

# UNIVERSITÀ DEGLI STUDI DI PAVIA

## DIPARTIMENTO DI STUDI UMANISTICI

## CORSO DI LAUREA MAGISTRALE IN LINGUISTICA TEORICA, APPLICATA E DELLE LINGUE MODERNE

## BUILDING A MULTILINGUAL OUTLIER DETECTION DATASET FOR THE EVALUATION OF DISTRIBUTIONAL THESAURI AND WORD EMBEDDINGS

RELATRICE

Prof.ssa Elisabetta Jezek

CORRELATORE

Ph.D. Miloš Jakubíček

Tesi di Laurea Magistrale di Emma Romani Matricola n. 484596

Anno Accademico 2021/2022

# Index

Introduction
Part 1. A Theoretical Introduction to Thesauri, Distributional Models and Evaluation $\dots 1$
Chapter 1. Thesauri: Semantic Relations, History, Applications, and Construction
1.1 Lexical Semantics and Semantic Relations1
1.1.1 Lexical Semantics
1.1.2 Semantic relations
<b>1.2 Definitions of Thesaurus, History of its Development, and its Applications</b>
1.2.1 A definition of Thesaurus
1.2.2 Thesauri and dictionaries
1.2.3 Thesauri's structure24
1.2.4 History of Thesauri and Roget's Thesaurus2
1.2.5 WordNet
1.2.6 Automatic Thesauri
1.2.7 Information Retrieval Thesauri
1.3 Automatic Distributional Thesaurus
1.3.1 Distributional hypothesis
1.3.2 Similarity and relatedness
1.3.3 Word Space Models
1.3.4 An overview on distributional Thesauri construction
1.4 Sketch Engine Distributional Thesaurus and Word Embeddings
1.4.1 A Distributional Thesaurus: Sketch Engine Thesaurus
1.4.2 Word Embeddings as Distributional Thesauri
1.4.3 How to assess the quality of automatic thesauri and to compare different approaches the issue of evaluation
Chapter 2. The Issue of the Evaluation and How to Evaluate Distributional Models 4
2.1 The Evaluation in General: Why and How to Evaluate
2.1.1 Human evaluation and Inter-Annotator Agreement

2.1.2 Evaluating the model against the gold standard: accuracy, precision, recall
<b>2.2 Extrinsic vs. Intrinsic Techniques for the Evaluation of Distributional Models</b> 55
2.2.1 Extrinsic evaluation: an overview of the techniques and issues
2.2.2 Intrinsic evaluation: an overview of the techniques and issues
2.2.3 How to correlate intrinsic and extrinsic evaluation methods
2.2.4 Issues with the intrinsic evaluation: how can it be improved?
<b>2.3</b> The Outlier Detection Task as a Technique for the Intrinsic Evaluation of Distributional Thesauri and Word Embeddings
2.3.1 Camacho-Collados & Navigli (2016): the first proposal and the task in brief
2.3.2 On the outlier detection task: a literature review
Part 2. HAMOD Dataset: History, Methodology, and Building Process
Chapter 3. HAMOD: a High Agreement Multilingual Outlier Detection Dataset
<b>3.1 Motivation, Purposes, and Background Study</b> 76
3.1.1 What HAMOD stands for and what it tells about its purposes
3.1.2 In what HAMOD differs from 8-8-8 dataset and why it improves it
<b>3.2 Dataset Formal Description</b> 79
3.2.1 Definition of the sets: semantic categories, topics, inliers, outliers
3.2.2 Formal requirements for the encoding of the dataset and format
<b>3.3 The Original Dataset: its Development, Extent, and What Needed to be Improved</b> 86
<b>3.4 HAMOD dataset: the Outcome</b>
Chapter 4. Methodology: How to Build and Implement HAMOD
4.1 First Step: How to Create New Sets
4.1.1 T-PAS ontology and the Semantic Types 105
4.1.2 Wikipedia lists and categories108
4.1.3 Verb sets: using Levin (1993) semantic classes for verb sets
4.2 Second Step: How to Translate and Adapt the Sets into other Languages
<b>4.3 Third Step: Do the Words Belong to a Basic Vocabulary?</b> 119
4.4 Fourth Step: How to Further Implement the Dataset, that is, the Guidelines 124
4.2.1 Guidelines for Clusters Creation

4.2.2 Guidelines for HAMOD dataset Creation and Translation	
Part 3. An Experiment with HAMOD Dataset	
Chapter 5. Evaluation of Distributional Thesauri and Word Embeddings throug Dataset	
5.1 Experiment Setup and Hypotheses on the Results of the Experiment	
5.1.1 The experiment setup: steps and requirements	
5.1.2 Hypotheses on the results of the experiment	
5.2 Human Evaluation	
5.2.1 The online interface for the experiment	
5.2.2 Selection and training of participants	
5.2.3 Metrics for the human evaluation: raw agreement	140
5.3 Evaluation of Word Embeddings and Sketch Engine Distributional Thesa	<b>urus</b> 141
5.3.1 Data and models	
5.3.2 Metrics for the evaluation of distributional models: accuracy and OPP	
5.3.3 Calculating the results	
Chapter 6. Quantitative and Qualitative Analysis of the Results of the Expe Further Improvements	
6.1 Quantitative Analysis: Human Evaluation	
6.2 Quantitative Analysis: Word Embeddings and Sketch Engine Thesaurus	
6.3 Qualitative Analysis: Human Evaluation	
6.4 Qualitative Analysis: Word Embeddings and Sketch Engine Thesaurus E	valuation
6.5 Human vs. Distributional Models' Performance	174
6.5.1 Quantitative analysis	
6.5.2 Qualitative analysis	
6.6 Final Remarks	
Conclusions	
Bibliography	
Websites	

Appendix 1. HAMOD Dataset	
Appendix 2. Guidelines for HAMOD Dataset Creation and Translation	
1. Dataset Creation	
1.1. Selection of the Semantic Categories	
1.2. Inliers Selection	
1.3. Outliers Selection	
1.4. Formal Requirements for the encoding of the dataset	
1.5. Format	
2. Translation and Adaptation of the Dataset	
Appendix 3. Quantitative Results of the Evaluation of Distributional Models .	
Acknowledgments	
Ringraziamenti	

# Introduction

In September 2021 I had the chance to undertake an Erasmus+ Traineeship at Lexical Computing s.r.o. in Brno (Czech Republic). During that five-months experience, I conducted this experimental thesis. Lexical Computing is a research company in the field of Corpus and Computational Linguistics, which implements a corpus query system, Sketch Engine, containing hundreds of corpora and various tools for linguistic analysis on these corpora. On some of these tools, namely the Sketch Engine distributional Thesaurus and Word Embeddings, we developed this thesis project.

The project consisted in the implementation of a multilingual dataset for the outlier detection task (HAMOD dataset),<sup>1</sup> which could then be used for an intrinsic evaluation of the tools mentioned above. The *outlier detection task* (originally proposed by Camacho-Collados & Navigli (2016), this thesis' background study) is an evaluation methodology that can be described as follows: given a set of words, the task is to find a "word that does not fit" (the *outlier*) to the others, instead, semantically related words.

This methodology was conceived to evaluate distributional models – such as those at the basis of the Sketch Engine distributional Thesaurus and the Word Embeddings. Distributional thesauri are lists words with similar or related meanings ranked according to some similarity measures. Word Embeddings are vector representations of words in a semantic space;<sup>2</sup> word's distance or proximity in the semantic space can be calculated according to similarity measures – the proximity in semantic space is also believed to be proximity (i.e., similarity or relatedness) in meaning.

The need to evaluate thesauri and embeddings quality can be explained in the automatic approach to their computation. While in the past thesauri were manually compiled by linguists and lexicographers, nowadays automatic procedures are preferred. As for word embeddings, extrinsic techniques (on downstream NLP tasks) are more often pursued and usually word embeddings perform well on the majority of these tasks. We believed that intrinsic methodologies, instead, could be worthier to investigate, in order to gain more insights on the nature of distributional models and to conduct linguistic analyses on the outputs of the task.

<sup>&</sup>lt;sup>1</sup> HAMOD dataset stands for "High-Agreement Multilingual Outlier Detection" dataset.

<sup>&</sup>lt;sup>2</sup> See Harris (1957), the distributional hypothesis.

There are several intrinsic evaluation techniques which can be pursued in order to evaluate distributional models, but one of the major strengths of – and the reason why we chose – the outlier detection task is that, in the intrinsic evaluation process, human evaluators performing the task on the dataset provide a high level of inter-annotator agreement, thus making this dataset a gold standard against which it is possible to evaluate distributional models. This is relevant in comparison with other kinds of intrinsic evaluation methods, such as word similarity and word analogy, which generally provide lower performances in the human evaluation in terms of Inter-Annotator Agreement.<sup>3</sup>

In order to perform the outlier detection task on distributional models, we implemented a dataset, HAMOD dataset, which consists of several sets (currently 128) of semantically related words and their corresponding outliers. The dataset was first created in 2019 at Masaryk University (Brno) - with 37 sets originally - and I increased its size and improved its quality. To do so, I designed a methodology which merged my background and expertise in different fields of linguistics, not only Computational Linguistics but also theoretical semantics (and more specifically, Lexical Semantics), as well my experience in a project developed at the University of Pavia (T-PAS, an online semantic-syntactic resource for Italian verbs).<sup>4</sup> The implementation of new sets in the dataset was conducted by exploiting the notion of semantic category (or semantic type, as in T-PAS) and domain, and I used sources such as T-PAS ontology and Wikipedia structures in order to retrieve potential topics and words to store in the sets. The dataset refinement was carried out by testing the difficulty of the words contained in it by performing a small experiment on a group of Czech young students: the reason for this is that we only wanted to include a range of basic vocabulary in the sets. The dataset translation was a collective step, in which some native speakers of the languages in the dataset (namely, Czech, German, English, Estonian, French, Italian, Slovak) were involved and coordinated with my supervision to make the new sets multilingual.

We used HAMOD dataset in a preliminary experiment of intrinsic evaluation, which was divided into two phases. First, a human evaluation was performed on a benchmark of 22 Linguistics students from the University of Pavia; the experiment resulted into high agreement between the evaluators, and the insights on the most common disagreements could help us to refine the dataset. Then, distributional models computed on the most recent Sketch Engine web corpora were evaluated, with the technical support of Lexical Computing. The models' performance proved to be significantly lower than the human; furthermore, Word Embedding models outperformed Sketch Engine Thesaurus in all the languages evaluated. It is therefore worth investigating further on these results with a view to improving Sketch Engine Thesaurus.

<sup>&</sup>lt;sup>3</sup> See Chapter 2.

<sup>&</sup>lt;sup>4</sup> See Jezek et al. (2014).

The thesis is structured as follows.

Part 1 is theoretical: we explore thesauri, distributional models and the issues related to the evaluation in Computational Linguistics.

In Chapter 1 we provide a theoretical introduction to semantic relations – which are the core structure of thesauri; we then define what is a thesaurus and trace the history and developments in their construction; after, we focus on the latest advance in thesauri construction, that is, distributional thesauri, by recalling the distributional hypothesis and the various distributional models (within which Sketch Engine Thesaurus and Word Embeddings fall).

In Chapter 2 we approach the issue of the evaluation in general and the standard procedures and metrics; then we focus on intrinsic and extrinsic evaluation of distributional models, providing a detailed review of the existing methods, their advantages and drawbacks; finally, we introduce the outlier detection task, as first conceived by Camacho-Collados & Navigli (2016), and review some studies which applied this task.

Part 2 is applied: we focus on the construction of HAMOD dataset and outline the methodology we conceived to implement it.

In Chapter 3 we discuss the motivations and purposes of HAMOD dataset, recalling its background study and highlighting major differences; then we describe the structure of the dataset, its content and its formal layout, while tracing the history of its creation since 2019; finally, we present the outcome of our dataset implementation, that is, the current state of HAMOD dataset with its 128 sets and 7 languages.

In Chapter 4 we discuss the methodology we conceived and applied in order to implement HAMOD dataset, which can be outlined in steps: new sets implementation, translation and adaptation, words difficulty refinement; finally, we present the guidelines we wrote for future contributors who would like to enlarge the dataset further.

Part 3 is also applied: we describe the intrinsic evaluation experiment we conducted using HAMOD dataset.

In Chapter 5 we describe the setup of the experiment on HAMOD dataset, specifying some hypotheses on the expected results; we then outline in detail the setup of the human evaluation in a controlled environment, the web interface used in the experiment, and the evaluation metrics employed; finally, we describe the distributional models' evaluation setup and the evaluation metrics.

In Chapter 6 we provide the results of the human evaluation and models' evaluation separately, in a quantitative and qualitative perspective, and jointly, as a comparison between the human and the models' performances.

# Part 1. A Theoretical Introduction to Thesauri, Distributional Models and Evaluation

# Chapter 1. Thesauri: Semantic Relations, History, Applications, and Construction

In this Chapter we provide a theoretical introduction to *thesauri*. The term *thesaurus* can refer to a number of different language resources, useful for a range of different linguistic purposes and applications in Natural Language Processing (henceforth, NLP). It can be defined, in general as a resource in which words with similar meanings are grouped together (Kilgarriff & Yallop, 2000). The criteria for grouping words relies on semantic relations among words, one of the subjects studied in Lexical Semantics, the area of Linguistics under which this thesis project falls. Among the various types of thesauri, our interest is towards those automatically built on the underlying distributional theory of meaning. Distributional thesauri can be related to distributional models in general and therefore part of this Chapter is dedicated to the discussion of the distributional hypothesis and what word embedding are as representations of meaning in the semantic space.

Here follows a brief outline of the Chapter.

In Section 1.1 we briefly introduce Lexical Semantics and the various types of semantic (paradigmatic) relations in general.

In Section 1.2 we try to define what a *thesaurus* is and in which terms semantic relations are involved. We also clarify the distinction among *thesauri* and *dictionaries*. Then, we trace the history and development of thesauri, in their various applications, from Roget's Thesaurus to thesauri in Information Retrieval, from WordNet to distributional thesauri.

In Section 1.3 we focus on automatic distributional thesauri, the kind of resources that are at the core of this thesis project and will be evaluated in the applied part.<sup>1</sup> We briefly recall the distributional hypothesis, and we discuss various approaches to the automatic construction of distributional thesauri.

In Section 1.4 we explore the distributional models we analyse and evaluate within this thesis project. We present one specific distributional thesaurus, Sketch Engine Thesaurus, mentioning how it is constructed and its main features. We then discuss word embeddings, that is representations of meaning in semantic space, in which terms they can be related to distributional thesauri and in what Sketch Engine Thesaurus and word

<sup>&</sup>lt;sup>1</sup> See Chapter 5 and Chapter 6.

embeddings differ. Finally, we introduce one of the major issues in automatically built thesauri, that is, how to assess the quality of their performance.<sup>2</sup>

# 1.1 Lexical Semantics and Semantic Relations

Before introducing thesauri, we define the scope of this thesis, that is Lexical Semantics, and one of its matters of study, (paradigmatic) semantic relations.

#### 1.1.1 Lexical Semantics

*Lexical Semantics* (Cruse, 1986) is a branch of semantics which studies the meanings of words, with a focus on *content*, rather than *grammatical* words (Cruse, 2000: 15).<sup>3</sup> Content words are those which carry a lexical meaning, tend to be semantically autonomous and usually constitute open word classes (e.g., adjectives, nouns, verbs) (Jezek, 2016: 14-15).

Lexical semantics is concerned not only with defining the meaning of words, but also with explaining its flexibility in context, and accounting for how it contributes to the meaning of sentences, and with how words may or may not be combined (Jezek, 2016: 3). Therefore, the unit of analysis is the word. It is worth tracing a distinction between *lexical* and *phrasal* (or *sentence*) semantics. While lexical semantics investigates words and their meaning, phrasal (or sentence) semantics focuses on how the meaning of complex linguistic expressions (i.e., phrases and sentences) is obtained starting from the meaning of the constituent words (Jezek, 2016: 55). As Jezek (2016: 55) points out, the two branches are complementary: on one side, words contribute to building the meaning of sentences; on the other, context<sup>4</sup> influences the way words are interpreted.

<sup>&</sup>lt;sup>2</sup> This topic is widely examined in Chapter 2.

<sup>&</sup>lt;sup>3</sup> The distinction among content and function (grammatical) words is outlined in Jezek (2016: 14-15).

<sup>&</sup>lt;sup>4</sup> The context can be generically defined as a set of words that immediately precede or follow it, that is, its immediate linguistic environment (Jezek, 2016: 55).

#### 1.1.2 Semantic relations

One of the focuses of Lexical Semantics is the study of the relations that occur among words, or more specifically, their meanings. We can distinguish various kinds of semantic relations between words, and a distinction can be traced between *paradigmatic* and *syntagmatic* relations.<sup>5</sup>

*Paradigmatic* relations reflect the semantic choices available at a particular structure point in a sentence (Cruse, 2000: 148), that is, these relations hold among words that can be grammatically substitutable for each other in the same context (Murphy, 2003: 8; Jezek, 2016: 162). We can consider an example from Cruse (2000: 148): in the following sentence, the gap can be equally filled by different words.

(1) I'll have a glass of \_\_\_\_\_ . beer wine water lemonade

These words stand in a paradigmatic relation among each other and form what is called a *lexical paradigm*, containing words of the same grammatical category that share some semantic features (Murphy, 2003: 8; Jezek, 2016: 162). As Jezek (2016: 163) points out, paradigmatic relations are primarily associations between *meanings*, and secondarily between *words*: indeed, we need to take into account polysemy, which activate different relations for each meaning.<sup>6</sup> Furthermore, paradigmatic relations occur between lexical items that belong to the same *word class* (noun-noun, verb-verb, adjective-adjective etc.).

*Syntagmatic* relations, instead, hold between terms that can occur together linearly in an expression, to form complex linguistic units such as phrases, sentences, and texts; particularly between terms that stand in an intimate syntactic relationship, such as "subject of," "object of," "modifier of," and so forth (Murphy, 2003: 8; Jezek, 2016: 161). For example, a syntagmatic relation occurs between *cold* and *water* (*cold* is "modifier of" *water*).

We now focus on *paradigmatic* relations, which are those that organize and structure the lexicon of a language (Jezek, 2016: 163) and are exploited in the construction of thesauri. Before addressing specific relations, a further distinction (Cruse, 1986; Cruse, 2000; Murphy, 2003; Jezek, 2016) has to be made between:

<sup>&</sup>lt;sup>5</sup> See Saussure (1916); Hjelmslev (1961).

<sup>&</sup>lt;sup>6</sup> Jezek (2016: 163) reports a clarificatory example: the adjective *big* associates with *substantial* or *massive* in its 'large' sense (as in "a big building"), with *bad* in its 'intense' sense (as in "a big headache"), with *important* in its 'significant' sense (as in "a big decision"), and so forth.

- 1. *Vertical relations* (or *hierarchical* relations). In this kind of relations, one term is superordinate to another, which is subordinate (Jezek, 2016: 163).
- 2. *Horizontal relations* (or *non-hierarchical* relations). In this kind of relations, the two terms are symmetric, or on the same level (Jezek, 2016: 163).

Vertical relations are mainly in relations of *inclusion* (hyponymy/hyperonymy, meronymy/holonymy). Horizontal relations can be distinguished in relations of *identity* (synonymy) or *opposition* (antonymy, complementarity, converseness).

#### Vertical relations

The first relation of inclusion we define is *hyponymy/hypernymy* (Cruse, 2000; Murphy, 2003; Jezek, 2016). It is a hierarchical relation between two words (especially verbs and nouns), one of which (the hyponym or *subordinate*) has a more specific meaning than the other (the hyperonym or *superordinate*).<sup>7</sup> It is the relation that occurs, for example between (the first item is the hyperonym, the second the hyponym):

(2) vehicle – car fruit – apple move – walk

This relation can be instantiated by the relation of entailment "is a" (*car* is a *vehicle* etc.); it is asymmetrical and unilateral: *car* is a type of *vehicle*, but we cannot say that *vehicle* is a type of *car* (Jezek, 2016: 165). This kind of relation is at the basis of *taxono-mies*,<sup>8</sup> that is, hierarchical classifications of words, in which multiple levels of hyponymy and hyperonymy are included (that is, a hyponym can also serve as a hyperonym for other terms).<sup>9</sup>

One relation that derives from hyponymy/hyperonymy is *co-hyponymy*: one hyperonym can have multiple hyponyms. For example, *car*, *airplane*, *bus*, *train* are co-hyponyms of *vehicle*.<sup>10</sup>

Another type of vertical relation is *meronymy/holonymy*, or *part-whole* relation. It is a hierarchical relations between two words (especially nouns denoting concrete entities),

<sup>&</sup>lt;sup>7</sup> That is, the meaning of the hyponym consists of the meaning of the hyperonym plus sum additional features.

