{DS}_{\mathbf{SLOPE(60,0)}}$ | **0.33** | **0.46** | **0.38** |

Candidate concept identification for German and Spanish showed results comparable to the results of SpaCy NER as it was for English (cf. Table 36). Thus, the proposed overall approach for concept extraction tested on English has a potential to work efficiently on other languages involved in the project, therefore it will be further applied to them to improve the preliminarily obtained measures.

## 5.3 Towards new metrics for dependency parser evaluation

Syntactic parsing is the first stage of sentence-level analysis in the V4Design pipeline; therefore, it is crucial in order to get good accuracy in the semantic analysis task. For an optimal performance of a downstream application, such as the population of a Knowledge Base, the question on the best parser is thus central. A first round of experiments on parser evaluation for downstream applications has been carried out but the results are inconclusive so far. We will report on the full experiments in the final deliverable, but give a short overview of the idea in the following.

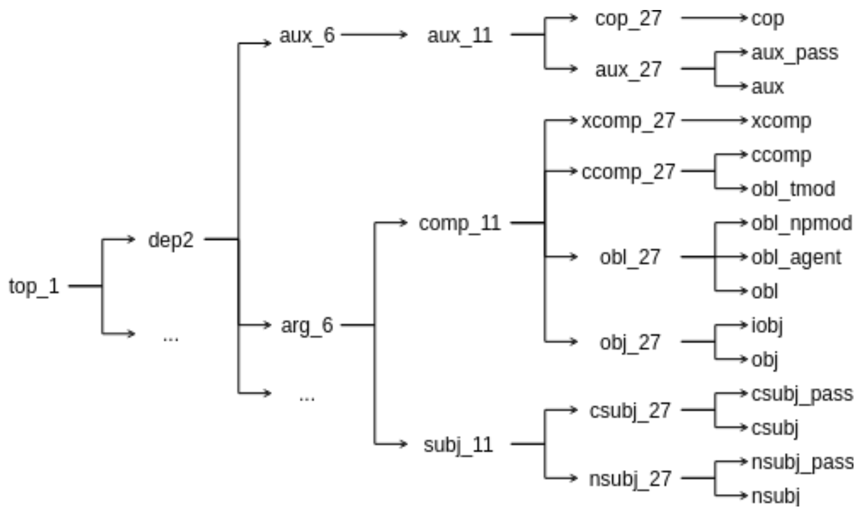
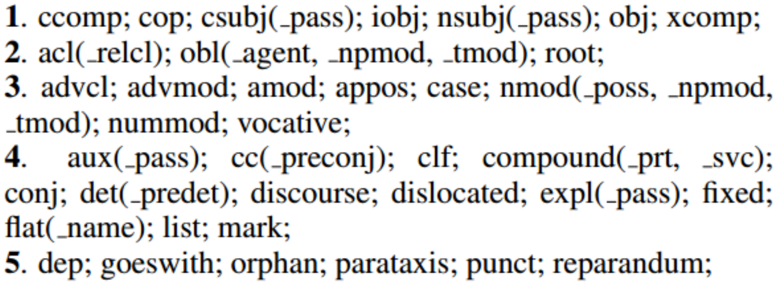


Figure 8: A partial hierarchy of UD relations

The three most commonly used evaluation metrics in dependency parsing are *labelled* and *unlabelled attachment scores* (LAS/UAS), and *label accuracy*. Their adequacy has been questioned, especially when applied across typologically different languages and different formal frameworks, and alternative metrics have been proposed. However, most of the proposed metrics still assess only the correctness of the structural topology of a dependency structure (with or without the morphological or lexical features of its nodes), which is also illustrated by the strong correlation between them in recent works. Furthermore, they count all errors equally, and failures related to the linguistic (including the semantic) interpretation of parsing glitches are not captured.

Research on language resources led to detailed hierarchical (and typed) annotation schemes, where more generic types of relations generalise over more specific ones and the most detailed relations are leaves; cf. a sample hierarchy for Universal Dependencies in Figure 8. In addition, dependencies have been grouped and ranked to reflect their importance in terms of relevant properties; cf., e.g., ’subj’ > ‘dobj’ > ‘iobj’, etc. In a first approximation, the relation labels from detailed typed hierarchies can be grouped as shown in Figure 9 for the UD hierarchy, according to the prominence of the individual relations.