<sup>&</sup>lt;sup>8</sup> See T-PAS System of Semantic Types (Jezek, 2019) for an example of a hierarchical structuring of concepts and lexical items.

<sup>&</sup>lt;sup>9</sup> For example, *mammal* is the hyperonym of *feline*, which is the hyperonym of *cat*; but cat is also hyponym of feline and feline is hyponym of mammal. This relation is transitive, as *mammal* is also hyperonym of *cat*, and *cat* is hyponym of *mammal*.

<sup>&</sup>lt;sup>10</sup> We include this relation here, even though it is more a horizontal relation, rather than a vertical one.

one of which (the meronym) denotes a *part* of the other (the holonym), which is the *whole* (Jezek, 2016: 167-168).<sup>11</sup>

It is the relation that occurs, for example, between (the first item is the holonym, the second the meronym):

(3) hand – finger tree – branch team – athlete

This relation can be instantiated by "is a part of", which also helps us differentiating hyponymy and meronymy. Indeed, we can say that *finger* is a part of *hand*, but we cannot say that *finger* is a type of *hand*; vice versa, we cannot say that *car* is a part of *vehicle*.

Likewise hyponymy/hyperonymy, also meronymy/holonymy has a *co-meronymy* relation (but in this case, the relation can be bilateral): a meronym can have multiple holonyms (e.g., *slice* is a meronym of *bread*, *cheese*, *cake*, *pizza* etc.) and a holonym can have multiple meronyms (e.g., *face* has *nose*, *mouth*, *eyes*, *cheeks* etc. as meronyms).<sup>12</sup>

#### Horizontal relations

Moving to horizontal relations, *synonymy* is the *identity* relation we discuss here. Several definitions of synonymy have been proposed and none of them is univocal; indeed, we can discern various kinds of synonymic relations, or, according to Cruse (2000: 156), "degrees of synonymy". A general definition of synonymy is provided by Cruse (2000: 156): he points out that synonymy is not simply about sameness in meaning, but instead synonyms are «words whose semantic similarities are more salient than their differences». More specifically, we can distinguish three different degrees of synonymy: absolute synonymy, propositional synonymy, and near-synonymy (Cruse, 2000: 156).

*Absolute synonymy* refers to complete identity of meaning (Cruse, 2000: 157) or semantic equivalence between two words that can always be substituted for each other in any context in which they occur (Cruse, 2000: 157; Jezek, 2016: 172). It concerns an extremely limited (if not inconsistent, according to Cruse, 2000: 157) set of – typically monosemous – words, such as (Jezek, 2016: 172):

(4) misery – poverty mix – blend enough – sufficient

<sup>&</sup>lt;sup>11</sup> This is a general definition, for further specifications we refer to Jezek (2016: 168-169).

<sup>&</sup>lt;sup>12</sup> We include this relation here, even though it is more a horizontal relation, rather than a vertical one.

*Propositional synonymy* (Cruse, 2000) or *contextual synonymy* (Jezek, 2016) refers to the contextual interchangeability of words; that is, it is the relation that occurs between words that, in at least one context of use and for a specific meaning (we need to acknowledge polysemy), can be substituted for each other without consequences for the interpretation of the sentence (Jezek, 2016: 173).<sup>13</sup> Examples of propositional synonyms are (Jezek, 2016: 173-174):

(5) box – package difficult – hard take place – happen

Finally, *near-synonymy* is a more complex type of synonymy, which interlaces with a generic notion of semantic distance or similarity (Cruse, 2000: 159; Jezek 2016: 174).<sup>14</sup> We can simply define it as the relation in which interchangeability is less effective: two near-synonyms cannot be substituted in a context without giving odd results (Jezek, 2016: 174). This definition can be better understood if we consider some criteria to distinguish near-synonyms are (Cruse, 2000: 160; Jezek, 2016: 174):

- 1. degree (higher or lower value of a property; e.g., *cold freezing*, *disaster catastrophe*)
- 2. manner (adverbial specialization of verbs, that is, different manners to perform the same type of action; e.g., *whisper mumble*, *chuckle giggle*)
- 3. connotation and register (expressive meanings differ according to formal, colloquial, vulgar registers; e.g., *mother mum, delicious yummy*)
- 4. geographic area (one item has different word forms according to the geographic area; e.g., *cookie – biscuit*; *fall – autumn* in American vs. British English).
- 5. gender (similar items are expressed in different word forms according to gender distinctions; e.g., *pretty* (F) *handsome* (M); *celibe* (M) *nubile* (F) Eng. 'unmarried' in Italian)

We finally address relations of *opposition*, that is, terms in contrast with each other with respect to one key aspect of their meaning. They cannot be substituted with each other without changing the interpretation of the sentence or expression in which they occur, and they cannot be simultaneously true for the same entity (e.g., we cannot say something like "\*the street is both *wide* and *narrow*") (Jezek, 2016: 176). Following Jezek

<sup>&</sup>lt;sup>13</sup> Or, in other terms, without affecting the truth-conditional properties of the sentence or expression (Cruse, 2000: 158).

<sup>&</sup>lt;sup>14</sup> On relatedness and similarity, see Section 1.3.2.

(2016) and Cruse (2000) we address here *antonymy*, *complementarity*, and *converseness* as types of relations of opposition.

Antonymy (or polar opposition) is described by Jezek (2016: 176) as follows:

Antonyms [or polar oppositions] are word pairs that denote a property [...] or a change in property [...] that has the characteristic of being gradual from a conceptual point of view. Two antonyms, therefore, oppose each other in relation to a scale of values for a given property, of which they specify the two poles (or bounds).

Here follows some examples of antonyms (Jezek, 2016: 176):

(6) easy – difficultwide – narrowto lengthen – to shorten

One peculiar property of antonyms is that the negation of one term is not equivalent to the other term (e.g., *not easy* does not straightforwardly mean *difficult*), as antonyms are polar gradations of a specific property on a scale which can include other intermediate lexicalized gradations. Let us consider, for example, properties of temperature (Jezek, 2016: 177):

(7) freezing - cold - lukewarm - warm - hot - boiling - burning

Therefore, *not freezing* does not necessarily mean its polar opposite *burning*, but it may mean *cold*, *lukewarm*, *warm* etc.

*Complementarity* (or *binary opposition*) differs from antonymy in the sense that the opposition is not polar but instead mutually exclusive (that is, there are not intermediate gradations): if anything (within the appropriate area) falls into one of the compartments, it cannot fall into the other, and if something does not fall into one of the compartments, it must fall into the other (Cruse, 2000: 168; Jezek, 2016: 178). Here follows some examples:

(8) dead – alivetrue – falseto continue – to stop

The property of antonyms we have enunciated above is therefore inverse: is the negation of one term *is* the equivalent to the other term (e.g., *not dead* straightforwardly means *alive*).

Finally, *converseness* refers to the relation between words whose meaning involves necessarily an asymmetrical relation between at least two elements, each of which express

the underlying relation in the opposite way from the other (Jezek, 2016: 178-179). Here follows some examples:

(9) husband – wifebuy – sellin front of – behind

Indeed, *husband* is such only in relation to a *wife*, and vice versa, being the relation "married to".

To sum up, following the relevant literature in this Section we have explored and classified a variety of semantic paradigmatic relations between words and their meanings, which are one of the objects of study of Lexical Semantics. In the following Sections we will introduce thesauri, peculiar linguistic resources which are structured exploiting semantic relations between words.

# **1.2 Definitions of Thesaurus, History of its Development, and its Applications**

This Section defines what thesauri are and traces a brief history of the evolution of thesauri, following Kilgarriff & Yallop (2000), and the encyclopaedic entry for *thesaurus* in the *Encyclopaedia of Language and Linguistics* (2006).

#### 1.2.1 A definition of Thesaurus

The term *thesaurus* (from Latin, *thesaurus* and Ancient Greek θησαυρός '*thesaurós*', Eng. 'treasury') has undergone several extensions in meaning over the years, and currently three distinct meanings can be distinguished (Hartmann, 2006):

1. *special word list, lexicon* (obsolete): it refers to the first historical dictionary projects or generically encyclopaedic works which aimed at covering vast and exhaustive knowledge collections in various field of interest.<sup>15</sup>

<sup>&</sup>lt;sup>15</sup> We will not discuss this specific acceptation here, as it is out of our scope. Please refer to Faloppa (2011).

- semantically organized dictionary, nomenclator: it is the most relevant sense nowadays, since its first use in a lexicographic product, *Thesaurus of English Word and Phrases* by Peter Mark Roget in 19<sup>th</sup> century, a conceptually structured word-finder.
- 3. *terminological database, index*: it refers to specific tools used in Information Retrieval, consisting in indexes to the vocabulary of specific fields in the study and codification of technical terminology.<sup>16</sup>

In this thesis we are exclusively concerned with the second use of *thesauri*, on which we will focus in the following part of this Section. In this sense, a general definition of thesauri may be the one reported in Hartmann (2006):

A linguistically oriented wordbook [better, resource] in which general language vocabulary is organized from a semantic point of view, designed to give guidance on alternative or related words for similar concepts.

Starting from Hartmann (2006) definition, we try to explore this kind of resources first contrasting thesauri and dictionaries. Then, we address the issue how "language vocabulary is organized from a semantic point of view" in a thesaurus, that is, how the thesaurus is generally structured (at least, in traditional approaches): this helps us connecting to the previous Section (Section 1.1), in which semantic relations are discussed. After, we trace a brief history of thesauri, from Roget's Thesaurus to automatic thesauri, by mentioning WordNet (Fellbaum, 1998). Finally, if we consider a broader and inclusive definition of thesauri as «a resource in which words with similar meanings are grouped together» (Kilgarriff & Yallop, 2000), it is worth briefly touching upon Information Retrieval thesauri (the third meaning outlined by Hartmann, 2006), which are a different kind of resource, and it is not examined within this thesis.

#### 1.2.2 Thesauri and dictionaries

In order to trace a distinction between *thesauri* and *dictionaries* we can consider their practical use, first. As Kilgarriff & Yallop (2000) underline, traditional thesauri are marketed as aids to help writers choose the appropriate word, and for this the critical consideration is to provide a wide range of possibilities and alternatives to a specific word searched. On the other side, the main purposes of (at least, monolingual) dictionaries are, for example to help with finding meanings for rare words and correct spellings. This distinction can also be depicted in these terms (Hartmann, 2006): dictionaries generally fulfil a (passive) decoding function (i.e., reading and comprehension); thesauri generally fulfil an (active) encoding function (i.e., writing).

<sup>&</sup>lt;sup>16</sup> See Section 1.2.7.

Given this difference in use, we can introduce a distinction which can help to better define thesauri and dictionaries in contrast (Hartmann, 2006): *semasiology* vs. *onomasiology*. Following Geeraerts (2003), they can be defined as two complementary perspectives (or point of view on the same object, the word):

- 1. *Semasiology* takes its starting point in the individual word and looks at the semantic information (or, concepts) that may be associated with that word: in other words, what are the meanings of a specific word?
- 2. *Onomasiology* takes its starting point in a concept and investigates which words may be associated with that concept: in other words, what are the words associated to a specific meaning?

According to Hartmann (2006), the traditional general monolingual dictionary is *se-masiological* in the sense that it provides an explanation of the meanings of, or concepts instantiated by a given word by means of definitions or examples. A thesaurus (which is called *onomasiological dictionary*) is *onomasiologic* in the sense that it starts with a given concept or meaning and it provides a range of lexical choices for expressing that concept. In general, in a dictionary, the direction is from word  $\rightarrow$  to meaning; in a thesaurus, the direction is from meaning  $\rightarrow$  to word.

As far as distinction is concerned, the two perspectives can be integrated, and dictionaries can – and usually do – include onomasiological information. Indeed, semantic information in dictionaries goes beyond the description of separate words and word meanings, as words do not exist in isolation but are related to each other in various ways (e.g., by lexical relations or by belonging to the same conceptual domain) (Geeraerts, 2003). Onomasiological information is usually included by adding synonyms and antonyms to a dictionary entry, or by specifying the conceptual domain to which a word meaning may pertain (e.g., medical, mathematics, psychology, sports etc.) (Geeraerts, 2003).

Further differences between thesauri and dictionaries are highlighted in Kilgarriff & Yallop (2000).

First, a relevant difference concerns the structure, or indexing, of the words in the resource: the dictionary is organized *alphabetically*, whereas the thesaurus is typically *thematically* organized by meaning or word group (this being semantic categories, topics, domains, or, in general, a taxonomy).

Second, another difference can be spotted in goals: in producing a dictionary entry, the goal is to provide a coherent analysis that separates out the distinct meanings and patterns of use the word has, with each part of the entry making sense in relation to the others (*distinctive* approach); when producing a thesaurus entry, the unit which must appear coherent is the thesaurus entry or word group (*cumulative* approach). As the authors point out, a dictionary distinction may be lost in the thesaurus, and, conversely, a single dictionary meaning is commonly found in more than one section of the thesaurus.

We can consider, in terms of comparison, the following example. We used the same online resource, Merriam-Webster, which offers both a thesaurus and a dictionary for the English language.<sup>17</sup> We compare the same entry, one meaning of the adjective *big* in the thesaurus  $(10)^{18}$  and in the dictionary (11).<sup>19</sup>

(10) *big* (adjective)

#### synonyms for big

consequential, earth-shattering, earthshaking, eventful, historic, important, major, material, meaningful, momentous, monumental, much, significant, substantial, tectonic, weighty

#### words related to big

decisive, fatal, fateful, strategic earnest, grave, heavy, serious, sincere, distinctive, exceptional, impressive, outstanding, prominent, remarkable, valuable, worthwhile, worthy distinguished, eminent, great, illustrious, noble, notable, noteworthy, outstanding, preeminent, prestigious famous, notorious, renowned, all-important, central, critical, crucial, essential, key, pivotal, seminal, vital

#### near antonyms for big

paltry, petty, worthless, anonymous, nameless, obscure, uncelebrated, unknown

#### antonyms for big

inconsequential, inconsiderable, insignificant, little, minor, negligible, slight, small, trifling, trivial, unimportant

#### (11) big adjective \'big\

bigger; biggest

1a: large or great in dimensions, bulk, or extent *a big house*also : large or great in quantity, number, or amount *a big fleet*1b: operating on a large scale *big government*

<sup>&</sup>lt;sup>17</sup> A peculiarity of this thesaurus is that it disambiguates the words for each entry, by separating synonyms, antonyms, related words according to the specific meaning of the headword. It also specifies the semantic relations, which are usually left implicit in thesauri such as Roget's.

<sup>&</sup>lt;sup>18</sup> From: <u>https://www.merriam-webster.com/thesaurus/big</u> (last access: 24/06/2022).

<sup>&</sup>lt;sup>19</sup> From: <u>https://www.merriam-webster.com/dictionary/big</u> (last access: 24/06/2022).

#### 1.2.3 Thesauri's structure

In the previous paragraphs, some peculiarities of thesauri structures have already emerged. We briefly address this issue here, keeping however in mind that there are considerable variations among thesauri, and several subtypes of thesauri can be distinguished (Hartmann, 2006).

The first characteristic we can address is the one concerning the *macro-structure* (that is, how the entries are listed, or indexed). While we have stated in the previous Section (Section 1.2.2) that dictionary entries are organized *alphabetically*, and thesauri entries *thematically*, there has been a strong influence by the semasiological tradition, resulting in thesauri organized alphabetically, assuming that the «universal spell of the alphabet is the most convenient device for arranging the information» (Hartmann, 2006).<sup>20</sup> Thematically structured thesauri group words in terms of semantic categories and conceptual domains (e.g., Roget's thesaurus). Categories conceptual domains are implicitly specified by paradigmatic semantic relations: hyponymy/hyperonymy (thus, taxonomically), synonymy, antonymy and meronymy.

The second characteristic, which we have already spotted in the previous Section (Section 1.2.2) is related to the *micro-structure* (that is, the internal organization of each entry). The organization of a thesaurus entry is usually *cumulative*, that is, each entry has lists of words as alternatives to the headword for that entry, with no explanation of what distinguishes a specific alternative from another<sup>21</sup> – being the semantic relations implicit and for the fact that no definitions are provided in thesauri. Dictionary entries, instead, are usually organized *distinctively*: for each meaning of the headword, explanatory discriminations by means of definitions and examples are provided.<sup>22</sup>

WordNet, as we will see below (Section 1.2.5), represents a combination of these two approaches (*cumulative* and *distinctive*), providing definitions and examples for each word listed with respect to a given headword.

<sup>&</sup>lt;sup>20</sup> As Hartmann (2006) notices, an alphabetic format may be easier from the point of view of access from the user, but it imposes on the lexicographer an even greater burden of sense discrimination.

<sup>&</sup>lt;sup>21</sup> This can be an issue in particular as far as non-native speakers are concerned.

<sup>&</sup>lt;sup>22</sup> This resulting in a better usability for non-native speakers consulting the dictionary.

We also mention here that thesauri can be bilingual and can be delimited to specific varieties of language (diachronic, social, literary) or domains.<sup>23</sup>

#### 1.2.4 History of Thesauri and Roget's Thesaurus

The second acceptation of the word thesaurus discussed in Section 1.2.1 can be grounded in a pioneer lexicographic work, the Thesaurus of English Word and Phrases by the British physician Peter Mark Roget (1779-1869), who undertook this project after his retirement.<sup>24</sup> First of its kind, the Thesaurus of English Word and Phrases was first published in 1852 by Roget, and he intended it as «a collection of the words it [the English language] contains and of the idiomatic combinations peculiar to it, arranged, not in alphabetical order as they are in a dictionary, but according to the *ideas* which they express» (Roget, 1852). The principles of organization which emerge from the preface of the thesaurus quoted in the previous lines are still valid nowadays: Roget's thesaurus is a catalogue of semantically related words and phrases, organized according to a system of classification of the ideas which are expressible by linguistic means (Jarmasz, 2003). The original classification included six major classes (abstract relations, space, material world, intellect, volition, sentiment, moral powers), within each of them finer categorizations are included: each class has several sections; under each section there are around 1 000 heads that represent various concepts; within these heads, words and phrases are organized in groups according to the part of speech (Jarmasz & Szpakowicz, 2012).<sup>25</sup> As we discussed in Section 1.2.2, the direction is from the concept to the word, as the words are the last level in the hierarchical structure of the thesaurus. Figure 1 shows the general classification of the thesaurus; Figure 2 shows a finer level of organization, taking as an example the class of the material world. Finally, Figure 3 shows how specific entries are organized within the thesaurus.<sup>26</sup>

<sup>&</sup>lt;sup>23</sup> These kinds of thesauri are not of our interest, as we focus on general language thesauri. Therefore, we refer to Hartmann (2006) for further discussion.

<sup>&</sup>lt;sup>24</sup> On the history of Roget's Thesaurus, see Hüllen (2005).

<sup>&</sup>lt;sup>25</sup> This classification can be clearly seen as an ontology, that is, a conceptual representation of the reality.

<sup>&</sup>lt;sup>26</sup> These three Figures and Figure 4 are taken from: <u>https://archive.org/details/Rogets-Thesau-rus/page/n23/mode/2up</u> (last access: 24/06/2022).



Figure 1. Roget's Thesaurus general classification



Figure 2. Roget's Thesaurus classification, with a focus on material world class

71	367-374
dritic, dendriform; woody, grassy; ver-dant, - durous; floral, mossy; lign-ous, -cous; wooden, leguminous; end-, ex-ogenous. "Extended lists of names of specific varieties of animals, vegetables, etc., are beyond the scope of this work.	namental –, flower –, kitchen –, truck –, market –, hop- garden; nursery; green-, hot-, glass-house; conservatory, cucumber frame, cloche, bed, border, seed-poit; grass-plat, lawn; park etc. (pleasure ground) 840; partere, shrubbery, plan- tation, avenue, arboretum, pinery, pinetum, of- chard, vineyard, vinery; orangery; farm etc.
368. Zoology. [ The science of animals.]—N. zoo-logy, -noniy, -graphy, -tomy; anitomy; comparative anatomy; animal –, comparative-physiology; morphology. anthrop-, ornith-, ichthy-, herpet-, ophi-, malac-, helminth-, entom-, oryct-, paleont-ology; ichthytectstomy; taxidermy. zo- etcologist. Adj. zoological etc. n.	<ul> <li>(abode) 189.</li> <li>V. cultivate; till, - the soil; farm, garden; sow, plant; reap, mow, cut; manure, dress the ground, dig, delve, dibble, hoe, plough, plow, harrow, rake, weed, lop and top, force, transplant, thin out, bed out, prune, graft.</li> <li>Adj. agr-icultural, -airan, -estic, arable; predial, rural, rustic, country, bucolic, Boeotian; horticultural.</li> </ul>
369. Botany. [The science of plants.]—N. botany; phyto-graphylogy, -tomy; vegetable physiology, herborization, dendr-, myc-, fung-, alg- ology; flora, pomona; botanist etc.; botanic garden etc. (garden) 371; hortus siccus, herbarium, her- bal. herb-ist, -arist, -alist, -orist, -arian etc. V. botanize, herborize. Adj. botanical etc. n.	372. Mankind.—N. man, -kind; human -race, - species, - nature; humanity, mortality, flesh, generation. [Science of man] anthropu-logy, -graphy, - wphy; ethno-logy, -graphy; humanitarinaism. human being; person, -age; individual, creature, lellow creature, mortal, body, somebody, one; such a -, somecone; soul, living soul; carthling; party, head, hand; dramatis personae; people, persons, folk, public, sciety, world; community at large; general public; mation, ality; state, realm; common weal, -wealth; republic,
370. Cicuration. [The economy or management of animals.]—N. taming etc. v.; iccuration, zoohygiantics; domestication, -ity; manège: veterinary art: breeding, pisciculture, spiculture etc. menagery, vivarium, zoological garden, zoo; sear-pit; aviary, apiary, hive; aquarium, fishery.	budy politic; million etc. (commonalty) 876; population etc. (inhabitant) 188. cosmopolite; lords of the creation; ourselves. Adj. human, mortal, personal, individual, national, civic, public, cosmopolitan; anthropoid.
fish hatchery: duck fish-pond; stud-farm; stock farm, dairy. [Destruction of animals] phthisozoics etc. (killing) 361. neat cow shep-herd, shepherdess: grazier: drover, cowboy, cowkeeper; trainer, breeder, groom, ostler etc. 746, veterinary surgeon, vet, horse doctor; farrier; keeper; game keeper, cage etc. (prison) 752; hen-coop, bird-cage, cauf; sheep-fuld etc. (inclosure) 232. V tame, domesticate, acclimatize, brecd, tend, break in, train, corral, round up; cage, bridle etc. (restrain) 751; le etc. 266.	373. Man.—N. man. male, he: manh+skd etc. (adolescence) 131; gentleman, sir, master, yeiman, wight, swain, fellow, guy, blade, beau, chap, galfer, gi+sk man; husband etc. (married man) 903; Mr., mister, monieur, sahib, Herr, señor, signor; buy etc. (youth) 129; Adinis, [Male animal] cock, drake, gander, dog, boar, stag, hart, buck, horse, entire horse, stallion; gib- tom-cat; he-, Billy-giat; ram, tup; bull, +ick; capon, ox, gelding; steer, stot. Adj. male, he: masculine; manly, virile; un- womanly, -feminine.
nilk; shear; hatch; incubate. Adj, pastoral, bucolic; tame, domestic, lomesticated, broken in, gentle, docile.	374. Woman.—N. woman, she, female, pet- ticoat, skirt, moll, broad, feminality, feminity, muliebrity; womanhood
371. Agriculture. [The economy or management of plants.]—N. agriculture. cultivation, husbandry, farming: georgics. geoponics; tillage, tilth, agronomy, gardening, spade husbandry, vintage; hort, arbor, silv-, citr-, vit-, flor-iculture; intensive culture; landscape gar- dening; forestry, afforestation. husbandman, horticulturist, citriculturist, gar- dener, florist; agricult-or, -urist; yeoman, farmer, cultivator, tiller of the soil, ploughman, sower, reaper, wondcutter, backwondsman, forester; vine grower, vintager, Boer; Triptolemus, field, meadow, garden; butanic –, winter –, or-	etc. (adolescence) 131: feminism; gynecology, gyniatrics, gynics. womankind; the -sex. – fair; fair – , softer-sex: weaker vessel; the distaff side. dame, madam, madame, mistress. Mrs., lady, mem-sahib. Frau, señora, signora, donna, belle, matron, dowager, goxdy, gammer; goxd -woman, – wife; squaw; wife etc. (marriage) 903; matron- age, -hoxd. Venus, nymph, wench, grisette; little bit of fluff girl etc. (south) 129. inamorata (hove) etc. 897; coartesan etc. 962. spinster, old maid, virgin, bachelor girl, new woman, amazon.

Figure 3. Roget's Thesaurus page showing entries organization

Another point can be made on the use Roget conceived of his thesaurus by considering the title of the first edition (Figure 4):<sup>27</sup> Thesaurus of English Words and Phrases Classified and Arranged so as to Facilitate the Expression of Ideas and Assist in Literary Composition. Indeed, since its first publication – followed by a substantial number of editions

<sup>&</sup>lt;sup>27</sup> The image is taken from: <u>https://taxodiary.com/2015/01/a-celebration-of-rogets-taxonomy/</u> (last access: 24/06/2022).

and enlargements (see Hartmann, 2006) countless writers, orators, and students of the English language have used it in writing (Jarmasz, 2003). As Roget himself states (Roget, 1852): «the assistance it gives is that of furnishing on every topic a copious store of words and phrases, adapted to express all the recognizable shades and modifications of the general idea under which those words and phrases are arranged» (Roget, 1852).



Figure 4. Roget's Thesaurus of English Word and Phrases (1st ed., 1852), cover page

Nowadays, Roget's thesaurus function has expanded outside of a strict lexicographic use. Since 1957, Roget's thesaurus machine tractable versions have served several computational applications: from machine translation to text classification, from word sense disambiguation, often together with, or as opposed to WordNet (for further discussion, see Jarmasz, 2003; Jarmasz & Szpakowicz, 2012; Kennedy & Szpakowicz, 2008).

#### 1.2.5 WordNet

WordNet (Fellbaum, 1998) is an online lexical database of English developed at Princeton University, designed according to psycholinguistic principles, and manually compiled by linguists and cognitive psychologists. We discuss WordNet here because, likewise thesauri, it organizes the lexicon according to semantic relations.

In WordNet, 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 (Miller et al., 1990). The main relation among concepts and words in WordNet is thus synonymy; other relations encoded are hyponymy/hyperonymy, meronymy/holonymy for nouns, troponymy<sup>28</sup> for verbs, and antonymy for adjectives.<sup>29</sup> We report here a Figure (Figure 5) exemplifying WordNet structure for the word *big*.<sup>30</sup>

<sup>&</sup>lt;sup>28</sup> We have not addressed troponymy in Section 1.1. Troponymy can be defined as a paradigmatic relation of inclusion that occurs between verbs (it corresponds to holonymy). With verbs, inclusion tends to be a matter of "manner" (e.g., move vs. run) (Jezek, 2016: 166).

<sup>&</sup>lt;sup>29</sup> See Section 1.1 for the definition and a discussion on these relations.

<sup>&</sup>lt;sup>30</sup> Taken from WordNet 3.1 online version: <u>http://wordnetweb.princeton.edu/perl/webwn</u> (last access: 24/06/2022).

WordNet Search - 3.1 - WordNet home page - Glossary - Help		
Word to search for: big	Search WordNet	
	Change tions, "W:" = Show Word (lexical) relations ample sentence"	
Adjective		
extent) "a large city"; "set out for th "a large family"; "big businesses"; " newspapers"; "a big group of scien • <u>similar to</u> • <u>attribute</u> • <u>antonym</u> • <u>W</u> : (adj) <u>small</u> [Opposed quantity or magnitude o <i>small car</i> "; "a little (or sr • <u>W</u> : (adj) <u>little</u> [Opposed quantity or magnitude o <i>small car</i> "; "a little (or sr • <u>derivationally related form</u> • <u>S</u> : (adj) big (significant) "graduation • <u>S</u> : (adj) big (very intense) "a b <i>shock</i> "; "a bad earthquake"; "a bad • <u>S</u> : (adj) big (loud and firm) "a big ver	d to: <u>large</u> ] (limited or below average in number or r extent) "a little dining room"; "a little house"; "a mall) group" to: <u>big</u> ] (limited or below average in number or r extent) "a little dining room"; "a little house"; "a mall) group" n was a big day in his life" bad headache"; "in a big rage"; "had a big (or bad) I storm"	

Figure 5. Screenshot from WordNet with some synsets for the word big

What is relevant herein is the difference between WordNet and thesauri in general. Indeed, WordNet resembles a thesaurus, in that it groups words together based on their meanings, and it is structured following semantic relations (Fellbaum, 1998: 7). However, there are some important distinctions (Fellbaum, 1988: 8):

- 1. WordNet interlinks not just word forms but specific senses of words, which are therefore semantically disambiguated: conceptual and lexical levels are separated, whereas in thesauri only lexicalized concepts are accounted for<sup>31</sup>
- 2. WordNet explicitly labels the semantic relations among words, while this is not the case for thesauri, in which semantic relations are implicit and unlabelled

<sup>&</sup>lt;sup>31</sup> As Fellbaum (1988: 8) states, «WordNet's particular structure therefore reveals a conceptual inventory that is only partially mapped onto the lexicon of English». This assumption can be referred to the conceptual approach to the nature of meaning (Jezek, 2016: 64-67).

3. WordNet provides brief definitions (called "glosses" per each synset, followed by one or more short sentences illustrating the use) whereas thesauri do not (in this, it resembles more a dictionary)

WordNet has several uses in computational linguistics: its applications include information retrieval, language generation, question answering, text categorization, text classification and word sense disambiguation.

#### 1.2.6 Automatic Thesauri

Hartmann (2006) already addressed the relevance of computational approaches to the construction of thesauri. Indeed, traditional lexicographic thesauri relied on *manual* approaches by experts. This ensured high quality of the entries but since human labor is expensive, the extent of the thesaurus – in terms of the number of words and concepts covered – was limited. It is clear that the quality of this kind of thesauri depended on the expertise and the opinions of the compilers for whom it was impossible to be aware of how millions of speakers of the language use each word.<sup>32</sup>

The most recent advance in thesauri construction, as an alternative to manually compiled thesauri, are *automatic* thesauri. These thesauri are generated automatically on the basis of large corpora according to various techniques (which we will explore in Section 1.3). These approaches have several advantages: the computations in order to compile the thesauri do not need any human intervention; the thesaurus size, in terms of the number of words covered is potentially unlimited, thus covering even rare or infrequent words;<sup>33</sup> it can also be rebuilt frequently as needed to reflect diachronic changes in word use, by applying the same automatic techniques; finally, a corpus-based automatic thesaurus relies on the analysis of language as produced by speakers. Nevertheless, an automatic approach raises some issues in the sense that it is not easy to assess the quality of its content, if there is no human intervention (we will deal with this point in Section 1.4.3 and, widely, in Chapter 2).

<sup>&</sup>lt;sup>32</sup> <u>https://www.sketchengine.eu/blog/automatic-thesaurus-synonyms-for-all-words/</u> (last access: 24/06/2022).

<sup>&</sup>lt;sup>33</sup> Thus, how useful this can be both in terms of lexicographic and computational uses has to be seen.

#### 1.2.7 Information Retrieval Thesauri

We briefly mention here thesauri in Information Retrieval, even though they are not in the scope of this thesis project.<sup>34</sup>

Information Retrieval (IR) is a field of computer and information science

concerned with the structure, analysis, organization, storage, searching, and dissemination of information. An IR system is designed to make a given stored collection of information items available to a user population. [...] In today's world, the information is more likely to be full-length documents, either stored in a single location, such as newspaper archives, or available in a widely distributed form, such as the World Wide Web (WWW) (Salton & Harman, 2003).

A thesaurus in Information Retrieval can be defined as a «vocabulary of controlled indexing language, formally organized so that a priori relationships between concepts are made explicit» (Aitchison et al., 2000). They are used to support information indexing (by compilers) and searching (by users) process in IR (Lassi, 2002; Clarke, 2019). As with WordNet, the organizing principles are synonymy and taxonomy, which make it possible for searches to be broadened or narrowed, and for searches to be matched against documents using synonyms of the search terms (Kilgarriff & Yallop, 2000). Originally, IR thesauri were manually constructed by domain experts; nowadays, automatic approaches are pursued (see Lassi, 2002; Clarke, 2019). Beside the use and the structure, another key feature that differentiates IR thesauri from lexicographic thesauri and WordNet is that they are typically domain-specific (e.g., medicine, psychology, engineering, chemistry etc.), whereas lexicographic thesauri and WordNet address general language (Kilgarriff & Yallop, 2000).

## **1.3** Automatic Distributional Thesaurus

We have already introduced automatic thesauri in Section 1.2.6. In this Section, we explore automatic thesauri, which are the kind of resources analysed within this thesis project. Hartmann (2006) points out the issue of which semantic theory is the most suitable in the construction of thesauri, and whether it can be supported by computational

<sup>&</sup>lt;sup>34</sup> For further information, see the works quoted in this Section (Aitchison et al., 2000; Lassi, 2002; Clarke, 2019).

techniques. In this respect, the distributional theory of meaning and a distributional approach to thesauri construction based on large corpora is the most pursued. We first introduce distributional semantics and some key notions related to it and then we explore distributional approaches to thesauri construction.

#### 1.3.1 Distributional hypothesis

Distributional semantics<sup>35</sup> relies on the *distributional hypothesis* on the nature of word meaning (Jezek, 2016: 73):

The basic tenet of this hypothesis is that the meaning of a word correlates with its distribution, i.e., the set of contexts in which it occurs, particularly its local context, that is, the words it stands in a grammatical relation with.

Therefore, words that occur in *similar contexts* – i.e., similar distributional properties – tend to have *similar meanings* (Jurafsky & Martin, 2021).

This hypothesis was first formulated in the 1950s by the American linguist Z. Harris (Harris, 1954) and in parallel work conducted in British lexicology by J.R. Firth (Firth, 1957) (Jezek, 2016: 73). The underlying idea is that the meaning of a word is inherently differential (Sahlgren, 2008; Jezek, 2016: 73) and it can be established by comparing words, that is, set of contexts.

A key notion in distributional semantics is the one of *similarity*, which we will discuss in the following Section. It is worth pointing out here that also dissimilarities – or differences – between meanings of words can be spotted through the distributional analysis: differences of contextual distribution can be interpreted straightforwardly as differences in meaning between the words in question (Jezek, 2016: 74).

#### 1.3.2 Similarity and relatedness

A key notion in the distributional hypothesis is the one of *similarity*, which we try to define in this Section, together with the notion of *relatedness*. These notions are notoriously obscure and subject to criticism, as seen as too broad to be useful (Sahlgren, 2008).

Budanitsky and Hirst (2006) differentiate semantic *similarity* (*narrow* concept) and *relatedness* (*broad* concept) as follow.

Semantic *similarity* is a narrower concept that holds between lexical items having a similar meaning and sharing common features, like *palm* and *tree*. It is usually defined via the lexical relations of synonymy and hyponymy,<sup>36</sup> and semantically similar words

<sup>&</sup>lt;sup>35</sup> For a comprehensive overview in a linguistic perspective, see Boleda (2020).

<sup>&</sup>lt;sup>36</sup> Canonical paradigmatic relations, see Section 1.1.
can be substituted for each other in context. As Gladkova & Drozd (2016) refer, *similarity* is also defined as co-hyponymy (e.g., *car* and *bicycle* as co-hyponyms of "means of transport") (Turney & Pantel, 2010), or as the relation exemplified by pairs of synonyms (Hill et al., 2015).

The broad concept of semantic *relatedness* (also called *association*, see Jurafsky & Martin, 2021) is a relation that holds between lexical items that are connected by any kind of lexical or functional association, or, in other words, it refers to any kind of semantic relation between words. For example, this is the kind of relation that occurs between *doctor* and *hospital*, *coffee* and *cup*, *train* and *departure*. As we can spot from these examples, semantically related words are not interchangeable in context (Kolb, 2009), but instead, they tend to co-occur in context (Turney & Pantel, 2010).

As Hadj Tajeb et al. (2020) underline, semantic *relatedness* includes *similarity*, for the fact that semantic relatedness can be intended as semantic proximity or association, that is, how much connection humans perceive between two words and their meaning. Semantic *similarity* is therefore a specific case of semantic *relatedness*, in that the sense of relatedness depends on the degree of synonymy, that is, the amount of shared properties. Consider, as further examples, the relation between two similar words, such as car - plane (which share the following properties: being means of transport, machines, powered by and engine through fuel), and the related words car - street (which do not share relevant properties).

Budanitsky and Hirst (2006) emphasize an additional differentiation between *semantic* similarity and *distributional* (or co-occurrence) similarity. Indeed, distributional methods that measure the similarity of the distributional behaviour of words do not take into account the different senses a word has, and therefore mix up the similar words for all the word senses. We can then define semantic similarity as a relation between *concepts* (or meanings), and distributional similarity is a relation between *words* (regardless of polysemy, that is, the number of meanings each word can have).

#### 1.3.3 Word Space Models

As Jezek (2016: 75) points out, recent developments in the field of computational semantics have led to a renewed interest in the distributional hypothesis, also thanks to the availability of large, digitalized corpora. A variety of computational techniques have been developed for the extraction of distributional profiles for words from texts, thus leading to the translation of the distributional hypothesis into a full-fledged computational model of meaning representation, called *distributional* or *word-space model* (Jezek, 75).

As the name suggests, distributional models are based on a *spatial* (or geometric) *metaphor*, according to which distributional similarity between two can be interpreted as spatial proximity (Jezek, 2016: 94). Furthermore, in this idea, word meanings are conceived as points in a multidimensional semantic space: in order for two meanings to be conceptualized as being close to each other, as the similarity-is-proximity metaphor suggest, they need to possess spatiality (Jezek, 2016: 94). At the heart of the word-space model there

is the idea that words that tend to combine with the same words occupy locations in the semantic space that are closer to each other than those of words with a different distribution (Jezek, 2016: 95).

With meaning defined by its contextual distribution, and being meaning a point in the multidimensional space, meaning can be represented as *vectors* which, in a distributional model, can be retrieved from a *matrix of co-occurrence*, a way of representing how often words co-occur (Jurafsky & Martin, 2021) (Table 1). In the matrix, each cell records the number of times the target word (rows) and the context<sup>37</sup> word (columns) co-occur in some context in some training corpus.

	computer	data	result	pie	sugar
cherry	2	8	9	442	25
strawberry	0	0	1	60	19
digital	1670	1683	85	5	4
information	3325	3982	378	5	13

Table 1. Co-occurrence matrix taken from Jurafsky & Martin (2021)

In algebra, vectors are elements in a vector space, consisting of an array of numbers determining their dimension and length. In the co-occurrence matrix, vectors are the rows, being the numbers the dimensions of the vectors (each dimension records the number of co-occurrences).<sup>38</sup> What follows (Figure 6) is a spatial bi-dimensional visualization of two of the vectors in Table 1 (*digital* and *information*), for two contexts (*computer* and *data*), from Jurafksy & Martin (2021):



Figure 6. Spatial visualization of word vectors (Jurafsky & Martin, 2021)

<sup>&</sup>lt;sup>37</sup> In this case the context is a word, but it can also be a region of text or a whole document.

<sup>&</sup>lt;sup>38</sup> In this case, vectors have 5 dimensions, but in real distributional models are exponentially more.

Within these models, to define the distributional similarity between two words we can measure the proximity of the vectors in the multidimensional semantic space by *cosine* distance, which is basically the dot product between vectors (Jurafsky & Martin, 2021).

Before moving to distributional thesauri, we now briefly introduce a major distinction in distributional models, as outlined in Baroni et al. (2014) and Almeida & Xexéo (2019), which allows us to introduce the notion of *word embeddings*.

Word embeddings are dense, distributed, fixed-length word *vectors*, built using word co-occurrence statistics as per the distributional hypothesis (Almeida & Xexéo, 2019). They are called "embeddings" because they are embedded in the semantic space, and they are currently the standard way to represent meaning in NLP tasks.

We can distinguish between two types of approaches for the computation of distributional models and word vectors: *predictive* models (that is, *word embeddings*) and *countbased models* (such as the approach at the basis of Word Sketch and Sketch Engine distributional thesaurus, see Section 1.4). *Predictive* models exploit single word contexts and predict the probability the next word given its context – a sequence of words (Baroni et al., 2014). *Count-based* models rely on global word context co-occurrence counts and leverage global co-occurrence statistics in word-context matrices (Almeida & Xexéo, 2019). The difference can be depicted as follows (Baroni et al., 2014):

Instead of first collecting context vectors and then reweighting these vectors based on various criteria [*count-based* models], the vector weights are directly set to optimally predict the contexts in which the corresponding words tend to appear [*predictive* models]. Since similar words occur in similar contexts, the system naturally learns to assign similar vectors to similar words.

The idea of predictive models, which is first turned out into *word2vec* model (Mikolov et al., 2013a), relies on the intuition that instead of counting how often each word co-occurs with another (count-based), we can predict the probability for that word to show up near the other word (predictive). We do not discuss this issues further, as we refer, for further discussion, Jurafsky & Martin (2021); Turney & Pantel (2010); Baroni & Lenci (2010); Almeida & Xexéo (2019).