**1**. ccomp; cop; csubj(_pass); iobj; nsubj(_pass); obj; xcomp;
**2**. acl(_relcl); obl(_agent, _npmod, _tmod); root;
**3**. advcl; advmod; amod; appos; case; nmod(_poss, _npmod, _tmod); nummod; vocative;
**4**. aux(_pass); cc(_preconj); clf; compound(_prt, _svc); conj; det(_predet); discourse; dislocated; expl(_pass); fixed; flat(_name); list; mark;
**5**. dep; goeswith; orphan; parataxis; punct; reparandum;

Figure 9: Relevance grouping of UD relations (from most to least prominent)

The two complementary perspectives on the comparison of dependency relations above suggest that:

(1) The difference between relations is not uniform; e.g., a direct and an indirect object are more similar than an object and an adverbial, and it is more harmful to confuse an object with an adverbial than with an object of another type. The actual difference between relations can be captured using the distance between dependencies, as derived from the label location in a balanced typed relation hierarchy (henceforth *distance* hierarchy, $H_{DIST}$). To instantiate $H_{DIST}$, we use the typed hierarchy in Figure 8.

(2) The presence/absence of the relation *A* does not necessarily have the same impact as the presence/absence of the relation *B*; e.g., the absence of the subject is more harmful than the absence of an adverbial. That is, we need to capture the absolute importance of each dependency relation (henceforth *absolute* hierarchy, $H_{ABS}$). To instantiate $H_{ABS}$, we use the five-group ranked list from Figure 9, where each group is assigned a cost according to its prominence. We set these costs in our experiments to 1, 0.65, 0.4, 0.2 and 0.05, but they could also be learned.

We have worked on the definition of experimental metrics that make use of $H_{DIST}$ and $H_{ABS}$, and the numerous experiments carried out did not allow us to reach any solid conclusion. We will thus report on the finalised experiments in the final deliverable (D3.5).

# 6 ONLINE DEMONSTRATOR

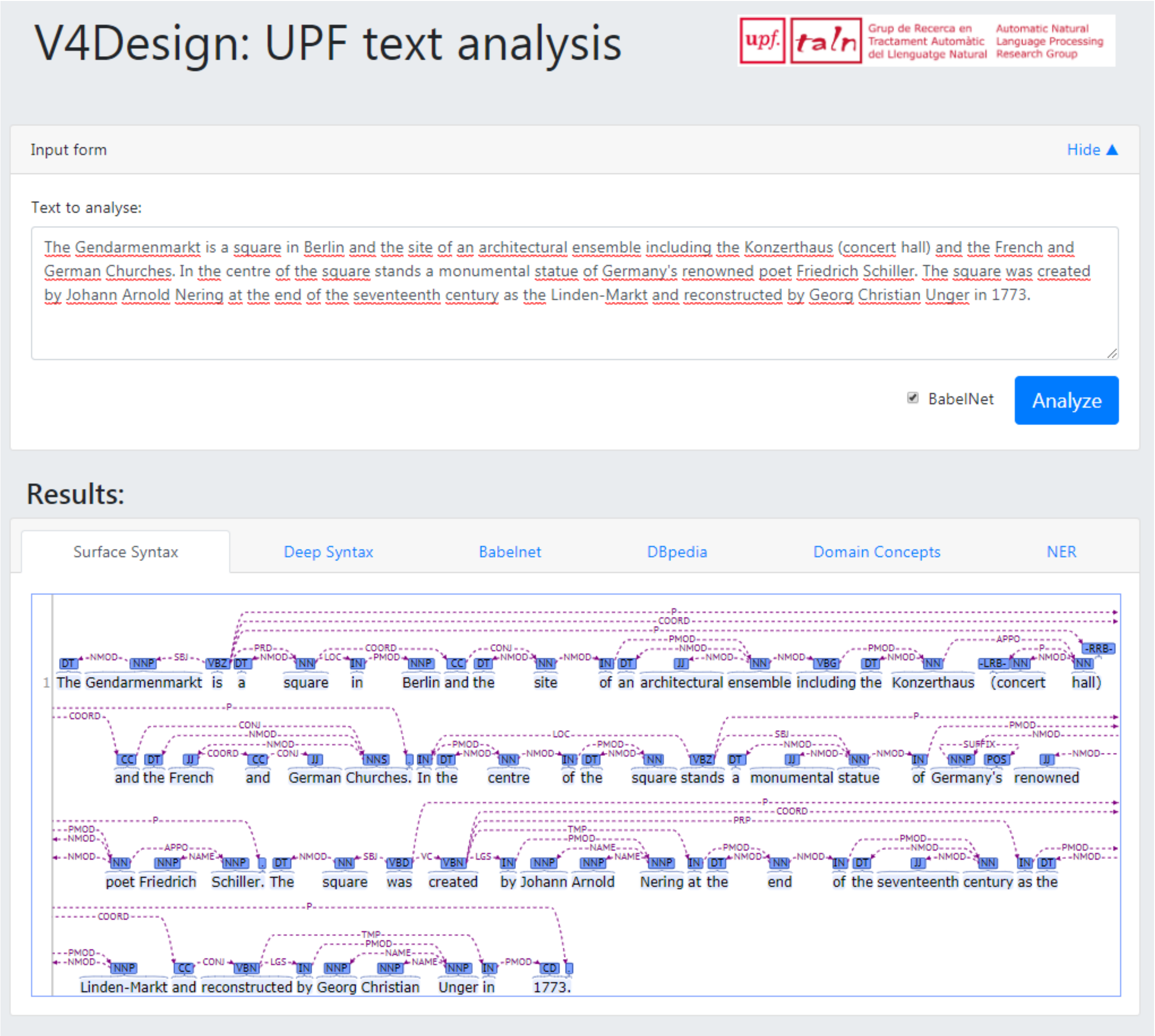
UPF's Language Analysis online demonstrator is available at http://taln.upf.edu/v4design. It performs the following analysis tasks:

- Syntactic analysis: prediction of grammatical relations between all the words of the sentence; this includes segmentation, lemmatisation, PoS tagging, morphological analysis, and syntactic analysis.
- Shallow semantic analysis: prediction of predicate-argument relations between the content words of a sentence.
- Word sense disambiguation: assignment of a sense to a word.
- Entity linking: linking of a word with a DBpedia entry (URI).
- Concept candidate detection: detection of words or groups of words that potentially correspond to an atomic concept.

In the following, we briefly go over each component.

**Surface-syntactic analysis**

- Current languages: English, Spanish
- Next languages: German (ready, to be integrated), Greek
- Current formalisms: Penn Treebank style (English), Universal dependencies (others).
- Tools and resources used: nlp4j v1.1.3 (dependency parsing), Mate Tools v3.5 (lemmatisation, morphological tagging), Stanford CoreNLP v3.8.0 (PoS tagging, segmentation), OpenNlp Tools v1.8.4 (segmentation)



V4Design: UPF text analysis

Input form

Text to analyse:

The Gendarmenmarkt is a square in Berlin and the site of an architectural ensemble including the Konzerthaus (concert hall) and the French and German Churches. In the centre of the square stands a monumental statue of Germany's renowned poet Friedrich Schiller. The square was created by Johann Arnold Nering at the end of the seventeenth century as the Linden-Markt and reconstructed by Georg Christian Unger in 1773.

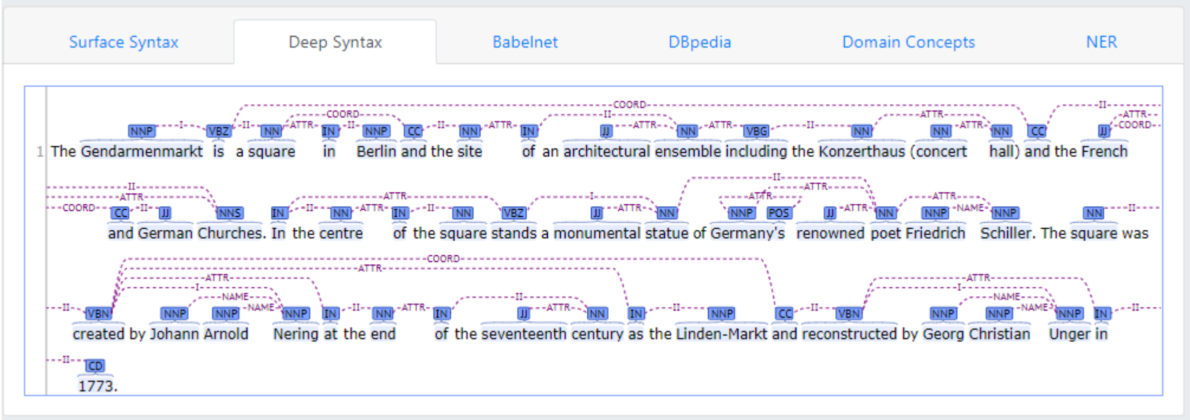
☑ BabelNet

Results:

upf. taln Grup de Recerca en Tractament Automàtic del Llenguatge Natural
Automatic Natural Language Processing Research Group
Hide ▲
Analyze
Surface Syntax
Deep Syntax
Babelnet
DBpedia
Domain Concepts
NER
1 The Gendarmenmarkt is a square in Berlin and the site of an architectural ensemble including the Konzerthaus (concert hall)
and the French and German Churches. In the centre of the square stands a monumental statue of Germany's renowned
poet Friedrich Schiller. The square was created by Johann Arnold Nering at the end of the seventeenth century as the
Linden-Markt and reconstructed by Georg Christian Unger in 1773.
DT NNP VBZ NN IN CC JJ VBG POS NNS VBD VBN CD -LRB- -RRB-
NMOD SBJ PRD LOC COORD CONJ PMOD P APPO NAME SUFFIX TMP VC LGS PRP

### Shallow-semantic analysis

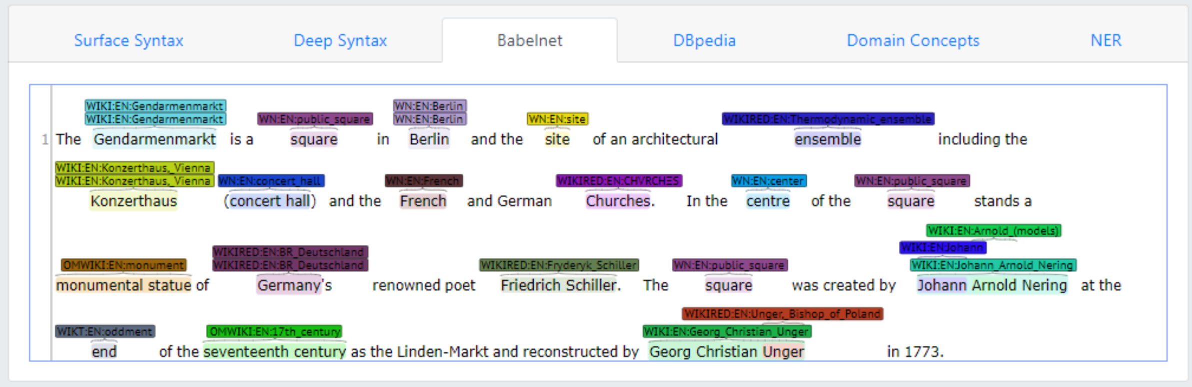
- Current language: English, Spanish
- Next language: Greek (ready, to be integrated) and German (to be developed)
- Current formalism: Meaning-Text Theory, Universal Dependency-based deep structures
- Current level of abstraction: Deep Syntax - language-specific
- Next target for level of abstraction: Conceptual - language-independent (under development)
- Tool and resources used: UPF graph-transduction grammars and lexical resources



1 The Gendarmenmarkt is a square in Berlin and the site of an architectural ensemble including the Konzerthaus (concert hall) and the French and German Churches. In the centre of the square stands a monumental statue of Germany's renowned poet Friedrich Schiller. The square was created by Johann Arnold Nering at the end of the seventeenth century as the Linden-Markt and reconstructed by Georg Christian Unger in 1773.

## Word Sense Disambiguation (part of concept extraction)

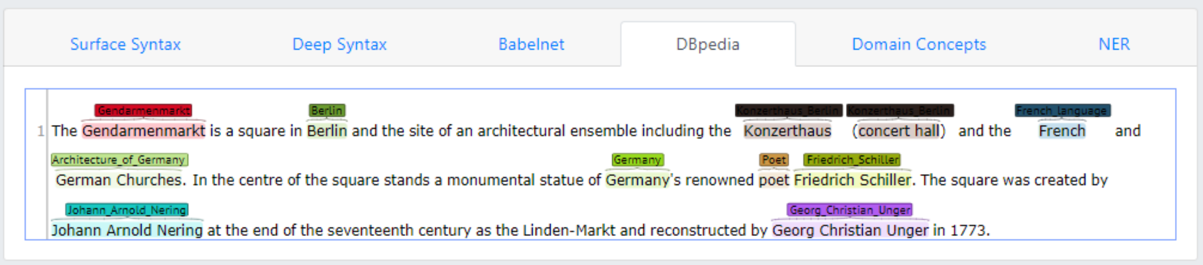
- Current language: English, Spanish
- Next languages: German (ready, to be integrated), Greek
- Current resources used: BabelNet, Wikipedia, WordNet
- Currently using our own BabelNet disambiguation tool fed with our candidate detection (see below)



1 The Gendarmenmarkt is a square in Berlin and the site of an architectural ensemble including the Konzerthaus (concert hall) and the French and German Churches. In the centre of the square stands a monumental statue of Germany's renowned poet Friedrich Schiller. The square was created by Johann Arnold Nering at the end of the seventeenth century as the Linden-Markt and reconstructed by Georg Christian Unger in 1773.