It is worth briefly mentioning the difference between *static* and *contextualized* (or *dynamic*) word embeddings, which are out of the scope of this thesis. The embeddings we evaluate in this thesis (namely, *word2vec*) are *static* in the sense that they do not change with the context once been learned, and, despite their efficiency, the static nature of these embeddings makes it difficult to cope with the polysemy problem, since the meaning of a polysemous word depends on its context and static word embeddings fail to distinguish meanings of polysemous words (Wang et al., 2020). *Contextualized* (or *dynamic*) word embeddings overcome this issue. They are representation of words in context (Jurafsky & Martin, 2021): while static embedding models learn a single vector embedding for each unique word in the corpus, in contextualized word embedding models each word is represented by a different vector each time it appears in a different context (Jurafsky & Martin, 2021). These models are pretrained on large amounts of text and then fine-tuned on a specific NLP task, demonstrating dramatic superiority on these tasks as compared to static

models (Wang et al., 2020). Nevertheless, this comes at a cost, as these models are computationally and environmentally expensive both for pretraining and fine-tuning (Andersen et al., 2020). We then focus on a static model and refer to the Wang et al. (2020) and Jurafsky & Martin, 2021) for further discussion.

What is interesting for us is that distributional models of this kind can be used to retrieve distributional thesauri, as the ones we explore in the following Section (Section 1.3.4). Furthermore, distributional models themselves such as *word2vec* can be seen as distributional thesauri (see Section 1.4), allowing us to compare and evaluate different approaches, namely a count-based one and a predictive one, as we will discuss further in this thesis.

#### 1.3.4 An overview on distributional Thesauri construction

As Baroni & Lenci (2010) mention, distributional models have found wide applications in computational lexicography, especially for automatic thesaurus construction based on corpora. An automatic distributional thesaurus is generally viewed as a set of entries (headwords) with, for each *entry*, a list of semantic *neighbours* ranked in descending order of distributional similarity with this entry (Ferret, 2017); therefore, the semantic neighbours of an entry word are words whose contexts are similar to that of entry (Claveau & Kijak, 2016).<sup>39</sup> Words in a thesaurus are linked by various implicit semantic relations which all fall under the notion of similarity (synonymy, antonymy, hyperonymy, hyponymy). Compared to traditional lexicographic thesauri, neither the macro-structure nor the micro-structure are an issue in the organization of the thesaurus, as each word for which the thesaurus is computed in the corpus is both an entry and a neighbour to other entries.

Rychlý & Kilgarriff (2007), briefly describe the steps in the procedure for the construction of an automatic thesaurus as follows:

- 1. take a corpus
- 2. identify contexts for each word
- 3. identify which words share contexts
- 4. for each word, the words that share most contexts are its nearest neighbours

We will discuss Rychlý & Kilgarriff (2007) approach in the following Section (Section 1.4), as it is at the basis of Sketch Engine Thesaurus. We now briefly recall some reference studies on the construction of distributional thesauri. Pioneer works in distributional thesauri construction are those by Grefenstette (1994) and Lin (1998), which exploited syntactic parsing and selected word pairs that are in a certain grammatical relation to each

<sup>&</sup>lt;sup>39</sup> We will better understand this description in the presentation of Sketch Engine Distributional Thesaurus (Section 1.4.1).

other (e.g., head-modifier, verb-object, subject-object etc.) as candidates for semantic neighbours. Other syntactic-based approaches are discussed in Yang & Powers (2008) and Riedl & Biemann (2013). Ferret (2012) proposes a method for improving distributional thesauri by combining it with WordNet lexical networks for building similarity measures.

# 1.4 Sketch Engine Distributional Thesaurus and Word Embeddings

In this Section we focus on two distributional models which are subject to our analysis and evaluation within this thesis project: Sketch Engine distributional thesaurus and word embeddings, namely *word2vec* (which are computed on Sketch Engine corpora).

#### 1.4.1 A Distributional Thesaurus: Sketch Engine Thesaurus

Sketch Engine (Kilgarriff et al., 2014) is a corpus management system and corpus query tool used by linguists, lexicographers, translators, and publishers worldwide.<sup>40</sup> It contains 600 corpora in more than 90 languages, each having a size of up to 60 billion words.<sup>41</sup> Sketch Engine main feature is the Word Sketch, a short summary of a word's collocational behaviour from the perspective of individual grammatical relations (e.g., noun's modifier, verb's subject etc.). Each word sketch item is a triple consisting of the *headword*, the *grammatical relation* and the *collocate* (e.g., *beer*, *OBJECT\_OF*, *drink*). As such a word sketch is basically a dependency syntax graph, calculated using a hybrid rule-based and statistical approach. The backbone word for computing word sketches represents a hand-written word sketch grammar, which selects collocation candidates using the corpus query language (CQL).<sup>42</sup> A sketch grammar typically makes heavy use of regular expressions<sup>43</sup> over morphological annotation of the corpus to select syntactically

<sup>&</sup>lt;sup>40</sup> See <u>https://www.lexicalcomputing.com/lexical-computing/</u> (last access: 24/06/2022).

<sup>&</sup>lt;sup>41</sup> See <u>https://www.sketchengine.eu/</u> (last access: 24/06/2022).

<sup>&</sup>lt;sup>42</sup> On CQL: <u>https://www.sketchengine.eu/documentation/corpus-querying/</u> (last access: 24/06/2022).

<sup>&</sup>lt;sup>43</sup> On regular expressions: <u>https://www.sketchengine.eu/guide/regular-expressions/</u> (last access: 24/06/2022).

viable collocation candidates. These candidates are subsequently subject to statistical scoring using a word association score.<sup>44</sup>

What follows (Figure 7) is a screenshot of a Word Sketch page for the lemma *book* in the corpus enTenTen13:

WORD SKE	EICH	English We	b 2013 (enTen]	(en13) Q	] ()						Get more spa	ce⊕ ⊕ (?)	
book as noun 9,47	'5,135× 👻	)									ঽ	<b>±</b> ⊙ <b>=</b> ()	<b>*</b> *
<i>₽</i> 14	(0) ×	¢	₩Ø×	¢	N Ø N	<b>₽</b>	N Ø N	e 141	Ø X	<i>₽</i> 8	.≝ Ø ×	<i>₽</i> I	N Ø N
modifiers of "be	ook"	nouns modifie	d by "book"	verbs with "bo	ok" as object	verbs with "be subjec		"book" and/or		prepositional p	hrases	adjective predio "book"	ates of
comic		review		read		contain		article		of "book"		full	
comic books		Book Review		write		book contains		books and articles		in "book"		book is full of	
history the history books		signing a book signing	•••	publish		cover book covers	•••	magazine books and magazines		"book" of		available books available	
picture		club		buy	•••	call		book		"book" on	••••	online	
picture book		book club		recommend		a book called		movie		"book" in		book online	
address		fair		recommend this	book	include		cd		for "book"		interesting	
address book		Book Fair		sell		book includes		iournal		to "book"		useful	
new	•••	store		finish		describe	•••			"book" for		helpful	
new book		book store		love		book describes		film	••••	from 'book'		early	
first the first book	••••	book cover	•••	purchase		provide book provides		paper	•••	on "book"		book early to avoid	
favorite		chapter		review		feature		toy				disappointment	
favorite books		and book chapter		illustrate		book features		video	••••	"book" about	•••	amazing book is amazing	•••
text		publisher		illustrated book		explore		newspaper		with "book"	•••	worth	
text books		book publishers		authore		book explores the		author		~	$\otimes$	book is worth	
phone the phone book		series book series		~	•	present This book presents		~ ≈				fair book fair in	
reference	••••	shelf				focus book focuses on						good	
audio		book shelves				book focuses on						great book is great	

Figure 7. Word Sketch for the lemma book showing its collocational behaviour in enTen13 corpus

Another available feature in Sketch Engine is the distributional thesaurus (Rychlý & Kilgarriff, 2007). Word Sketches make it possible to automatically derive distributional thesauri by calculating the similarity of word sketch contexts: this identifies words that occur in similar contexts as a target word that should therefore be similar or related to it. As the others mentioned in the previous Section (Section 1.3), Sketch Engine Thesaurus is an automatically generated list of synonyms or similar words that should belong to the same category with respect to the target word (only nouns, adjectives, verbs, and adverbs are supported). The words in the list are ranked on the basis of a numeric *similarity score* (or conversely, a dissimilarity score, i.e., a distance) yielding in the first place the most similar words for the target word. Following Jakubíček et al. (2021),<sup>45</sup> to compute a similarity score between word  $w_1$  and word  $w_2$ , we compare  $w_1$  and  $w_2$ 's word sketches in this way:

<sup>&</sup>lt;sup>44</sup> For further information regarding this Sketch Engine tool, see <u>https://www.sketchen-gine.eu/guide/word-sketch-collocations-and-word-combinations/</u> (last access: 24/06/2022).

<sup>&</sup>lt;sup>45</sup> See also Kilgarriff et al. (2014) and <u>https://www.sketchengine.eu/documentation/statistics-used-in-sketch-engine/</u> (last access: 24/06/2022).

- find all the overlaps, i.e., where w<sub>1</sub> and w<sub>2</sub> share a collocation in the same grammatical relation, e.g.: (*beer/wine*, *OBJECT\_OF*, *drink*), where the association score is > 0
- 2. let  $ws_{w1}$  and  $ws_{w2}$  be the set of all word sketch triples (*headword*, *relation*, *collocation*) for  $w_1$  and  $w_2$ , respectively, where the association score > 0
- 3. let  $ctx(w_1) = \{(r, c) | (w_1, r, c) \in ws_{w_1} \}$
- 4. let  $AS_i$  be the association score of a word sketch triple (logDice)
- 5. then the similarity distance between  $w_1$  and  $w_2$  is computed as:

$$Dist(w_1, w_2) = \frac{\sum_{(r,c) \in ctx(w_1) \cap ctx(w_2)} AS_{(w_1,r,c)} + AS_{(w_2,r,c)} - (AS_{(w_1,r,c)} - AS_{(w_2,r,c)})^2 / 50}{\sum_{i \in w_3} AS_i + \sum_{i \in w_{32}} AS_i}$$

Here follows a screenshot from Sketch Engine Thesaurus, which can be consulted by searching for a specific lemma-entry and it can be visualized in two ways: as a list of semantic neighbours ranked by the similarity score outlined in the previous lines (Figure 8); as a word cloud or bubble chart, giving a more visual idea of the semantic distance/proximity of the neighbours to the entry word (Figure 9). In the list, words with higher similarity score are those which should be more similar to the entry. Frequency of the word in the reference corpus is also reported. In the bubble chart, the "bubble" sizes for each word refers to frequency of the word: *work* is more frequent than *piece*; the distance from the centre of each "bubble" indicates the similarity score: *story* is more similar to *book* than *report* is.

Τŀ	IESA	URUS	English Web 201	3 (enTe	enTen13) 🔍	(j)					Get more	e space 🕀	Ð	0 🖪	1
bo	ook as nou	n 9,475,135× 👻									10	ৎ ≛	• <del>•</del>	: :	) 1
	Word	Frequency ?	Similarity $^{?} \downarrow$		Word	Frequency ?	Similarity $^{?} \downarrow$	Word	Frequency ?	Similarity $^{?} \downarrow$	Word	Frequence	cy ? Sin	nilarity <sup>?</sup> ↓	
	story	5,094,257	0.563		<sup>14</sup> paper	3,391,593	0.493	 27 today	8,145,727	0.479 .	 40 lot	11,293,	226	0.469	
	article	5,316,970	0.539		15 video	4,422,373	0.490	 28 music	5,463,514	0.478 .	 41 history	4,938,	421	0.468	
	film	4,090,854	0.525		16 project	8,123,356	0.490	 29 site	10,395,423	0.478 .	 42 question	7,454,	112	0.467	
	work	13,918,242	0.523		17 something	10,069,694	0.488	 30 game	10,960,883	0.475 .	 43 card	5,167,	248	0.466	
	5 piece	3,843,154	0.510		18 program	10,968,768	0.484	 31 song	3,059,026	0.475 .	 44 record	3,614,	164	0.466	
	word	7,511,058	0.510		<sup>19</sup> experience	8,359,286	0.483	 32 part	13,542,324	0.473 .	 45 material	4,506,4	416	0.465	
	idea	7,324,365	0.508		20 website	6,863,922	0.483	 33 collection	2,652,731	0.473 .	 46 information	12,050,4	409	0.465	
	thing	18,769,726	0.504		21 picture	3,831,254	0.483	 34 product	10,551,787	0.471 .	 47 design	6,217,	311	0.465	
	series	3,755,011	0.503		22 storey	2,017,590	0.482	 35 list	4,916,758	0.470 .	 48 case	10,190,	271	0.464	
	10 post	4,700,749	0.501		23 blog	3,861,240	0.482	 36 item	4,022,480	0.470 .	 49 number	11,274,	518	0.464	
	11 page	6,847,041	0.501		24 report	4,948,416	0.480	 37 form	5,997,434	0.470 .	 50 art	5,037,4	473	0.462	
	12 show	4,944,507	0.498		25 class	5,297,426	0.479	 38 world	15,477,129	0.469 .					
	3 movie	3,238,617	0.497		26 event	8,197,530	0.479	 39 course	8,697,482	0.469 .					

Figure 8. Sketch Engine Thesaurus for the lemma book with list of similar items ranked according to a similarity score



Figure 9. Sketch Engine Thesaurus for the lemma book with bubble chart visualization

#### 1.4.2 Word Embeddings as Distributional Thesauri

Word embeddings (as we discussed in Section 1.3.3) and distributional thesauri share the same underlying distributional hypothesis, thus making them comparable, as we will discuss in the following Chapter (Chapter 2).

In this project, we use Sketch Engine word embeddings, which are created using a modified version of the *fastText* package, an extension of *word2vec* (Mikolov et al., 2013a), with the ability to read corpora and allows to calculate models with the same tokenization and format as the source corpora (Herman, 2021) and they are trained on skip-gram models of 100 dimension. A peculiarity of Sketch Engine word embeddings is that multiple models are available, namely one for *word* (i.e., raw corpus), one for *lemma* and one for *lemma* + *part of speech* (*lempos*). For each, a lowercase (*lc*) version can be selected (Herman, 2021).

The embeddings for around 15 languages are also accessible through a web interface (<u>https://embeddings.sketchengine.eu/</u>) which clearly resembles the one of the Sketch Engine Thesaurus (Figure 10).

Query book		Maximum Rank 100000	
Language English (Web, 2013)	Attribute * Lemma (lowercase)	×	
	SEARCH		
	Similarity	Rank	
paperback	0.819	10750	
non-fiction	0.784	15443	
books	0.780	6623	
hardcover	0.761	16674	
hardback	0.761	28181	
ebook	0.757	4631	
novels	0.756	45800	
e-book	0.756	8425	
anthology	0.750	10627	

Figure 10. Sketch Engine Embedding Viewer, lemma book and similar words

Indeed, for each entry (i.e., word queried) a list of similar words ranked according to the similarity score (in this case, cosine distance)<sup>46</sup> is generated, with the highest being the most similar to the headword. In the Figure, "Rank" refers to the position of the word in the list of all words in the reference corpus when ordered by frequency (it is not the frequency value in the corpus itself; this allows comparison between different corpora with different sizes in words).

# *1.4.3 How to assess the quality of automatic thesauri and to compare different approaches: the issue of evaluation*

Various approaches have been proposed for the automatic construction of distributional thesauri, as we have discussed in Section 1.3.

One issue that we can consider is how good an *automatic* technique for the construction of thesauri is compared to traditional approaches (that is, thesauri *manually* built by linguists and lexicographers).

Another point is how to compare different automatic techniques in distributional thesauri construction. It is quite obvious that various approaches relying on diverse techniques and interpretations of the distributional hypothesis provide different outputs in terms of the kind of the list of related words compiled for each target word in the thesaurus.

Let us consider a simple example: we can visualize Sketch Engine Thesaurus and Word Embeddings by searching the same target word and by comparing the list of related

<sup>&</sup>lt;sup>46</sup> See Section 1.3.3.

words that are printed out. The first Figure is from Sketch Engine Thesaurus (Figure 11): we can see a list of the first 20 similar<sup>47</sup> lemmas to the searched lemma *big* (as adjective) in enTenTen13 corpus.<sup>48</sup>

THESA		-	L										
big as adject	ive 9,07	72,553×	• …						2	<u>+</u> @	• =	()	☆
		Word	Frequency ?	Similarity $^{?} \downarrow$			Word	Frequency?	Simila	arity?↓	•		
	1	large	9,624,894	0.634		11	easy	7,048,496		0.45	1		
	2	small	9,994,762	0.576		12	different	10,762,805		0.44	5		
	3	great	17,905,210	0.498		13	powerful	1,900,263		0.44	4		
	4	strong	4,384,326	0.496		14	many	24,189,879		0.44	2		
	5	huge	2,771,633	0.477	•••	15	much	6,641,531		0.44	1		
	6	heavy	1,652,877	0.468		16	long	8,028,354		0.44	1		
	7	few	11,991,654	0.466		17	high	13,191,376		0.43	9		
	8	hard	5,170,194	0.462		18	expensive	1,538,805		0.43	9		
	9	bad	6,377,012	0.461		19	cheap	3,042,593		0.43	8		
	10	short	4,332,985	0.457		20	nice	3,148,355		0.43	8		

Figure 11. Screenshot from Sketch Engine Thesaurus, enTenTen13 corpus, adjective big

The second Figure is from Embedding Viewer in Sketch Engine<sup>49</sup> (Figure 12): we selected the same search word in its lemma form (to make the results more comparable, as Sketch Engine Thesaurus only works with lemmas) and what we can see is a list of the first 20 similar lemmas, ranked according to the similarity score (that is, cosine distance).

<sup>&</sup>lt;sup>47</sup> We use similar as this is how the functioning of Sketch Engine Thesaurus is explained. We are aware that similarity is a fuzzy concept and various semantic relations are instead retrieved by the Sketch Engine Thesaurus (not only synonymy). See Budanitsky & Hirst (2006).

<sup>&</sup>lt;sup>48</sup> EnTenTen13 is the English Web 2013 corpus, for more information see: <u>https://www.sketchen-gine.eu/ententen-english-corpus/</u> (last access: 24/06/2022). The reason why we selected this corpus as it is the only other English corpus available in Sketch Engine Embedding Viewer. The corpora on which we focused our analysis, instead, are more recent (namely, enTenTen20).

<sup>&</sup>lt;sup>49</sup> These models were trained using fastText, an extension of *word2vec* from the corpora available in the Sketch Engine using the SkipGram model with dimension 100 (<u>https://embeddings.sketchen-gine.eu/</u>, last access: 24/06/2022).

Query		Maximum Rank
Language English (Web, 2013)	Attribute	vercase)
	SEARCH	
	Similarity	Rank
huge	0.867	872
humongous	0.748	33748
ginormous	0.747	46143
super-sized	0.722	49397
massive	0.697	1996
gigantic	0.690	9402
large	0.688	238
enormous	0.684	3345
small	0.679	220
biggest	0.673	18363
big-time	0.671	26684
decent-sized	0.664	90646
helluva	0.663	35173
mega	0.659	8037
bigtime	0.656	57962
heckuva	0.653	85212
big-ass	0.651	91504
lot	0.651	188
little	0.650	181

Figure 12. Screenshot from Sketch Engine Embedding Viewer, enTenTen13 corpus, lemma big

If we compare the outputs of the Sketch Engine Thesaurus (column 1) and of the Word Embeddings (column 2), as it is visible from Table 2, we can notice that similar words in

common (in bold) are not much and, furthermore, they are ranked differently. Moreover, different semantic relations (e.g., antonymy, which is instantiated by *small* and *little* in the list) have different weights among the two models (*small* is in a higher position in Sketch Engine Thesaurus with respect to Word Embeddings).

Sketch Engine Thesaurus	Word Embeddings
large	huge
small	humongous
great	ginormous
strong	super-sized
huge	massive
heavy	gigantic
few	large
hard	enormous
bad	small
short	biggest
easy	big-time
different	decent-sized
powerful	helluva
many	mega
much	bigtime
long	heckuva
high	big-ass
expensive	lot
cheap	little
nice	great

Table 2. Sketch Engine Thesaurus and Word Embeddings in comparison for the target lemma big

Therefore, the issue is how to compare these two kinds of approaches (or, in general, various distributional models) and, subsequently, how to assess, and consequently evaluate, their quality. This leads us to Chapter 2, in which we approach the issue of evaluation of computational models, in general, and distributional ones (which include the two models we analyse in this thesis).

# Chapter 2. The Issue of the Evaluation and How to Evaluate Distributional Models

Evaluating computational models is fundamental to assess the quality of their performance. The evaluation of word embeddings has received a considerable amount of attention in recent years. As one of the goals of this thesis is to evaluate distributional models, in this Chapter we discuss the issue of evaluation.

Here follows a brief outline of the Chapter.

In Section 2.1 we discuss the purpose of evaluation in general, and the approach to evaluation. We then focus on human evaluation of computational models, and we briefly mention the standard evaluation metrics used.

In Section 2.2 we trace the distinction between intrinsic and extrinsic evaluation methodologies. We focus on the intrinsic evaluation tasks, among which is the *outlier detection* (the one we pursue within this thesis project). We also highlight the drawbacks of intrinsic methodologies and the issue of combining intrinsic and extrinsic techniques.

In Section 2.3 we present the outlier detection task as originally conceived by Camacho-Collados & Navigli (2016), and we review some subsequent studies which tried to improve and apply this task in various directions, before our contribution.

# 2.1 The Evaluation in General: Why and How to Evaluate

Computational models of various kinds – among which distributional models – have to be evaluated in order to assess the quality of their performance. The basic idea of the evaluation can be traced back to Alan Turing, who, in an article in the review *Mind* (Turing, 1950), introduces the well-known *Turing test*. Putting it simply, given three participants – a human judge, a machine and another human being tested – with no possibility to see each other (i.e., in separate rooms) and only written communication allowed (through a terminal), the human judge has the task to determine who, among the two, is the computer and who is the human. If the human judge cannot tell them apart, the machine succeeds in the test, indirectly proving that the machine performance is satisfying compared to the human one.

Although the test has been widely discussed and criticized, and several methods have been proposed in order to overcome it (for a survey, see Pinar Saygin et al., 2000), still the core idea at its basis is valid: compare machines/models/algorithms results with those produced by a human and analyse how close the model is to the human performance. In current NLP approaches this can be done mainly in two ways:

- 1. by comparing the model performance and the human performance on the same task, with humans providing a gold-standard dataset against which the model can be assessed (*intrinsic evaluation*)
- 2. by applying the model to a downstream NLP task (such as named entity recognition, sentiment analysis, etc.) (*extrinsic evaluation*)

We will focus on this in the following Section (Section 2.2), especially as far as distributional models are concerned (being this the kind of models we evaluate in this thesis). As a first step, let us now focus on the human side of the evaluation, taking our specific experiment as an example.

#### 2.1.1 Human evaluation and Inter-Annotator Agreement

Typically, at the basis of any evaluation method is an annotated dataset (in our case, *HAMOD dataset*), on which a task is defined (in our case, the *outlier detection task*) and which has to be performed by the model (in our case, by *distributional models*). The dataset can be collected automatically or manually (in our case, *manually*) and then, in order to assess its quality and usability in the task, human annotators/evaluators<sup>1</sup> are asked to annotate the dataset or to perform the task itself (in our case, we asked the evaluators to *perform the task itself*, that is, the same task that we made the models perform later).

Human judgments and their choices in annotation/evaluation procedures are of course subjective.<sup>2</sup> As Artsein (2017) points out, «there exists *variation* in annotator performance, and this variation needs to be examined in order to understand what the annotators are doing, and to be able to make meaningful use of the annotators' output». An annotation process is reliable (and thus, the usability of the dataset in the experimental task) if it is reproducible, that is, if the annotations yield consistent results (Artstein, 2017). In order to achieve consistency and then reproducibility, the human annotation/evaluation

<sup>&</sup>lt;sup>1</sup> We keep this distinction because in our specific experiment, humans were not asked to annotate some data (e.g., assigning some labels), but instead to evaluate the dataset itself that this thesis author has annotated/retrieved, through the task itself.

<sup>&</sup>lt;sup>2</sup> This part is inspired from some notes and slides of Elisabetta Jezek's *Dati Empirici e Teorie Linguistiche* held at the University of Pavia in 2020 (which was followed by this thesis' author).

tasks have to follow these requirements: the same dataset (or portion of dataset) has to be annotated or evaluated *independently* by *multiple* participants following some common *guidelines*. Let us explain this last point by following Artstein (2017):

To check for consistency, we need to apply the annotation process several times to the same source, and we need to use different annotators because a single person might remember their annotations from a previous round.

We also add that annotators should not influence each other in the decisions, and that one single annotator is not enough to overcome the subjectivity of the judgments in the process (therefore, the more they are, the more reliable the process is).

> The annotators should follow written guidelines, to make sure that the annotation process relies on knowledge that is transferable. They must work independently, so that agreements come from a shared understanding of the annotation guidelines rather than individual discussions on case points. Annotators should be drawn from a well-defined population in order for the researchers to know what shared assumptions they bring to the annotation process prior to reading the guidelines. The sample material must be representative of the totality of the material in terms of the annotated phenomena (Artstein, 2017).

After the annotation/evaluation process, human judgments are compared to find out whether and in which terms the annotators agreed or disagreed with each other. Indeed, «agreement among annotators on the same source data gives a measure of the extent to which the annotation process is consistent, or reproducible» (Artstein, 2017), and therefore the annotated or evaluated data can be used in further steps of the experiment.

Agreement between human annotators needs to be measured and there exist various formal metrics for that, generally referred as IAA (Inter-Annotator Agreement) measures. The simplest way is *raw* or *observed agreement*, that is, counting the number of items for which the participants provided the same answer or the same label in the annotation against the overall number of items in the dataset to be evaluated or annotated (Pustejovsky & Stubbs, 2012: 126; Artstein, 2017). As both Pustejovsky & Stubbs (2012: 126) and Artstein (2017) point out, these kinds of measure do not take into account random chance (or *accidental*) agreements that are likely to occur: agreement in itself does not imply that the annotation process is reliable. Therefore, there is one way to measure meaningful agreement, which is commonly used: *Cohen's Kappa*. This metrics is intended to calculate the amount of agreement that was attained above the level expected by chance or arbitrary coding (Artstein, 2017). It can be summarized as follows (Pustejovsky & Stubbs, 2012: 127):

$$k = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

In the equation, Pr(a) is the relative observed agreement between annotators, and Pr(e) is the expected agreement between annotators, if each annotator was to randomly pick a category for each annotation.

We do not focus on this aspect in this thesis, as we do not use this metrics in the evaluation task made by the human evaluators.<sup>3</sup> As discussed in Jakubíček et al. (2021), in our project we simply calculate *raw agreement*, as agreement by chance does not statistically play a relevant role.<sup>4</sup>

Once agreement is calculated, it can be compared to a general scale of agreement levels (such as the one mentioned in Pustejovsky & Stubbs, 2012: 131-132) or, depending on the task, to the high standard reached for that task in the relevant literature.

When the human evaluation is assessed by high agreement and, therefore, a gold standard is reached, we can proceed with a further step, that is, evaluating the models by comparing its performance against the gold standard.

#### 2.1.2 Evaluating the model against the gold standard: accuracy, precision, recall

A common way to evaluate a model is through a *confusion matrix*, a table which allows to confront the performance of the model with the gold standard (the human performance) and assess how good the model has done. In order to explain what a confusion matrix is and what it is useful for, we can consider a simple binary task<sup>5</sup> (our task, the *outlier detection*, is more complex in terms of combinations): the model has to give as an answer "yes" or "no" (the only two possible answers) to any input it receives (for example, an input can be a word and the task consist in telling whether the word is – yes – or is not – no – a Named Entity, or, putting it in another way, the model has to assign the label "Named Entity" to a word). We can intersect the model's answers to this task, and the human gold standard on the same task as follows (

Table 1):

Table 1. Confusion matrix for a binary task comparing model and human performance

gold standard						
model		yes	no			
	yes	true positive	false positive			
	no	false negative	true negative			

<sup>&</sup>lt;sup>3</sup> For more information on this metrics and analogous ones, see Artstein (2017) and Pustejovsky & Stubbs (2012). For some examples using Cohen's Kappa, see Pustejovsky & Stubbs (2012).

<sup>&</sup>lt;sup>4</sup> See the paper (Jakubíček et al., 2021) for further details.

<sup>&</sup>lt;sup>5</sup> This part and the example of the confusion matrix is inspired and adapted from some notes and slides of Bernardo Magnini's Computational Linguistics course held at the University of Pavia in 2021 (which was followed by this thesis' author).

In this matrix, there are four possibilities:

- 1. both the model and the gold standard assign the label "yes" (true positive)
- 2. both the model and the gold standard assign the label "no" (true negative)
- 3. the model assigns "yes" when the gold standard assigns "no", that is, the model assigns a label that it should not have assigned (*false positive*)
- 4. the model assigns "no" when the gold standard assigns "yes", that is, the model fails to assign a label that had to be assigned (*false negative*)

On the basis of these combinations, some metrics can be calculated in order to evaluate the model performance against the gold standard. We focus on three of them, specifically: *accuracy, precision,* and *recall*.

The simplest metrics, which can be compared to the raw agreement among the IAA metrics, is *accuracy*. Accuracy is the percentage of items (in the example above, words) correctly identified by the model. Following the previous Table, the items correctly identified by the model are those in which the model and the gold standard agree: the true positive and the true negative. Their sum is simply divided by the overall number of items, as can be instantiated by the following formula:

 $accuracy = \frac{true \ positive + true \ negative}{true \ positive + true \ negative + false \ positive + false \ negative}$ 

As Pustejovsky & Stubbs (2012: 173) point out, «while accuracy is easy to calculate, it can only give us a general idea of how well an algorithm performed at a task: it cannot show specifically where the task went wrong, or what aspects of the features need to be fixed». There are other metrics which come from Information Retrieval,<sup>6</sup> that give relevance to the items correctly identified by the model with respect to all the items identified by the model (*precision*) and to the items correctly identified by the model with respect to all the items in the gold standard (*recall*). In other words, precision determines how exact, recall how complete the performance by the model is with respect to the gold standard.

*Precision*, being the measure of how many items were accurately identified by the model (Pustejovsky & Stubbs, 2012: 174), can be instantiated by the following formula:

 $precision = \frac{true \ positive}{true \ positive + false \ positive}$ 

<sup>&</sup>lt;sup>6</sup> See Section 1.2.7 for a definition.

being *true positive* + *the true negative* the total number of the items correctly (*true positive*) or incorrectly (*false positive*) classified by the model.

*Recall*, being the measure of how many relevant items were identified by the model (Pustejovsky & Stubbs, 2012: 174) – relevant in the sense that we consider all the items that should have been identified against those which were actually correctly identified, can be instantiated by the following formula:

$$recall = \frac{true \ positive}{true \ positive + false \ negative}$$

being true positive + false negative the total number of items in the gold standard.

As we can see comparing precision and recall formulas to accuracy, *true negatives* are not taken into account neither for precision nor for recall: the reason for this is that is that true negatives are often exponentially more than the true positives in large datasets, thus computing the true negatives invalidates the accuracy measure. Precision and recall overcome this, by excluding the true negatives.

Finally, we mention that applying both measures to the same model can lead to opposite results (e.g., high precision and low recall, or vice versa). The ideal for a model is to be high both in precision and in recall, having the most in terms of exactness and completeness in the performance.

In order to have a global estimation of the performance of the model, precision and recall can be combined in a single measure, the *F-measure*.<sup>7</sup> The *F-measure* is calculated by finding the harmonic mean of the precision and recall (Pustejovsky & Stubbs, 2012: 175):

$$F = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

We do not discuss this metrics here, as it is not relevant for our evaluation. We will recall some of the metrics outlined in the previous paragraphs in the experimental part of this thesis (Chapter 5).

To sum up, we can outline the steps that should be following in an evaluation procedure:

- 1. select a task (e.g., the *outlier detection*)
- 2. create a dataset for the evaluation (e.g., *HAMOD dataset*)
- 3. assess the quality of the dataset through human annotation/evaluation (e.g., benchmark human evaluation on the outlier detection task)

<sup>&</sup>lt;sup>7</sup> F-measure also comes from Information Retrieval.

- 4. calculate the IAA and reach a gold standard
- 5. evaluate the model on the task
- 6. calculate the performance through the evaluation metrics (precision, recall, F-measure)

If any of the first steps lead to poor results, it has to be reiterated, following the specific requirements for each step (e.g., the human evaluation issues raised in Section 2.1.1), until the quality is good enough to proceed with the other steps.

Keeping this standard approach for the evaluation, we now move towards our specific case – the evaluation of distributional models, starting from a distinction between extrinsic and intrinsic evaluation methodologies.

# 2.2 Extrinsic vs. Intrinsic Techniques for the Evaluation of Distributional Models

As Bakarov (2018) points out in his survey, distributional (semantic) models (and their representations, word embeddings) are one of the most popular tools in NLP, but their nature and limitations are still not well understood. Therefore, the issue is how to *evaluate* their quality. As the author points out,

there is still no consensus in the scientific community about which evaluation method should be used: NLP engineers who are interested in dealing with downstream tasks (for instance, *semantic role labelling*) usually evaluate the performance of embeddings on such tasks [as we will see below, we refer to this as *extrinsic* evaluation methods], while computational linguists exploring the nature of semantics tend to investigate word embeddings through experimental methods from cognitive sciences [that is, *intrinsic* evaluation methods].

Schnabel et al. (2015) stressed the importance of the evaluation of distributional models and first systematized the existing approaches, by covering a wide range of evaluation criteria. For the first time, they also set the distinction among *intrinsic* and *extrinsic* evaluation methods. Since then, several similar surveys have been written on this subject, investigating extrinsic and intrinsic techniques, highlighting their pros and cons, and suggesting new research paths, especially as far as the combination of the two paradigms (*intrinsic* and *extrinsic*) is concerned.

Following these surveys (Schnabel et al., 2015; Bakarov, 2018; Wang et al., 2019; Torregrossa et al., 2021) in this Section we try to outline the distinction between the two

and we focus on *intrinsic* evaluation methods – that is, the one that we pursue within this thesis' project – and on the most common intrinsic techniques being exploited nowadays.

Before moving on, it is important to emphasize here that in the following part of this Section we will mainly talk about *word embeddings*, as most of the research has been done on how to evaluate word embeddings. Nevertheless, *distributional thesauri* – which is the main focus of this thesis and which we will compare to word embeddings – are distributional models themselves<sup>8</sup> and can be evaluated exploiting the same techniques used for word embeddings, especially as far as the intrinsic methods are concerned.<sup>9</sup>

First, we provide some definitions of *extrinsic* and *intrinsic* evaluation methods, following Schnabel et al. (2015) and Bakarov (2018).

By *extrinsic* evaluation we mean those techniques in which word embeddings are used as input features to a downstream NLP task (e.g., part-of-speech tagging, named entity recognition)<sup>10</sup> and in which differences among various word embedding models are measured in performance metrics specific to that task (Schnabel et al., 2015). Therefore, the performance of the model being measured on a dataset for NLP tasks functions as a measure of word embedding *quality for that specific task* (Bakarov, 2018).<sup>11</sup>

By *intrinsic* evaluation we mean those experiments in which word embeddings are compared with human judgments on word relations (Bakarov, 2018), or, in other words, intrinsic evaluation reflects the coherence between word embeddings and human judgement (Schnabel et al., 2015). Therefore, the quality of word embeddings is tested independent of specific NLP tasks (Wang et al., 2019) and – differently from extrinsic evaluation methods – those evaluations try to assess the *global quality* of the language representation (Torregrossa et al., 2021).

We will provide further sub-distinctions of intrinsic evaluation methods in the relevant Section (Section 2.2.2), by mentioning different approaches in these distinctions (Schnabel et al., 2015; Bakarov, 2018; Hadj Taieb et al., 2020).

#### 2.2.1 Extrinsic evaluation: an overview of the techniques and issues

According to Bakarov (2018), any downstream task could be considered as an evaluation method, given the definition of *extrinsic evaluation* as methods that measure the

<sup>&</sup>lt;sup>8</sup> As we highlighted in Section 1.4.

<sup>&</sup>lt;sup>9</sup> We will demonstrate this in the applied part of the thesis (Chapter 5) when we present the experiment that we have pursued exploiting the outlier detection task, which was originally intended for the evaluation of word embeddings only.

<sup>&</sup>lt;sup>10</sup> *Downstream tasks* in NLP are self-supervised-learning tasks (that is, that do not use human-annotated datasets to learn representations of the data used to solve the task) that utilize a pre-trained model or component.

<sup>&</sup>lt;sup>11</sup> We will discuss this latter assumption in the relevant Section (Section 2.2.1).

contribution of a word embedding model to an NLP downstream task (Schnabel et al., 2015). As extrinsic evaluation methods are not our focus of this thesis, we only mention here the most common tasks in which word embeddings are used (and thus intentionally or unintentionally evaluated)<sup>12</sup> that are reported in all the surveys we reviewed. Here follows a list of extrinsic evaluation methods-downstream tasks and a brief definition for each task (we keep the names used in Bakarov, 2018):<sup>13</sup>

- 1. *Part-of-Speech Tagging*. The goal is to is to assign a part of speech (noun, verb, adjective, etc.) of each word/token in the context of a sentence (Bakarov, 2018; Wang et al., 2019).
- 2. *Phrase Chunking (or Shallow Syntax Parsing).* The goal is to label segments of a sentence with syntactic constitutes (that is, to identify noun, verb, adjective phrases and their boundaries) (Bakarov, 2018; Wang et al., 2019).
- 3. *Named Entity Recognition.* The goal is to identify particular entity class (a person, an organisation, a brand or other) for a word (or multi-word) in the sentence (Bakarov, 2018; Torregrossa et al., 2021).
- 4. *Sentiment Analysis*. It is a sentence-level classification task where the model is asked to give a sentimental class for each sentence (Torregrossa et al., 2021) with a binary/ multi-level label representing positive or negative polarity of text's sentiment (Bakarov, 2018; Wang et al., 2019).
- 5. *Semantic Role Labelling*. The goal is to identify thematic roles<sup>14</sup> of arguments for various predicates within the sentence (Bakarov, 2018).
- 6. *Text Classification*. It a sentence-level/document-level classification task where the model has to classify text into different categories (Torregrossa et al., 2021).

We now briefly spot some issues with the extrinsic evaluation.

As we mentioned in the introductory part of this Section, while defining extrinsic evaluation, the performance of the model being measured on a dataset for NLP tasks functions as a measure of word embedding *quality for that specific task* (Bakarov, 2018). While it is sometimes assumed that word embeddings showing a good result on one task will show a good one also on others, this may not always be the case. As Schnabel et al. (2015) suggests, the assumption that there is a global ranking of word embedding quality and

<sup>&</sup>lt;sup>12</sup> As Bakarov (2018), many studies do not directly mention the problem of word embeddings evaluation, even though they use word embeddings in their experiment with a downstream task.

<sup>&</sup>lt;sup>13</sup> This is not a comprehensive list and the order in which we propose the tasks is random. We refer to Bakarov (2018) for specific papers on each specific technique and further examples.

<sup>&</sup>lt;sup>14</sup> A *thematic role* is the role that the referent of an argument plays in the event expressed by the verb. For example, thematic roles are "agent", "patient", "experiencer", "recipient", etc. (Jezek, 2016: 108, 117).

that high quality improves results on any kind of downstream task does not hold, as different task favour different embeddings. Furthermore, the authors prove that word embeddings performance on downstream tasks is not consistent across tasks.

Also, Bakarov (2018) points out that when word embeddings are used only to resolve *a specific task*, the evaluation will provide adequate insights on the word embeddings performance. On the other hand, if word embeddings are trained to serve a *wide range of different tasks*<sup>15</sup> which do not correlate between themselves (Schnabel et al., 2015 demonstrated this) no global evaluation for word embeddings can be obtained through extrinsic evaluation methods and none of these techniques can be used as an absolute metrics of word embeddings quality (Bakarov, 2018).

#### 2.2.2 Intrinsic evaluation: an overview of the techniques and issues

As we discussed in the previous Section (Section 2.2.1), one of the main drawbacks of extrinsic evaluation methods is that different tasks rely on diverse aspects of word embeddings, and satisfactory performance in one task does not necessarily imply equally satisfactory performance on another (Gladkova & Drozd, 2016). Therefore, intrinsic evaluation methods are pursued independent of specific tasks, in an attempt to assess the global quality of word embeddings (Wang et al., 2019).

Bakarov (2018) suggests some distinctions among intrinsic evaluation techniques. First, among absolute and comparative intrinsic evaluations:

- 1. *Absolute intrinsic evaluation*. In this type, manually created datasets of words are used to get human assessments,<sup>16</sup> and these assessments are compared with word embeddings. This approach relies on the idea that lexical semantics inferred by embeddings can be reported to lexical semantics determined by humans.<sup>17</sup>
- Comparative intrinsic evaluation. In this type, users give direct feedback on the embeddings themselves (Schnabel et al., 2015) and humans are asked to assess which model works better according to the comparable models outputs. The goal is not to estimate the absolute quality, but to find the most adequate embeddings in a given set (Bakarov, 2018).