## Entity linking (part of concept extraction)

- Current language: English, Spanish
- Next language: German, Greek
- Current resource used: DBpedia
- Currently using off-the-shelf DBpedia Spotlight fed with our candidate detection (see below)

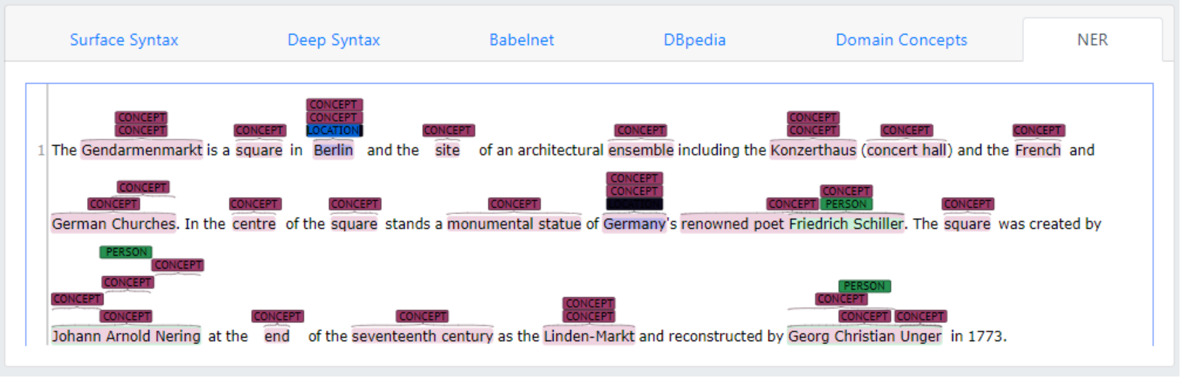


1 The Gendarmenmarkt is a square in Berlin and the site of an architectural ensemble including the Konzerthaus (concert hall) and the French and German Churches. In the centre of the square stands a monumental statue of Germany's renowned poet Friedrich Schiller. The square was created by Johann Arnold Nering at the end of the seventeenth century as the Linden-Markt and reconstructed by Georg Christian Unger in 1773.

## Concept candidate detection (part of concept extraction)

- Current language: English, Spanish

- Next languages: German (ready, to be integrated), Greek
- Tool: UPF concept candidate detection tool



The Gendarmenmarkt is a square in Berlin and the site of an architectural ensemble including the Konzerthaus (concert hall) and the French and German Churches. In the centre of the square stands a monumental statue of Germany's renowned poet Friedrich Schiller. The square was created by Johann Arnold Nering at the end of the seventeenth century as the Linden-Markt and reconstructed by Georg Christian Unger in 1773.

# 7 CONCLUSIONS

In this deliverable, we report on the advances on the Language Analysis components achieved during the first half of the V4Design. This includes (i) an extensive overview of the annotated corpora, the lexical resources, and the available open-source tools that can be useful for the analysis pipeline, (ii) the development and evaluation of new statistical algorithms for multilingual concept candidate detection in English, Spanish and German, (iii) advances on the automatic compilation of high-quality lexical resources in Spanish; (iv) the training and evaluation of statistical modules for sentence segmentation, lemmatisation, part-of-speech tagging, morphological tagging and dependency parsing in English, Spanish, German, and Greek; (v) the development of new graph-transduction grammars to be used for multilingual semantic and conceptual relation extraction in English, Spanish and Greek; (vi) the integration of all aforementioned components into the V4Design architecture; (vii) the preliminary definition of the foundations of the conceptual model used as interface with the KB; (viii) preliminary experiments towards the definition of new evaluation metrics for dependency parser evaluation.

All work has been carried out as planned, and all Language Analysis components are now ready to process all the data as required by the V4Design use cases. During the second half of the project, the work will focus on the following:

- the work carried out in the framework of Task 3.4 will be completed so as to allow for using the targeted conceptual representation in the pipeline;
- given the restrictions due to the previous bullet, the mapping with the KB is currently ad hoc; we will thus define and implement a thorough connection with the KB;
- we will develop the semantic analyser in German, and the concept candidate detection in Greek;
- more properties will be extracted from text.