<sup>&</sup>lt;sup>15</sup> Which differ in the kind of word embeddings feature they exploit, as different linguistic levels are targeted – morphology vs. syntax vs. semantics and so on.

<sup>&</sup>lt;sup>16</sup> Human judgments can be collected in laboratory on a limited set of examinees (*judgments collected in-house*) or on crowd-sourcing web platforms collecting an unlimited number of participants (*judgments collected through crowdsourcing*).

<sup>&</sup>lt;sup>17</sup> See the Section 2.2.4 for a discussion on this latter point.

Another distinction which interlaces with the previous one is the one between:

- 1. *Intrinsic conscious evaluation*. These methods which comprise most of the absolute and comparative techniques are designed to collect assessments which are results of conscious processes in a human brain. Some researchers point out that these answers may be biased by subjective factors and thus introduce variability (e.g., interpretation of meaning, words relations) (Bakarov, 2018).
- 2. *Intrinsic subconscious evaluation*. These (often interdisciplinary) methods collect assessments from the subconscious level of cognition by exploiting neuroimaging techniques. These methods should avoid the biases introduced by conscious evaluation (Bakarov, 2018).<sup>18</sup>

Finally, further intrinsic evaluation categories can be traced, beyond those already mentioned:

- 1. *Intrinsic thesaurus-based evaluation*. These methods do not compare word embeddings with experimental dataset, but with knowledge bases, semantic networks, dictionaries and manually-constructed thesauri (Bakarov, 2018).
- 2. *Intrinsic language-driven evaluation*. These methods are based on a comparison with data underlying in a language itself, such as graphematic representations of words, speech sound signals, or frequency of occurrence of words in corpora.

We do not discuss intrinsic subconscious, thesaurus-based and language-driven evaluation further (we refer to Bakarov, 2018, which is the most comprehensive survey on the topic that we have found). We focus, instead, on some of the *conscious intrinsic evaluation* techniques, and among those, on *absolute intrinsic evaluation*.

## Word (Semantic) Similarity

The first intrinsic evaluation we report is one of the most popular methods.<sup>19</sup> This method is based on the idea that distances between words in an embedding space could be evaluated through the human heuristic judgments on the actual semantic distances between these words (Bakarov, 2018). Here is how the experiment works.

<sup>&</sup>lt;sup>18</sup> We will not discuss these techniques beyond, as it is out of the scope of this thesis. We refer to Bakarov (2018), in which neuroimaging techniques (e.g., functional magnetic resonance imaging, electroencephalography, eye-tracking) are exploited.

<sup>&</sup>lt;sup>19</sup> This method roots go back to 1965 when the first experiments on human judgments on semantic similarity were conducted to test the distributional hypothesis (Rubenstein & Goodenough, 1965).

The human judge is given a set of pairs of words (which can be similar, such as *cup* and *glass*, or related, such as *song* and *musician*) and is asked to assess the degree of similarity/relatedness between these words, assigning a score on a numerical scale (usually from 0 to 1). The performance of the model is evaluated in terms of correlation<sup>20</sup> between the average scores by the human judgments<sup>21</sup> and the cosine distance between the word embeddings for the corresponding words in the model space (Baroni et al., 2014).

Therefore, the goal is to measure how well the notion of human perceived similarity is captured by the word vector representations and validate the distributional hypothesis where the meaning of words is related to the context they occur in (Torregrossa et al., 2021).

Not only word similarity is the most common method, but it is also the most criticized (Bakarov, 2018). Here we briefly report a list of some of the problems on which Faruqui et al. (2016) shed a light on:<sup>22</sup>

- 1. Human judgments are subjective, thus introducing biases in the assessments
- Connotative words contrary to denotative words tend to cause subjectivity based on cultural or personal criteria
- 3. The task is ambiguous, since different experiments tend to propose different definitions of semantic similarity
- 4. There is no clear distinction between semantic similarity and relatedness, and the two are often addressed jointly or, even worse, confused
- 5. Being the word pairs out of context, it is not easy to account for polysemy
- 6. These experiments tend to result into low inter-annotator agreement between human judges<sup>23</sup>
- 7. Numerical scores assigned by humans do not fully describe all the types of relations between words

<sup>&</sup>lt;sup>20</sup> Being this kind of experiment strongly rooted in psycholinguistics (that is, the study of the psychological processes involved in the use of language) (Bakarov, 2018), the correlation between human judgments and cosine distance is traditionally measured by Pearson and Spearman coefficients (Baroni et al., 2014).

<sup>&</sup>lt;sup>21</sup> IAA procedures are applied, and therefore many human evaluators need to be enrolled to get consistent results (see Section 2.1.1).

 $<sup>^{22}</sup>$  It is important to mention them, because in our work we try to address and overcome some of these problems (see Chapter 3).

<sup>&</sup>lt;sup>23</sup> This, in particular, is the issue we try to overcome by proposing the outlier detection task for the intrinsic evaluation of distributional models.

#### Word Analogy Task

It is the second most popular intrinsic evaluation method, after word semantic similarity. This method is based on the idea that arithmetic operations (namely, dot product between vectors) in a word vector space could be predicted by humans (Bakarov, 2018). Here is how the experiment works (following Bakarov, 2018):

given a set of three words, *a*, *b* and *c*, the task is to identify such word *d* that the relation *c*:*d* is the same as the relation *a*:*b*. For instance, one has words a = Paris, b = France, c = Moscow. Then the target word would be *Russia* since the relation *a* : *b* is *capital* : *country*, so one needs need to find the capital of which country is Moscow.

It is therefore a matter of solving the proportion:

a:b=c:dParis:France = Moscow:?

Then, the human performances on this task are compared to those by the embeddings and this should prove how good the model is at identifying those which are called linguistic regularities in Mikolov et al. (2013b), or analogies. The evaluation of the model is done in terms of proportion of questions where the nearest neighbour from the whole semantic space is the correct answer (Baroni et al., 2014).

What follows is a brief list of problems with this task:

- 1. There are no precise evaluation metrics (Bakarov, 2018)
- 2. The test can be subjective (Wang et al., 2019)
- 3. Analogies in human reasoning and logic can be rather different from those of word embeddings, as they do not encode our sense of reasoning (and thus the relationships found can be different (Wang et al., 2019)

## Concept Categorization (or Word Clustering)<sup>24</sup>

This method evaluates the capacity of word embeddings to distinguish semantic clusters among a set of words (Torregrossa et al., 2021). Here is how the experiment works.

Given a set of words, the goal is to split it into different categorical subsets of words. For example, given the set of words

sandwich, tea, pasta, water

<sup>&</sup>lt;sup>24</sup> We report this method here because it is slightly connected to the task we pursue, the outlier detection task, for the fact that they both focus on the ability to cluster, thus eliciting different aspects of this property (grouping words with similar features vs. finding the intruder in a cluster of words).

Both the human and the model (whose results on the task are then compared), should be able to split it as:

#### sandwich, pasta tea, water

being sandwich and pasta kinds of food, and tea and water kind of beverages.

The main problem with this task is subjectivity: humans can group words by inference using concepts that word embeddings can gloss over (Wang et al., 2019). Furthermore, it is important to specify the number of clusters that need to be distinguished within the set of words (Bakarov, 2018).

#### **Outlier** Detection

This method evaluates the same feature of word embeddings as the word categorization method - i.e., semantic clustering, but the task is different: the goal is to find a semantically anomalous word in an already formed cluster (Bakarov, 2018).

We do not discuss this method here, as an entire Section (Section 2.3) is dedicated to it. We only point out here that this method is less subjective and there is less amount of research on this evaluator as compared with that of word similarity and word analogy (Wang et al., 2019).

## 2.2.3 How to correlate intrinsic and extrinsic evaluation methods

Several authors (Bakarov, 2018; Wang et al., 2019; Torregrossa et al., 2021) acknowledge that correlations between intrinsic and extrinsic methods and evaluation metrics are not clear: it is not obvious to tell the quality of an embedding on a specific extrinsic task regarding its global performance on several intrinsic tasks (Torregrossa et al., 2021). Performance scores of word embeddings, when measured with intrinsic and extrinsic evaluation approaches, do not correlate between themselves. It is unclear what class of methods is more adequate (Bakarov, 2018).

Correlation studies are indeed necessary. To our knowledge, Wang et al. (2019) conducted the widest consistency study of extrinsic and intrinsic evaluation methods via correlation analysis. They compared – using correlation metrics – various intrinsic and extrinsic techniques on various distributional models, among those we mentioned in the previous paragraphs: word semantic similarity, word analogy, concept categorization, outlier detection as for *intrinsic* methods; Part-of-Speech Tagging, Phrase Chunking, Named Entity Recognition, Sentiment Analysis as for *extrinsic* ones. What they discovered is that by now word similarity, word analogy, and concept categorization are more effective intrinsic evaluators and they should be used jointly when testing a new embedding model; as for extrinsic methods, Sentiment Analysis correlates with word analogy, but the others do not correlate in a significant manner with the intrinsic tasks. Overall, evaluation of embeddings remains significantly complex as interaction mechanisms across different kinds of tasks and different methods are not well understood (Torregrossa et al., 2021) or still not significantly correlated (Wang et al., 2019). As Wang et al. (2019) acknowledge, there is still no perfect evaluation method because it is difficult to understand exactly how the embeddings spaces encode linguistic relations. While word embedding models perform well in downstream tasks, more work needs to be done in developing better metrics for the evaluation of the overall quality of word models (that is, intrinsic methods), which could shed light on the way linguistic relations are encoded in word space models, and therefore assess their quality.

## 2.2.4 Issues with the intrinsic evaluation: how can it be improved?

Follow Gladkova & Drozd (2016), focus on the common methodology behind intrinsic evaluation methods. They criticize the methodological premise of these methods, that is, the *interpretability* of word embeddings by humans as a measure of their quality. For example, with word similarity tasks we define the best word embedding model as the one that is closest to the human judgments. According to the authors, intrinsic methods attempt to transfer the traditional linguistic model of discrete word meanings and linguistic features onto the continuous semantic space. Assuming that a good embedding produces results that makes sense only in terms of traditional linguistic categories is limiting and it simply avoids unique – and *non-interpretable* – word embeddings features, first of all fluidity of meaning that is unattainable by traditional linguistic analysis. Indeed, by focusing on the structures that we expect the word embeddings to have, we might be missing the structures that they actually have. And, furthermore, actual intrinsic evaluation techniques do not address polysemy properly (excluding ambiguous words from evaluation datasets is not a solution, it simply avoids the matter).

What the authors hope for as an alternative is to embrace ambiguity and non-interpretability as an intrinsic characteristic of word embeddings, and to take a more exploratory approach, identifying the properties of a model rather than aiming to establish its superiority to others.<sup>25</sup>

 $<sup>^{25}</sup>$  As we will see in Chapter 6, we partly address this issue in our approach to the intrinsic evaluation technique we pursue, by focusing on a specific property of distributional models to create semantic clusters and by providing an in-depth analysis on qualitative – not only statistical – results of this task applied to the models.

# 2.3 The Outlier Detection Task as a Technique for the Intrinsic Evaluation of Distributional Thesauri and Word Embeddings

Among the *intrinsic* evaluation methodologies reviewed in the previous Section, here we focus on the *outlier detection* task. The reason for this is that, as we have already mentioned, the outlier detection task is the one we pursue within this thesis project. The task consists in identifying (*detection*), given a group of words, the word that does not belong in the group (that is, the *outlier*). In the following Sections we review the existing literature on this topic: we comment on the original proposal by Camacho-Collados & Navigli (2016), and on some subsequent works inspired by it, highlighting their drawbacks.

Before moving on, we briefly introduce here a note on terminology. The name *outlier detection* is also widely used in data mining<sup>26</sup> techniques: in these tasks, an *outlier* is «as an observation in a data set which appears to be inconsistent with the remainder of that set of data» (Ben-Gal, 2005; Boukerche et al., 2020). Common to our outlier detection task and data mining outlier detection is the idea of finding an anomaly within a set of items.

#### 2.3.1 Camacho-Collados & Navigli (2016): the first proposal and the task in brief

The *outlier detection* task is conceived by Camacho-Collados & Navigli (2016), who propose it as an alternative to the classical word similarity task, which, as we discussed in the previous Sections, suffers from low Inter-Annotator Agreement. Another drawback, according to the authors, is that the word similarity task is too simple: «words are simply viewed as points in the vector space. Other interesting properties of vector space models are not directly addressed» (Camacho-Collados & Navigli, 2016: 43).

The *outlier detection* task, instead, focuses on a specific property of distributional models which, according to the authors, has not been addressed properly: the *semantic coherence*, that is, «the capability of vector space models to create semantic clusters (i.e., clusters of semantically similar items)» (Camacho-Collados & Navigli, 2016: 43). The reason why this approach is preferable with respect to the other is that it provides a clear gold standard thanks to the high human performance on the task and thus its usability in applied tasks such as the evaluation of distributional models. An innovative aspect of this task is that it tests an interesting language understanding property not fully addressed to

<sup>&</sup>lt;sup>26</sup> Data mining is an interdisciplinary subfield of computer science and statistics with an overall goal of extracting information (with intelligent methods) from a data set and transforming the information into a comprehensible structure for further use (Gorunescu, 2011).

date: the ability to create semantic clusters in the vector space (Camacho-Collados & Navigli, 2016).

It is worth already mentioning here some weak spots regarding the latter statements in this paragraph.

As for the first point, the fact that their dataset is defined a gold standard (even though they present their experiment as a pilot study) seems a bit far from the reality, as we will see below: the dataset size, the number of evaluators involved, as well as the way in which they conducted the experiment, is, in our opinion, not enough to state this. A bigger dataset and a higher number of evaluators is needed, instead, and this is the goal of this thesis project.

As for the second, while the authors only mention this ability regarding vector space models, we point out here that we are also interested in the human linguistic ability to create and detect semantic clusters, on which we will focus in the analysis of the results of an experiment led using the outlier detection dataset (see Chapter 5 and Chapter 6). Comparing the quantitative results of the outlier detection task performance by human evaluators and distributional models is not enough, in our opinion: we want to examine and provide some insights on the qualitative differences between the two performances. This, to our knowledge of the studies we reviewed, has not been addressed enough.

Back to the task described in Camacho-Collados & Navigli (2016), the authors briefly explain how they designed the task. They took inspiration from a type of exercise used in language exams to assess the standard vocabulary, which is discussed in Richards (1976). Richards (1976) is concerned with the role of vocabulary in the syllabus in the light of the assumptions and findings of theoretical and applied linguistics and, among the teaching techniques he suggests, he proposes an exercise which relies on the idea that «words do not exist in isolation. Their meanings are defined through their relationships with other words, and it is through understanding these relationships that we arrive at our understanding of words» (Richards, 1976: 81). This exercise he proposes is aimed at understanding the meaning of a word by discriminating between members of a lexical set (a set of words that share many semantic properties in a specific syntagmatic context – context which, however, is not relevant here (Jezek, 2016: 162)). He proposes the following example:

In each of the following groups of words one word does not belong. other words have something in common which excludes this particular word. Please underline the word that does not belong in the group.

- 1. 1 swelling, lump, bump, mass, discoloration
- 2. ribs, skull, spine, femur, bone, kneecap, hair
- 3. stain, wart, blotch, discoloration, spot, mark

That is, basically, the outlier detection task: among a set of words with "something in common" (the *inliers*), an intruder (the *outlier*) has to be identified. The authors provide this example: among *apple*, *banana*, *lemon*, *book*, *orange* the outlier is book, as it is not a fruit like the others.

For their pilot study, the authors developed a dataset, the 8-8-8 *dataset*, consisting of eight different topics each made up of a cluster of eight words belonging to the chose topic and eight possible outliers – that had to be heterogeneous, varying their similarity to and relatedness with the elements of the cluster<sup>27</sup> (Camacho-Collados & Navigli, 2016). Most of the words contained are named entities: "IT Companies", "German Car Manufacturers", "Apostles of Jesus Christ", "South American Countries" (Camacho-Collados & Navigli, 2016).<sup>28</sup> These clusters were created by 4 annotators and then they organized the dataset for the evaluation by creating 64 sets of 8 words + 1 outlier.<sup>29</sup>

The first step of the evaluation was assessing the human performance of the eight annotators in the task, measuring it through accuracy.<sup>30</sup> Each annotator was given eight randomized clusters (8 + 1) without any other information (i.e., the specification of the topic) and asked to detect the outlier in each set of words.

We may notice that if (as it seems to be)<sup>31</sup> Camacho-Collados & Navigli (2016) involved as evaluators the same people who also served as annotators in the creation of the dataset, this would be methodologically problematic. The results of the human performance evaluation may be conditioned by the fact that the evaluators may knew the content of the clusters and which the outliers were. Furthermore, the annotators could take a second turn of the task and use an external help to detect the outlier (web search). We believe that also this is methodologically problematic: undertaking a second turn and, in addition, use an external source, could, again, affect the evaluation result.

Unsurprisingly, the results of the experiment in terms of the accuracy (and therefore, inter-annotator agreement) is extremely high:

- 1. 98.4% in accuracy without any external help (first turn)
- 2. 100% in accuracy with external help (second turn)

Another reason for this striking success may be the limited size of the dataset. We believe that a bigger dataset can lead to more statistically relevant and significant results. These results, as Camacho-Collados & Navigli (2016) highlight, are in contrast with the

<sup>&</sup>lt;sup>27</sup> As we will discuss in the following Chapters (Chapter 3 and Chapter 4), the authors do not specify, neither exemplify these notions, which are notably obscure.

<sup>&</sup>lt;sup>28</sup> As we will see in the following Chapters (Chapter 3, in particular), we want to avoid named entities sets, as we are more interested in basic vocabulary (and thus common nouns).

<sup>&</sup>lt;sup>29</sup> Being the outliers 8, for each topic, there are 8 possible combinations of cluster words + 1 outlier (overall: 64).

<sup>&</sup>lt;sup>30</sup> See Section 2.3.

<sup>&</sup>lt;sup>31</sup> The authors do not really mention this clearly: they refer to both the people involved in the creation of the cluster and those who evaluated the dataset as "annotators". We therefore assume here that the two groups coincide.

evaluation performed in word similarity (low inter-annotator agreement). Assuming that the success of the human performance guarantees the validity and quality of the dataset, they proceed with the second step, that is testing the performance of word embeddings models in the outlier detection task on various corpora, in comparison.<sup>32</sup> We refer to the original paper for the results, and we will present the evaluation metrics they used in detail in Chapter 5.<sup>33</sup> What emerges from the evaluation of these word space models is the capability of word embeddings to create semantic clusters is worth investigating further, although the embeddings performance in the task (40%-73% in accuracy) is significantly lower than the human one (98.4%-100%).

Finally, the authors provide a brief qualitative analysis of the errors produced by the models, discovering that the main reasons for the wrong selection of the outliers are:

- 1. the lack of meaningful occurrences for a given word in the reference corpus: the more frequent the word is, the more accurate the word vector representation
- 2. the closeness of the outlier to the cluster of semantically similar/related items: the closer the word is to the cluster, the more difficult is to distinguish the outlier from the cluster
- 3. ambiguity in meaning and part of speech
- 4. synonyms (that is, items with more than one lexicalization, especially as far as named entities are concerned)

In Chapter 6, we will address our qualitative analysis of the human and word space models performance following these directions of interpretation.

To sum up, what Camacho-Collados & Navigli (2016) stress is that the word embeddings perform reasonably well in the task, but still far from human performance, which is relevant as opposed to the word similarity task, proving the reliability of the gold standard, which can be exploited in intrinsic evaluation tasks.<sup>34</sup>

<sup>&</sup>lt;sup>32</sup> We will not discuss here which kind of models and corpora they used, please refer to the study (Camacho-Collados & Navigli, 2016).

<sup>&</sup>lt;sup>33</sup> One of them is accuracy, which we have already presented in the previous Section (Section 2.1.2). the other one is OPP (outlier position percentage), a custom-made metrics which is more accurate as it gives the average percentage of the right answer. We will discuss this further in Chapter 5.

<sup>&</sup>lt;sup>34</sup> The 8-8-8 *outlier detection dataset*, the guidelines given to the annotators as part of the pilot study, and an easy-to-use Python code for evaluating the performance of word vector representations are available at: <u>http://lcl.uniroma1.it/outlier-detection</u> (last access: 24/06/2022).