# 8 REFERENCES

Akbik A, Bergmann T, Blythe D, Rasul K, Schweter S, Vollgraf R. FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP. InProceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations) 2019 Jun (pp. 54-59).

Aparicio, J., M. Taulé, Martí M.A., "AnCora-Verb: A Lexical Resource for the Semantic Annotation of Corpora" Proceedings of 6th International Conference on Language Resources and Evaluation, 2008, Marrakesh, Morocco.

Baker, Collin F., Charles J. Fillmore, and John B. Lowe. "The berkeley framenet project." In Proceedings of the 17th international conference on Computational linguistics-Volume 1, pp. 86-90. Association for Computational Linguistics, 1998.

Bohnet, Bernd and Joakim Nivre. 2012 A Transition-Based System for Joint Part-of-Speech Tagging and Labeled Non-Projective Dependency Parsing. EMNLP-CoNLL, pages 1455-1465, Jeju Island, Korea.

Brants, Sabine, Stefanie Dipper, Peter Eisenberg, Silvia Hansen, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith, and Hans Uszkoreit. 2004. TIGER: Linguistic Interpretation of a German Corpus. Journal of Language and Computation, 2004 (2), 597-620.

Brants, Sabine, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius and George Smith. The TIGER Treebank. In Proceedings of the Workshop on Treebanks and Linguistic Theories, September 20-21, Sozopol, Bulgaria, 2002

Burchardt, K. Erk, A. Frank, A. Kowalski, S. Padó and M. Pinkal, The SALSA Corpus: a German Corpus Resource for Lexical Semantics, Proceedings of LREC 2006, Genoa, Italy

Cordeiro, S., Ramisch, C., and Villavicencio, A. 2016. UFRGS&LIF at SemEval-2016 task 10: Rule-based MWE identification and predominant-supersense tagging. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), pages 910-917.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Fader, Anthony, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In Proceedings of EMNLP.

Fellbaum, Christiane (2005). WordNet and wordnets. In: Brown, Keith et al. (eds.), Encyclopedia of Language and Linguistics, Second Edition, Oxford: Elsevier, 665-670.

Foth, K., Köhn, A., Beuck, N., Menzel, W.: Because size does matter: the Hamburg dependency treebank. In: Proceedings of the Language Resources and Evaluation Conference (LREC 2014), pp. 2326–2333 (2014)

Fotopoulou,A. et al., "Encoding MWEs in a conceptual lexicon.", Proceedings of the 10th Workshop on Multiword Expressions (MWE), 43-47, 2014.

Gangemi, A., Presutti, V., Reforgiato Recupero, D., Nuzzolese, A. G., Draicchio, F., & Mongiovì, M. (2017). Semantic web machine reading with FRED. Semantic Web, 8(6), 873-893.

García-Miguel,J.M. et al.,"ADESSE. A Database with Syntactic and Semantic Annotation of a Corpus of Spanish", Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC), Valletta (Malta), 17-23 de mayo, 2010.

Geurts, P., D. Ernst, and L. Wehenkel. Extremely randomized trees. Machine learning, 63(1):3–42, 2006.

Giouli,V.; Fotopoulou, A.; "Emotion verbs in Greek. From Lexicon-Grammar tables to multipurpose syntactic and semantic lexica", Proceedings of the 15th EURALEX International Congress (EURALEX 2012), Aug 2012, Oslo, Norway.

Goutsos,D., "The Corpus of Greek texts: a reference corpus for modern Greek", Corpora, Volume 5 Issue 1, page 29-44, May 2010.

Gu, J., Lu, Z., Li, H., Li, V. O.: Incorporating copying mechanism in sequence-tosequence learning. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pp. 1631–1640 (2016).

Gurevych, I., J. Eckle-Kohler, S. Hartmann, M. Matuschek, Ch.M. Meyer, Ch. Wirth: UBY – A Large-Scale Unified Lexical-Semantic Resource Based on LMF, in: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL), pp. 580–590, 2012. Avignon, France.

Hamp, B. and H. Feldweg: GermaNet – a Lexical-Semantic Net for German, in: Proceedings of the ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications, pp. 9–15, 1997. Madrid, Spain

Hartrumpf, Sven, Hermann Helbig, and Rainer Osswald: The Semantically Based Computer Lexicon HaGenLex - Structure and Technological Environment. In: Traitement automatique des langues, 44(2), 2003, pp. 81-105.

Hatzigeorgiu, Nick et al. "Design and Implementation of the Online ILSP Greek Corpus." LREC (2000).

Henrich, Verena, Erhard Hinrichs, and Tatiana Vodolazova: WebCAGe -- A Web-Harvested Corpus Annotated with GermaNet Senses. In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012), Avignon, France, April 2012, pp. 387-396.

Honnibal, M., Montani, I.: spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing (2017).

Richard Johansson and Pierre Nugues. 2007. Extended constituent-to-dependency conversion for English. In Joakim Nivre, Heiki-Jaan Kalep, Kadri Muischnek, and Mare Koit, editors, NODALIDA 2007 Proceedings, pages 105–112, Tartu, Estonia. University of Tartu.

Kingsbury, P. and M. Palmer, "From TreeBank to PropBank," in Proceedings of LREC, Las Palmas, 2002.

Klein, G., Kim, Y., Deng, Y., Nguyen, V., Senellart, J., Rush, A. M.: OpenNMT: Neural Machine Translation Toolkit. Proceedings of the 13th Conference of the Association for Machine Translation in the Americas, pp. 177–184 (2018).

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural Architectures for Named Entity Recognition. In Proceedings of NAACL-HLT (pp. 260-270).

Liu, J., Shang, J., Wang, C., Ren, X. and Han, J., 2015, May. Mining quality phrases from massive text corpora. In Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data (pp. 1729-1744). ACM.

Luong, M. T., Pham, H., Manning, C. D.: Effective approaches to attention-based neural machine translation. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 1412–1421 (2015).

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55-60.

Marimon, M.; Bel, N., "Dependency structure annotation in the IULA Spanish LSP Treebank". Language Resources and Evaluation 2015; 49(2): 433-454.

Martínez Alonso, H. ; Zeman,D., "Universal Dependencies for the Ancora Treebanks", Procesamiento del Lenguaje Natural, Revista nº 57, septiembre de 2016, págs. 91-98

McDonald, R., K. Lerman, and F. Pereira. Multilingual Dependency Parsing with a Two-Stage Discriminative Parser Tenth Conference on Computational Natural Language Learning (CoNLL-X) (2006)

Mel'čuk I. Dependency syntax. State University of New York Press, Albany, NY. 1988

Mendes, P. N., Jakob, M., García-Silva, A., & Bizer, C. (2011, September). DBpedia spotlight: shedding light on the web of documents. In Proceedings of the 7th international conference on semantic systems (pp. 1-8). ACM.

Meyers, A. R. R., C. Macleod, R. Szekely, V. Zelinska, B. Young and R. Grishman, "The NomBank project: An interim report," in Proceedings of HLT-NAACL 2004 workshop: Frontiers in corpus annotation, Boston, 2004.

Mille, S.; Wanner, L., "Syntactic Dependencies for Multilingual and Multilevel Corpus Annotation", Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), 2010, Valetta, Malta

Mintz, Mike, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In Proceedings of ACLAFNLP, pages 1003–1011.

Navigli R. and S. Ponzetto, "BabelNet: Building a very large multilingual semantic network," in Proceedings of the 48th annual meeting of the association for computational linguistics, Uppsala, 2010.

Nivre, Joakim, et al. "Universal Dependencies v1: A Multilingual Treebank Collection." In Proceedings of LREC, 2016.

Papageorgiou,H et al., "Adding multi-layer semantics to the Greek Dependency Treebank", LREC 2006, Genoa, Italy.

Peris, A.; Taulé,M., "AnCora-Nom: A Spanish lexicon of deverbal nominalizations", Procesamiento del Lenguaje Natural, Vol. 46:11-18, 2011.

Petasis, G. et al., "A Greek morphological lexicon and its exploitation by natural language processing applications", 2001, Panhellenic Conference on Informatics, pp 401-419

Prokopidis et al,. Theoretical and Practical Issues in the Construction of a Greek Dependency Treebank. In Montserrat Civit, Sandra Kubler, and Ma. Antonia Marti, editors, Proceedings of The Fourth Workshop on Treebanks and Linguistic Theories (TLT 2005), pages 149-160, Barcelona, Spain, December 2005. Universitat de Barcelona.

Prokopis Prokopidis and Harris Papageorgiou. Universal Dependencies for Greek. In Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017), pages 102-106, Gothenburg, Sweden, May 2017.

Roller, Stephen, Sabine Schulte im Walde (2013). A Multimodal LDA Model integrating Textual, Cognitive and Visual Modalities. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). Seattle, WA.