#### 2.3.2 On the outlier detection task: a literature review

After Camacho-Collados & Navigli (2016), a few studies – which we briefly review herein – have addressed and exploited the outlier detection task, undertaking the following directions:

- 1. enlargement of the dataset (Blair et al., 2017; Gamallo et al., 2018; Andersen et al., 2020)
- 2. automation of the creation of the dataset (Blair et al., 2017)
- 3. combination with an extrinsic evaluation technique (Blair et al., 2017)
- 4. addition of new languages (multilingual dataset) (Blair et al., 2017; Gamallo, 2018; Andersen et al., 2020)

Blair et al. (2017) report the fact that dataset for the intrinsic evaluation «require manual annotations that suffer from human subjectivity and bias and are not multilingual». Therefore, they present WikiSem500, an outlier detection dataset which is fully automated using Wikidata and Wikipedia as graphs to derive semantic clusters, and which is also diverse in the number of included topics, words and phrases, and languages (Blair et al., 2017). Covering the dataset a wide range of domain knowledge from Wikipedia, with its 500 clusters, the human evaluation has been restricted to a small portion of it: first, 60 clusters were tested, with a pretty low accuracy (68.9%, based on 447 responses – it is not clear how many participants were involved); second, a smaller portion of it (15 clusters) was evaluated by 6 participants, resulting into 93% of accuracy. As for the embedding evaluation, we refer to the paper (Blair et al., 2017), but we point out that the results are worse than those in Camacho-Collados & Navigli (2016) in terms of the accuracy. Finally, in an attempt to correlate intrinsic evaluation and downstream tasks, they investigate the correlation between the outlier detection dataset performance and one type of extrinsic evaluation, that is, sentiment analysis: this results to be substantial.

Gamallo (2018) expands the 8-8-8 *dataset* (Camacho-Collados & Navigli, 2016) by proposing the *12-8-8 dataset* (thus manually adding 4 new clusters) and translates it into Portuguese. Unfortunately, their evaluation of the human performance is not appropriate, from a methodological point of view. What the author defines as 100% inter-annotator agreement is described as follows:

two annotators were asked to create four new topics, and for each topic to provide a set of eight words belonging to the chosen topic, and a set of eight heterogeneous outliers. One of them proposed all the words in less than 15 minutes, and the other annotator just agreed with all the decisions taken by the first one. As we can see, the task itself is not performed properly<sup>35</sup> with the human evaluators (who, anyway, are not numerically relevant) and the author proceeds straight to the distributional model evaluation.

The latest study we have found, Andersen et al. (2020), addresses what they consider 8-8-8 *dataset* (Camacho-Collados & Navigli, 2016) limitations:

- 1. small number of clusters: this affects the accuracy (the errors have a significant impact in its computation) and the coverage of concepts in the vector space
- 2. ambiguity: the presence of semantically and morphologically<sup>36</sup> ambiguous words may be wrongly classified as outliers, even when they are not
- 3. multi-word expression: the mismatch between the purely compositional meaning (or, the meaning of the single tokens) and the actual meaning of the expression can affect the results
- 4. low frequency of words in the reference corpora (which had already been underlined in Camacho-Collados & Navigli, 2016; see the previous Section; Section 2.3.1).

They also mention Blair et al. (2017) WikiSem500, pointing out that the automatic path they have followed in the construction of the dataset does not lead to remarkable results. The main problems of WikiSem500 due to the automatic procedure are, according to Andersen et al. (2020):

- 1. vague semantic connection within the clusters due to inconsistencies within Wikipedia graphs
- 2. ambiguous words
- 3. repetition of the same outlier in the same test group
- 4. repetition of the same word in different spellings in the same test group
- 5. infrequent words
- 6. lack of correspondences between the data for each language included

Andersen et al. (2020) propose the 50-8-8 dataset as a manually annotated dataset, containing unambiguous single-token words (this meaning that there are no multi-word expressions), chosen if they reached a minimum frequency of 350 occurrences in the reference corpora. The dataset is multilingual and parallel, as it covers three languages and the versions for each language overlap. Although the experiment they have conducted in the three languages comparing the outlier detection task and the word analogy task<sup>37</sup> on

<sup>&</sup>lt;sup>35</sup> We recall the human evaluation aspect in Section 2.1.1.

<sup>&</sup>lt;sup>36</sup> That is, that belong to various parts of speech.

<sup>&</sup>lt;sup>37</sup> See Section 2.2.2.

various word space models and corpora show remarkable results in the diverse performance of the three languages, it is important to notice that the human performance on the outlier detection task on this dataset has not been tested.

In the following Table (Table 2), we summarize the main aspects of the studies we reviewed in this Section (adding the first study, Camacho-Collados & Navigli, 2016). We compare the four datasets (column 2 to 5) according to the parameters listed in column 1.

Table 2. Main differences between the various datasets reviewed (Camacho-Collados & Navigli, 2016; Blair et al.,2017; Gamallo, 2018; Andersen et al., 2020)

	8-8-8 dataset	WikiSem500	12-8-8	50-8-8
	(Camacho-	(Blair et al., 2017)	(Gamallo, 2018)	(Andersen et al.,
	Collados &			2020)
	Navigli, 2016)			
dataset creation	manual	automatic	manual	manual
number of	8	500	12	50
sets/clusters				
number of test	64	~2 700 per language	96 per language	400 per language
cases		(13 314 overall)	(192 overall)	(1 200 overall)
number of	1	5	2	3
languages	(English)	(English, Spanish,	(English,	(English,
		German, Chinese,	Portuguese)	German, Italian)
		Japanese)		
number of	8	/	2	not specified
annotators				
number of	8 (on the overall	6 (on 15 clusters)	2	/
evaluators	dataset)			
human	98.4%-100%	68.9%-93%	100% (unreliable)	/
performance				
kind of words	mainly Named	mainly Named	mainly Named	Named Entities,
	Entities	Entities	Entities	common nouns

To sum up, what we notice is that, in general:

- 1. automatic techniques, although being less time-consuming, yet do not lead to satisfying results, and manual approaches are preferred
- 2. there is some need to increase the size of the datasets
- 3. multilingual datasets are preferred (also for multilingual comparisons)
- 4. most of the dataset mainly contain Named Entities in the clusters
- 5. ambiguity and multi-word expressions tend to be seen as an issue, affecting the results of the task

As far as the latter point is concerned, we believe that ambiguity and multi-word expressions are instead an interesting focus, being one of the significant issues that affect static distributional models quality. We believe that removing them from the datasets (as
done in Andersen et al., 2020), takes away the opportunity to have an insight on how distributional models deal with semantic/part of speech ambiguity and multi-word expressions.

In the following Chapter (Chapter 3) we will introduce how we conceived our dataset for the outlier detection (HAMOD dataset), what it shares with the work by Camacho-Collados & Navigli (2016), and in what it is different.

# Part 2. HAMOD Dataset: History, Methodology, and Building Process

# Chapter 3. HAMOD: a High Agreement Multilingual Outlier Detection Dataset

This Chapter marks the beginning of the applied parts of this thesis. That is, the technical description of a dataset to perform the outlier detection task as previously described in Section 2.3<sup>1</sup> (Chapter 3); the methodology followed in order to build HAMOD dataset and the steps that led to its implementation (Chapter 4); its application in an experiment aimed at evaluating humans, distributional thesauri, and word embeddings performances in the outlier detection task (Chapter 5); finally, the analysis of the results in a multilingual perspective, and some consequent remarks which may lead to some further developments of the project (Chapter 6).

Back to this Chapter, here follows a brief outline of its content.

In Section 3.1 we discuss the motivations for and the purposes of this project, and we recall the differences from its reference study (Camacho-Collados & Navigli, 2016)<sup>2</sup> which served as background and starting point for our work.

In Section 3.2 we provide some details on the formal layout of the dataset and, in Section 3.3, we describe its initial state before this thesis author's implementation, in an attempt to trace its development from 2019, when the project started, and all the contributors, who we cite in the respective parts of this Section.

In Section 3.4 we present the outcome of the dataset implementation,<sup>3</sup> that is, the current state of the dataset with its 128 sets, 2048 words per language and 7 languages covered, for a total of 14336 words, as well as some statistics in this respect.

<sup>&</sup>lt;sup>1</sup> With specific adaptations from Camacho-Collados & Navigli, 2016 to our project, as we will see thereafter in this Chapter (Section 3.1.2).

<sup>&</sup>lt;sup>2</sup> See Section 2.3 for a wider discussion on this work.

<sup>&</sup>lt;sup>3</sup> The methodological steps we undertook in order to achieve this are described in Chapter 4.

# 3.1 Motivation, Purposes, and Background Study

#### 3.1.1 What HAMOD stands for and what it tells about its purposes

HAMOD is the acronym standing for *High Agreement Multilingual Outlier Detection* and it refers to the dataset for exercising the outlier detection task that aims at high Inter-Annotator Agreement and is built in a multilingual perspective (Jakubíček et al., 2021). The name of the dataset itself reflects its purposes and scopes:

- 1. *High Agreement:* the goal is to achieve a reliable, reproducible evaluation methodology based on high Inter-Annotator Agreement (IAA)<sup>4</sup> among human evaluators.<sup>5</sup>
- 2. *Multilingual:* the dataset includes several languages. This allows comparisons between human performances with diverse native languages, as well as between different distributional models in different languages.
- 3. *Outlier Detection:* i.e., the name of the task<sup>6</sup> we ask the human evaluators and the distributional thesauri to perform, which can also be considered as intrinsic evaluation methodology. We briefly recall it here as the task where a human or machine is presented with a set of words, out of which one is a so-called *outlier*: a word that "does not fit" to the others (Jakubíček et al., 2021).

As for the first point, in Section 2.2.4 we have already addressed the issues related to finding reliable evaluation methodologies as far as intrinsic techniques are concerned. Here we only recall that the many intrinsic evaluation methodologies – word similarity task, first – suffer from rather low Inter-Annotator Agreement (Jakubíček et al., 2021; Bakarov, 2018).

Therefore, the main goal of this project is to build a dataset which can then be used in a task performed by human evaluators whose outcome results in a high agreement between them, thus providing a benchmark or gold standard for intrinsic evaluation techniques. On this basis, the distributional thesaurus evaluation can be consequently carried out, and human performances can be compared to distributional models' performances exactly on the same task.

<sup>&</sup>lt;sup>4</sup> See Section 2.1.1.

<sup>&</sup>lt;sup>5</sup> This is fundamental if we want to use these data in the evaluation of distributional thesauri or word embeddings.

<sup>&</sup>lt;sup>6</sup> See Section 2.3 for an extensive explanation and discussion of the outlier detection task in theory.

Regarding the second point, a dataset containing several languages was needed, for several reasons. As in many other fields, most of the intrinsic evaluation methodologies are built for English language only (Bakarov, 2018; Hadj Taieb, 2020; Andersen et al., 2020). Even Camacho-Collados & Navigli (2016) only focused on English language (and on a small-sized dataset, as we have already discussed in Section 2.3). Moreover, Sketch Engine<sup>7</sup> core characteristic is its multilingual perspective. Its corpora – and consequently, their distributional thesauri – cover almost one hundred languages: developing an evaluation methodology that could be useful not only for comparison of different approaches to distributional thesauri, but also for the comparison of the same approach in different languages was crucial. Indeed, the dataset could be potentially and easily expanded by translation and adaptation to several other languages, if necessary.

It is worth mentioning here that the dataset is not strictly parallel, but comparable instead (Jakubíček et al., 2021): this means that the word it contains can sometimes lack of exact corresponding translation in the other languages, due to ambiguity or mismatch.<sup>8</sup>

Finally, for what concerns the third point, in Sections 2.2 and 2.3 we have already discussed widely what the outlier detection is in the scope of the intrinsic evaluation methods, what are its main advantages compared to other techniques and also its drawbacks. Still, we believe that – and we have already proved it in preliminary studies (Rychlý, 2019; Jakubíček et al., 2021)<sup>9</sup> – the outlier detection is a promising path in the evaluation of distributional thesauri and word embeddings.

To sum up, the main motivation for building HAMOD dataset is to assess human Inter-Annotator Agreement on the outlier detection task and – if high – to use it as a solid, reliable, and reproducible basis for the evaluation and comparison of diverse distributional models, among which we focus on distributional thesauri and word embeddings.

#### 3.1.2 In what HAMOD differs from 8-8-8 dataset and why it improves it

In Section 2.3 we extensively discussed the work done in Camacho-Collados & Navigli (2016). While it is unquestionable that our project is grounded on theirs, however there are several differences for what concerns the aim of the study, the dataset construction, its size, its application, and the kind of experiment that is held using that data.

While Camacho-Collados & Navigli (2016) work is more projected in a computational perspective, ours is more lexicographically oriented. Their main goal is to provide a new

<sup>&</sup>lt;sup>7</sup> <u>https://www.sketchengine.eu/</u> (last access: 24/06/2022). See also Section 1.4.1 for a brief introduction to Sketch Engine.

<sup>&</sup>lt;sup>8</sup> We will discuss this issue in more detail and explain how we address this in Section 4.2.

<sup>&</sup>lt;sup>9</sup> See also Section 2.3.2 when discussing the results of these previous work that relied on the original dataset by Camacho-Collados & Navigli (2016).

reliable framework for an intrinsic evaluation of word vector representations, to test the capability of vector space models to create semantic clusters in the space (Camacho-Collados & Navigli, 2016). Our goal is to apply their idea to other distributional models (namely, the distributional thesaurus in Sketch Engine, not only to word embeddings) and test a relevant range of basic, high-frequency vocabulary. This because Sketch Engine Thesaurus (and its underlying distributional model) serves mainly lexicographic purposes (that is, the retrieval of synonyms and similar words), thus the target of words in Camacho-Collados & Navigli (2016), that is, named entities, was not relevant for our scope.

As we will argue in the following part of this Section, we decided to accurately conform to the criteria that lead to their dataset construction. However, we diverted from them for several reasons.

First, we only involved one human annotator (that is, this thesis' author) for the creation of the dataset.<sup>10</sup> Also, the overall dataset is significantly different in size, number and kind of words included, and languages covered.<sup>11</sup>

Moreover, its application differs: Camacho-Collados & Navigli (2016) involve eight human evaluators and apply the task only to word embeddings.<sup>12</sup> By contrast, our project's final goal is to engage a high number of human evaluators – at least 80 per language (Jakubíček et al., 2021)<sup>13</sup> – and to evaluate both distributional thesauri built from Word Sketches and word embeddings.

In short, we outline the main differences between the two projects in the following Table (Table 1). We compare 8-8-8 *outlier detection dataset* in column 2 (Camacho-Collados & Navigli, 2016) to *HAMOD dataset* in column 3 according to the parameters in column 1.

<sup>&</sup>lt;sup>10</sup> This can be explained by the fact that the author was fully dedicated to this project, that the initial part of the dataset was already built and that finding other annotators would have lengthened the timing of this thesis. Also, keeping track of all the parts of the dataset, avoiding repetitions, and enhancing more homogeneity, have proved to be easier if done by one single annotator.

<sup>&</sup>lt;sup>11</sup> See the Table below for some quantitative data.

<sup>&</sup>lt;sup>12</sup> See Section 2.3.1.

<sup>&</sup>lt;sup>13</sup> This goal goes beyond the scope of this thesis, where the focus is on the dataset construction and the first evaluation experiment.

 Table 1. Main differences between HAMOD dataset and 8-8-8 outlier detection dataset (Camacho-Collados & Navigli, 2016)

	8-8-8 outlier detection dataset	HAMOD dataset
number of annotators	8	1
number of evaluators	8	22
number of	8	128
sets/clusters		
number of words	128 words	2048 words per language
number of languages	1 (English)	7 (Czech, German, English, Estonian,
		French, Italian, Slovak)
kind of words	mainly Named Entities	common vocabulary
type of evaluated data	word embeddings	distributional thesauri, word
		embeddings

In the following Section (Section 3.2) we will focus on HAMOD dataset formal structure: how sets are defined, how inliers and outliers are selected and which specific requirements we added in order to collect the words for the dataset.

# **3.2 Dataset Formal Description**

#### 3.2.1 Definition of the sets: semantic categories, topics, inliers, outliers

HAMOD dataset structure is simple. It consists of *sets* of words (also called *exercise sets* in Jakubíček et al. (2021) – the terms are interchangeable). Our sets correspond to those called *clusters* in Camacho-Collados & Navigli (2016); there is a correspondence between the two in their base structure and their composition (in terms of the relation between the words contained in each one); however, there are some differences in the kind of words and group of words selected, as we will shortly explain below.

Each set is a made by a group of sixteen words, distributed in two sub-groups:

1. the *inliers*, i.e., 8 words belonging to a *semantic category* or pertaining to a specific *topic*<sup>14</sup>

<sup>&</sup>lt;sup>14</sup> We will discuss these keywords in the following part of the Section.

2. the *outliers*, i.e., 8 words which *do not fit* to the first 8 at different degrees (that is, they do not have relevant or enough properties to belong to the semantic category or the topic of the inliers)

In short, we call the words belonging to the semantic category/topic *inliers* and the words which do not fit *outliers*. Each set has a *label* assigned, defining the content of the set, or more specifically, which kind of superordinate item the inliers instantiate.

As for the inliers, a key point now is to define the concepts of *semantic category* and *topic*. As far as our project is concerned, we define a *semantic category* as sets of words referring to entities, properties or events sharing some common features. The notion of semantic categorization (or classification) is well grounded in lexical semantics' literature, and it lays on the assumption that the vocabulary of a language is not just a collection of words, but it is structured at various levels, among which one is the semantic level (Cruse, 2000: 179). Also, as Jezek (2016: 102) debates, «words may be grouped into coherent semantic classes by looking at the category of things they refer to (the so-called ontological category)». As the author recalls, this idea relies on Lyons (1977) suggestion that an adequate basis for semantic classification of words is the ontological correlate.

In this project we use the notion of semantic category to define more or less prototypical sets (or clusters) of items (and thus, words that refer to these items). Let us consider an example. "Means of Transport" (one of the sets in the dataset) contains items which share the feature of being human-made artifacts, used by human beings to move around, fuelled by some kind of source of energy. The features are implicitly reflected in the label assigned to each set by the annotator (as we mentioned above), which is only needed to identify the set.<sup>15</sup> Examples of semantic categories (included in our dataset) are: "School Subjects", "Means of transport", "Clothes", "Parts of Skeleton", "Trees".

Let us take as an example, the set "Means of Transport". Here are its 8 inliers:<sup>16</sup>

#### Means of Transport

- 01 motorbike
- 02 ship
- 03 car
- 04 tram
- 05 bus
- 06 train
- 07 plane
- 08 helicopter

<sup>&</sup>lt;sup>15</sup> The label is not used in the task nor signaled to the human evaluators, otherwise knowing which kind of topic or category is being used would facilitate the evaluators' task to recognize the outlier among the inliers.

<sup>&</sup>lt;sup>16</sup> This and the following examples in this Section are reported in English for the sake of simplicity. The corresponding sets in the other languages can be found in Appendix 1.

Even if there are several types of means of transport (road, water, and flying vehicle), they all share the property of being human-made artifacts used by human beings to move around. Another example of a set, which is more specific,<sup>17</sup> derived from this, is "Road Means of Transport":

## **Road Means of Transport**

01	car
02	bus
03	taxi
04	bike
05	motorbike
06	trolleybus
07	van
08	scooter

Λ1

In this set, *train*, *plane* and *boat* would be outliers because they are not means of transport used on the road (they are used on rails, air, and water, instead).

Another kind of set is the one based on *topics*. We intend *topics* as domains or even semantic fields, a broader and looser concept which can include items of different nature and do not necessarily require items to share several relevant properties as we mean with the semantic categories. Examples of topics are: "Music", "Informatics", "Linguistics", "Cooking".

Sets based on *topics* are less strict: they can contain words that do not necessarily shared relevant features (e.g., abstract and concrete, human-made and natural objects, events, and properties can be mixed). For example, "Music" is a set based on a topic. Here are its 8 inliers:

	Music
01	note
02	song
03	guitar
04	rock
05	flute
06	sound
07	microphone
08	singer

In this case, items may belong to different semantic categories: for example, *guitar*, *flute* are musical instruments, *singer* and is an artistic profession, *rock* and *note* are abstract entities etc. All the inliers clearly pertain to the same topic.

<sup>&</sup>lt;sup>17</sup> From a taxonomic point of view: road means of transport are types of means of transport.

One critical issue is to define which kind of semantic categories and topics we want to select in our dataset. What follows is a list of requirements that restrict the set candidates and the kind of words included.

- Named Entities and Proper Names. As opposed to Camacho-Collados & Navigli (2016), we avoid Named Entities and proper names. "Solar System Planets", "South American Countries", "Presidents of Czech Republic" are not suitable candidates for a set, whereas "Musical Instruments", "Shapes" and "Professions" are suitable candidates for a set.<sup>18</sup>
- General Knowledge. Categories and topics should belong to some general knowledge; thus, we avoid narrow and domain-specific categories or topics. "Farm Animals" (cow, pig, goose, dog, etc.) is a suitable category, but "Dog Breeds" (basset hound, bohemian shepherd, poodle, bulldog, etc.) may be too specific and may not belong to some shared knowledge.
- *3. 12-year-old Vocabulary.* Words chosen as inliers and outliers should be easily understood by a 12-year-old person and be part of their vocabulary. What can be part of their vocabulary and what cannot is hard to define, therefore we tested this (see Section 4.3). Another possibility is to try to use frequent (compared to a reference corpus) and not too domain-specific vocabulary.
- 4. *Semantics, nothing else.* Although it has been already specified, criteria for the identification of the sets must be semantic. "Interrogative Pronouns" or "Time Preposition" are grammatical categories, but not semantic: they are not suitable candidates for a set, as the criterion would be syntactic or morphological.