Scheible, Silke, Sabine Schulte im Walde, Marion Weller, Max Kisselew: A Compact but Linguistically Detailed Database for German Verb Subcategorisation relying on Dependency Parses from a Web Corpus: Tool, Guidelines and Ressouce. In: Proceedings of the 8th Web as Corpus Workshop. Lancaster, UK, July 2013.

Schenkel R, Suchanek FM, Kasneci G. YAWN: A Semantically Annotated Wikipedia XML Corpus. InBTW 2007 Mar (Vol. 103, pp. 277-291).

Schmitz, M., Bart, R., Soderland, S., & Etzioni, O. (2012, July). Open language learning for information extraction. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (pp. 523-534). Association for Computational Linguistics.

Schuler, K.. VerbNet: A broad-coverage, comprehensive verb lexicon, Univeristy of Pennsylvania, 2005.

See, A., Liu, P. J., Manning, C. D.: Get to the point: Summarization with pointer-generator networks. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pp. 1073–1083 (2017).

Shang, J., Liu, J., Jiang, M., Ren, X., Voss, C. R., & Han, J. (2018). Automated phrase mining from massive text corpora. IEEE Transactions on Knowledge and Data Engineering, 30(10), 1825-1837.

Speer, Robyn, Joshua Chin, and Catherine Havasi. 2017. "ConceptNet 5.5: An Open Multilingual Graph of General Knowledge." In proceedings of AAAI 31.

Taulé, M., M.A. Martí, M. Recasens "Ancora: Multilevel Annotated Corpora for Catalan and Spanish",Proceedings of 6th International Conference on Language Resources and Evaluation, 2008, Marrakesh (Morocco).

Taulé, M. et al., "AnCora-Net: Integración multilingüe de recursos lingüísticos semánticos", Procesamiento del Lenguaje Natural, Vol. 47:153-160, 2011.

Tzortzi,K.; Markantonatou,S., “Development of a Conceptual Lexicon with ontological techniques”, Proceedings of Terminology and Ontology: theories and applications TOth 2014, Chambéry, France.

Wu, Fei and Daniel S. Weld. 2010. Open information extraction using Wikipedia. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL ’10).

Yosef MA, Hoffart J, Bordino I, Spaniol M, Weikum G. Aida: An online tool for accurate disambiguation of named entities in text and tables. Proceedings of the VLDB Endowment. 2011;4(12):1450-3.

# APPENDIX

## User requirements related to Language Analysis by use case

Table 39: HLUR extracted from PUC1

**PUC1**

| HLUR | HLUR Title | HLUR Description |
|---|---|---|
| HLUR_1.3 | Architectural design tool to form innovative ideas | Architects and designers have a tool that can assist in formulating new, innovative architectural ideas |
| HLUR_1.4 | Multiplicity of assets | Assets can be 3D objects, 2D videos/images, textual information, audio etc. |
| HLUR_1.7 | Asset accessibility and searching refinement | Architects and designers can have access to a variety of extracted assets and have the ability to filter and refine their search results. |

**Table 40**: HLUR extracted from PUC2

**PUC2**

| HLUR | HLUR Title | HLUR Description |
|---|---|---|
| HLUR_2.3 | Architectural design tool to form innovative ideas | Architects and designers have a tool that can assist in formulating new, innovative architectural ideas |
| HLUR_2.4 | Multiplicity of assets | Assets can be 3D objects, 2D videos/images, textual information, audio etc. |
| HLUR_2.7 | Asset accessibility and searching refinement | Architects and designers can have access to a variety of extracted assets and have the ability to filter and refine their search results. |
| HLUR_2.8 | Related and suggested assets | Architects/Designers and game developers can have access to a variety of other related or suggested assets to the asset they are working on. |

Table 41: HLUR extracted from PUC3

**PUC3**

| HLUR | HLUR Title | HLUR Description |
|---|---|---|
| HLUR_3.1 | Multiplicity of assets | Assets can be 3D objects, 2D videos/images, textual information, audio etc. |
| HLUR_3.2 | Related and suggested assets | Game developers can have access to a variety of other related or suggested assets to the asset they are working on. |

Table 42: HLUR extracted from PUC4

**PUC4**

| HLUR | HLUR Title | HLUR Description |
|---|---|---|
| HLUR_4.1 | Multiplicity of assets | Assets can be 3D objects, 2D videos/images, textual information, audio etc. |

| | | |
|---|---|---|
| HLUR_4.4 | Related and suggested assets | Game developers can have access to a variety of other related or suggested assets to the asset they are working on. |
| HLUR_4.6 | Data about the initial asset | Get data about the video that an asset is extracted from |