Once the semantic categories and topics are defined, and the inliers (that is, the 8 words belonging to the categories) are set, *outliers* need to be identified. Inliers are among themselves similar and related; outliers have instead a lower degree of similarity and relatedness, at different hierarchical levels (see the sub-sets below), and they do not share all the relevant properties that the inliers share among each other. This is therefore the criterion for the differentiation of the inliers and the outliers, and thus the key to solve the outlier detection task when looking for the word that does not fit.

We follow Camacho-Collados & Navigli (2016) in dividing the 8 outliers in 4 sub-sets (thus, 2 words per each sub-set), defining each sub-set as follows:

1. sub-set 1: two words that are closely related to the inliers, thus sharing a high number of features with them, but not enough to be part of the inliers. For example, in "Road Means of Transport" (see above), *skates* and *airplane* are means of transport, but an airplane is a flying vehicle, not a road one, and the

<sup>&</sup>lt;sup>18</sup> The motivation for this is explained in Section 3.2.

skates are more like a sport equipment, therefore they are distinguished from the inliers.

- 2. sub-set 2: two words that are less related by sharing less features. For example, *road* and *roundabout* are always human-made entities but cannot be said to be means of transport, but they are still related to the semantic category of driving.
- 3. sub-set 3: two words that are even less related, but still pertaining to the semantic category. They may be words referring to different kind of entities (from concrete to abstract, in the case of our examples) or to events or properties. For example, *traffic* and *car\_crash* refer to kinds of events that can involve road vehicles.
- 4. sub-set 4: two words that are not related at all to the inliers. They can be random words. For example, *rugby* and *toaster* do not share any feature with the inliers nor pertain to the semantic category.

To sum up, the outliers of the set "Road Means of Transport" are:

#### **Road Means of Transport**

01	skates
02	airplane
03	road
04	roundabout
05	traffic
06	car_crash
07	rugby
08	toaster

For what concerns topics, the selection of the first six outliers does not have to follow the same hierarchy as for semantic categories, as different entities, events, and properties may be included among the inliers. The last two outliers always need to be random words. See for example the outliers for the set "Music" (mentioned above):

#### Music

- 01 letter
- 02 colour
- 03 drawing
- 04 sculpture
- 05 writer
- 06 painter
- 07 picnic08 pocket

#### 3.2.2 Formal requirements for the encoding of the dataset and format

Besides from the selectional requirements (that is, which kind of semantic categories and words can be included in the dataset), we also have some formal requirements which mainly concern the encoding of the words in the sets.

1. Parts of Speech. Sets only contain words with the same part of speech, considering inliers and outliers altogether. There are only-noun, only-verb and only-adjectives sets in the current state of the dataset. See the following example in parallel:

part of speech	nouns	verbs	adjectives
	Astronomical Objects	Verbs Cognition	Colours
inliers	star	know	red
	planet	believe	blue
	black hole	think	green
	satellite	understand	yellow
	galaxy	remember	purple
	asteroid	forget	pink
	meteorite	meditate	orange
	comet	interpret	brown
ouliers	orbit	love	dark
	eclipse	hate	bright
	astornaut	listen	wooden
	telescope	hear	glass
	gravity	quarrel	striped
	light_year	confirm	dotted
	fountain	wash	sad
	peace	remove	low

- Multiword Expressions. In order to be processed by the evaluation script in the experiments using the dataset,<sup>19</sup> multiword expressions which are allowed in the dataset need to be encoded with an underscore ("\_") joining each word of the term. See, for example, *peanut\_butter*, *salle\_de\_bain* (Eng., 'bathroom'), *cambiamento\_climatico* (Eng, 'climate change').
- *3. Lemmas.* Words have to be encoded as their lemma. For example, singular form is the lemma for a noun, singular masculine for an adjective, infinitive form for a verb in Italian. Plural forms are not accepted unless the words only have a plural form (e.g., 'trousers' in English).

<sup>&</sup>lt;sup>19</sup> See Chapter 5.

Finally, for computational purposes our dataset needs to be stored in an online repository (GitLab,<sup>20</sup> in our case) and thus the material is organized as follows. Each set has its own folder; in each folder there is a simple *.txt* file containing the inliers and the outliers of each single language (therefore, in our case, 7 files per folder). The structure of each .txt file is as follows: inliers can be in a random order, as there is no hierarchical relation among the 8 elements; outliers need to follow the order outlined in the paragraphs above (Section 3.2). Inliers and outliers must be separated by an empty line:

inlier 1 inlier 2 inlier 3 inlier 4 inlier 5 inlier 6 inlier 7 inlier 8 <Empty line> outlier from sub-set 1 outlier from sub-set 1 outlier from sub-set 2 outlier from sub-set 2 outlier from sub-set 3 outlier from sub-set 3 outlier from sub-set 4 outlier from sub-set 4

See the set "Means of Transport" as an example:

**Means of Transport** motorbike ship car tram bus train plane helicopter exercise\_bike treadmill pavement road driver pilot needle

<sup>&</sup>lt;sup>20</sup> GitLab: <u>https://about.gitlab.com/</u> (last access: 24/06/2022).

shoe

As we will recall in Section 3.4, the dataset can be consulted also in a more readable format in Appendix 1.

# **3.3** The Original Dataset: its Development, Extent, and What Needed to be Improved

HAMOD dataset building project has started in 2019 and was initially developed by a group of students at the Masaryk University in Brno within the scope of a course taught by doc. Mgr. Pavel Rychlý (Ph.D.), associate professor at the Faculty of Informatics.<sup>21</sup> The overall number of sets originally put together by the students was 48, divided into 6 groups of 8 sets each.<sup>22</sup> What follows (Table 2) is a list of the sets (more precisely, their label) as originally submitted by the students. The items in bold in the second column are the sets that were kept after the post-selection described in the following paragraphs.

group number		set name
group_1	1	Illnesses
	2	Colours_II
	3	Materials
	4	Sport_Verbs
	5	Colours_I
	6	Human_Feature_Positivity
	7	Number
	8	School_Subjects
group_2	9	Furniture
	10	States_of_USA
	11	Pronouns

Table 2. List of the original 48 sets submitted by the students, grouped in 6 sub-groups

<sup>21</sup> Doc. Mgr. Pavel Rychlý (Ph.D.) is also the co-founder of Lexical Computing. Students were from the Faculty of Arts' degree in Computational Linguistics. The course was Language Modeling (PA154), held in spring 2019.

<sup>22</sup> The reason for this division in groups of sets is just for the sake of organization, no relation between the sets and no semantic criterion is applied.

	12	Musical_Instruments
	13	Countries_of_Europe
	14	Conifers
	15	Nationality
	16	Verbs_Animal_Sounds
	10	
group_3	17	Film_Genres
<u> </u>	18	Book_Genres
	19	Famous_Painters
	20	Family_Members
	21	States_Of_Asia
	22	Fashion_Stores
	23	Herbs
	24	Font_Types
group_4	25	Means_Of_Transport
	26	Electronics
	27	Clothes
	28	Zodiac_Signs
	29	Languages
	30	Even_Numbers
	31	Currency
	32	Serials
group_5	33	Music
	34	Interrogative_Pronouns
	35	Parts_of_Speech
	36	Time_Prepositions
	37	Spirits
	38	Parts_of_Skeleton
	39	Location_Prefixes
	40	Verb_Plants
group_6	41	Presidents_of_Czech_Republic
	42	Professions
	43	Rooms_in_the_House
	44	Deciduous_Trees
	45	Fruit
	46	Shapes
	47	Fruit_Trees
	48	Vegetable

After this first phase, the 48 sets underwent a post-selection, supervised by Professor Pavel Rychlý, in which 20 of them were discarded, for two main reasons. The first reason is that some sets may have been questionable or inconsistent, or contained mistakes according to who reviewed them. The second – and primary – reason is that they did not

fulfil the requirements for a semantic category or specific words to be part of the dataset, that is:

- 1. some sets contained Named Entities or proper nouns (e.g., "States of USA")
- some sets did not reflect some general knowledge, but were instead too specific in general (e.g., "Conifers") or too culture-specific (e.g., "Presidents of Czech Republic")<sup>23</sup>
- 3. some sets did not correspond to a semantic category, but instead to a syntactic or morphological criteria were followed in creating them (e.g., "Interrogative\_Pronouns", "Location\_Prefixes")

In the following Table (Table 3) we gather together all the sets that were discarded<sup>24</sup> (first column) and – if possible – we try to trace back to the reason for their exclusion (second column). Unfortunately, we only have the labels of the sets and not their content (the sixteen words, 8 inliers and 8 outliers); therefore, sometimes none of the three requirements listed in the paragraph above can be mentioned as a reason for them to be kept out.<sup>25</sup>

set name	<i>motivation(s) for the exclusion</i>
Colours_II	
Number	not a semantic category
States_of_USA	contains Named Entities
Pronouns	not a semantic category
Countries_of_Europe	contains Named Entities
Conifers	too specific
Nationality	redundant (same as "Languages")
Film_Genres	
Famous_Painters	contains Named Entities
States_Of_Asia	contains Named Entities
Fashion_Stores	contains Named Entities
Font_Types	
Even_Numbers	not a semantic category
Currency	
Serials	
Interrogative_Pronouns	not a semantic category
Time_Prepositions	not a semantic category

Table 3. List of the 20 discarded sets from the original 48 sets dataset, with motivation(s) for the exclusion

<sup>23</sup> As the goal in the creation of the dataset is to capture general-knowledge vocabulary ad semantic categories in a multilingual perspective, culture-specific sets need to be avoided.

<sup>&</sup>lt;sup>24</sup> The order of the items in the Table is not alphabetical: it is the same as in the previous Table.

<sup>&</sup>lt;sup>25</sup> Thus, the corresponding line in the second column is left empty.

Location_Prefixes	not a semantic category
Presidents_of_Czech_Republic	contains Named Entities, too specific (culture-specific)
Deciduous_Trees	too specific

These 28 sets were originally conceived in Czech by the group of students involved in the project and only later translated and adapted to other languages. The first languages to be added were Slovak and English, then French and German came later, always keeping Czech as source language for the translations.<sup>26</sup>

The last<sup>27</sup> contribution to the dataset was given in March 2020 by Ph.D. Kristina Koppel, a Senior Computational Lexicographer at the Institute of Estonian Language (Tallin, Estonia), who collaborates with Lexical Computing. She added 8 new sets in Estonian (which were later adapted to the languages already included) and she also translated the existing sets into Estonian, thus adding a sixth language (after Czech, English, Slovak, German and French). In the following Table (Table 4) we list the 8 sets added by Ph.D. Kristina Koppel.<sup>28</sup>

Table 4. List of the 8 sets added by Ph.D. Kristina Koppel

	set name
1	Birds
2	Bugs
3	Dishes_and_Cutlery
4	External_Body_Parts
5	Internal_Body_Parts
6	Office_Supplies
7	Parts_of_Head
8	Shoes

To sum up, before this thesis' author intervention, the dataset consisted of 37 sets, out of which 28 were originally submitted by the Masaryk University students, 8 were added by Ph.D. Kristina Koppel, and 1 ("Trees") was probably retrieved and readapted from the discarded "Deciduous\_Trees" set. See the following Table (

<sup>&</sup>lt;sup>26</sup> We will address the issues related to the translation and adaptation of the dataset in the methodological Section (Section 4.2).

<sup>&</sup>lt;sup>27</sup> Last before this thesis' author intervention.

<sup>&</sup>lt;sup>28</sup> The list follows an alphabetical order.

## Table 5) for the detailed counts.<sup>29</sup>

Table 5. Numbers of sets in the various phases of the original set construction

status	number of sets
initial sets (Masaryk University students)	48
excluded sets (in post-selection)	20
sets added by Kristina Koppel	8
additional sets	1
overall original dataset	37

As already mentioned before, in this original stage of the dataset, 6 languages were covered: English, German, French, Czech, Slovak, Estonian. The first five were implemented by the Masaryk University students and the latter by Ph.D. Kristina Koppel.

The following Table (Table 6) contains the overall 37 sets listed in alphabetical order. Notice that some labels of the semantic categories have been changed for the sake of uniformity.<sup>30</sup>

Table 6. List of the 37 final sets in the original dataset, before the author's intervention

	set name
1	Birds
2	Book_Genres
3	Bugs
4	Clothes
5	Colours
6	Dishes_and_Cutlery
7	Electronics
8	External_Body_Parts
9	Family_Members
10	Fruit
11	Fruit_Trees
12	Furniture
13	Herbs
14	Human_Features_Positivity
15	Illnesses
16	Internal_Body_Parts
17	Languages

<sup>&</sup>lt;sup>29</sup> The first column contains the status (or phase) of the project; the second column contains the number of sets included in the dataset at that point.

<sup>&</sup>lt;sup>30</sup> E.g., upper/lower case, plural forms etc.

18	Materials
19	Means_of_Transport
20	Music
21	Musical_Instruments
22	Office_Supplies
23	Parts_of_Head
24	Parts_of_Skeleton
25	Parts_of_Speech
26	Professions
27	Rooms_in_the_House
28	School_Subjects
29	Shapes
30	Shoes
31	Spirits
32	Trees
33	Vegetables
34	Verbs_Animal_Sounds
35	Verbs_Plants
36	Verbs_Sport
37	Zodiac_Signs

This dataset was assigned to this thesis's author within the scope of her Erasmus+ Traineeship project at Lexical Computing. The first task (and goal) was to provide a translation of the 37 sets for the Italian language, the author's native language. Although the time required for this was supposed to be little, several problematic issues emerged from the outset while pursuing it. We noticed that the provided translations for English, German and French were not accurate and, in general, there was a lack of uniformity in the dataset, which was probably due to the contribution of several people to the project and no one coordinating or double-checking that the implementations were coherent with each other. Indeed, we realised that:

- there were frequent misspellings (e.g., with French diacritics) which, if left uncorrected, would have compromised the experiment results especially as far as distributional thesauri evaluation is concerned<sup>31</sup>
- 2. not all words were lemmatized correctly (e.g., there were often plural forms with nouns)
- 3. French and German, in particular, were not always aligned with the other languages: not only because there were some mismatches in the translations,<sup>32</sup>

<sup>&</sup>lt;sup>31</sup> Typos or misspelled words may have no occurrence in the reference corpora used for the experiment, while we needed high-frequency words.

 $<sup>^{32}</sup>$  That is, there were – too often – cases in which inliers or outliers in French and German did not correspond at all to the other languages (even keeping into account the fact that there may not be an exact translation for a word). It often happened that there was a completely different word even were a

but also because some inliers were put among the outliers or vice versa, thus nullifying the usefulness of these sets

- 4. there were still some words referring to Named Entities or proper nouns which needed to be substituted
- 5. the criteria for the selection of the outliers and/or their order (as those in Camacho-Collados & Navigli, 2016)<sup>33</sup> were not always respected
- 6. in general, there were incorrect translations (especially of ambiguous words) or translation that were too literal (especially of multi-word expressions)

Therefore, after translating the dataset into Italian, a significant amount of time has been dedicated to solving the problematic issues listed above, with the final goal of reaching an aligned and coherent dataset in all its parts, which could be easily increased with new sets and adapted to other languages.

After fixing the original dataset and implementing the seventh language (Italian), another phase of the project regarded its extension in terms of the number and type of sets included, as well as the distribution of the parts of speech covered. This leads to the methodological Chapter of this thesis (Chapter 4). But before this, in the following Section (Section 3.4) we will anticipate the final outcome of the dataset implementation, by presenting the current status of the dataset, with its 128 sets.

## **3.4 HAMOD dataset: the Outcome**

*HAMOD dataset* refers to the final and latest version of the dataset, that is, the one which was implemented by this thesis' author.

The current dataset consists of 128 sets, out of which 11 are based *topics* and 117 on *semantic categories*. The dataset includes both the original core dataset presented in Section 3.3 (37 sets) and 91 new sets that were added by this thesis' author following the methodology described in Chapter 4. In the dataset, 7 languages (Czech, German, English, Estonian, French, Italian, Slovak) and three parts of speech (nouns, verbs, and

good translation of the word in the other languages (that was the same in the majority of the dataset, i.e., in Czech, English, Estonian and Slovak) was available. If the translation is available and frequent enough in the language, we prefer to keep it, instead of choosing an equivalent (see Section 4.2, for the translation and adaptation of the dataset to other languages).

<sup>&</sup>lt;sup>33</sup> Which we also embrace, see Section 3.2.1.

adjectives) are covered. In the following Table (Table 7) we present some statistics regarding the current status of the dataset: in the second column, the distribution of the parts of speech (nouns, verbs and adjectives) in the original dataset, in the additional sets and overall; in the third column the overall number of sets in the dataset, divided per phase of the project.<sup>34</sup>

status	part	number of sets		
	nouns	verbs	adjectives	
original dataset	31 (0.85)	3 (0.09)	2 (0.06)	37 (1.00)
added sets	54 (0.59)	27 ( <b>0.30</b> )	10 ( <b>0.11</b> )	91 (1.00)
overall	<b>86</b> (0.67)	<b>30</b> (0.23)	<b>12</b> (0.10)	<b>128</b> (1.00)

Table 7. Statistics regarding the number of sets and their part of speech distribution in the original dataset, in the additional sets and overall

What can be spotted here (Table 7) is not only the fact that the overall dataset is more than three times larger, but also that there is a significant variation in the distribution of the parts of speech: verbal and adjectival sets, in particular, have highly increased (as we will discuss widely in Chapter 4, Section 4.1).

Moreover, we calculated the number of words contained in the dataset, which is 2048 per language. It is clear that, as the dataset languages are aligned, each language has the same number of words included in the sets. The following Table (Table 8) shows the distribution of the parts of speech according to the number of words<sup>35</sup> contained in each set, per each phase of the dataset (column 2), and the overall number of words regardless of the part of speech.

status	part	number of words		
	nouns	verbs	adjectives	
original dataset	512	48	32	592
added sets	864	432	160	1456
overall (per language)	1367	480	192	2048
overall (all 7 languages)	9632	3360	1344	14336

Table 8. Statistics on the number of words in the dataset

<sup>&</sup>lt;sup>34</sup> The percentages are in *italics*, in brackets, next to the absolute numbers.

<sup>&</sup>lt;sup>35</sup> These numbers are calculated by multiplying the number of sets by 16 (that is, the number of words contained in each set). We did not calculate the percentages, because they would be exactly the same as in the Table above.

In the previous Table (Table 8), multi-word expressions were counted as one-token words. We believe it is worth discussing here the distribution of the multi-word expressions in the dataset. As mentioned in Sections 3.1 and 3.2, multi-word expressions are allowed, but as shown in Table 9, their impact in the dataset is limited, with a similar distribution among the various languages (< 0.1).<sup>36</sup> In the following Table (Table 9), we outline the distribution of the multi-word expressions among the various languages (column 2), according to the part of speech (line 3, 4 and 5), and calculate the overall number of multi-word expressions in the dataset, per language and in total (column 3).<sup>37</sup>

		multi-word expressions distribution							
language	CS	DE	EN	ЕТ	FR	IT	SK		
nouns	107	7	134	10	111	97	89	555 (0.77)	
verbs	38	9	20	37	11	2	39	156 (0.22)	
adjectives	0	0	0	0	3	4	0	7 (0.01)	
overall	145	16	154	47	125	103	128	718	
percentage	0.07	0.01	0.08	0.02	0.06	0.05	0.06	0.05	

Table 9. Statistics on the number of multi-word expressions in the dataset

Finally, we present here the list of the 128 sets labels (Table 10),<sup>38</sup> with the specification of the part of speech (column 3) and the specification of whether the set is based on a topic or on a semantic category (column 4). The new sets we added are signalled with their set name highlighted in bold, in the Table. We also provide a brief explanation of the content of each set (column 5). The idea of adding descriptions to the set labels comes from the definition of the Semantic Types used in T-PAS (Jezek et al., 2014),<sup>39</sup> which we will discuss extensively in the following Chapter (Chapter 4).<sup>40</sup> Here, the definitions are intended only to facilitate the reader comprehension of the labels used in the dataset.

<sup>&</sup>lt;sup>36</sup> On the issue of multi-word expressions, see Chapter 6, discussion of the results.

<sup>&</sup>lt;sup>37</sup> The percentages are calculated: per each language, over the total number of words per language (2048); overall, over the total number of words in the dataset (14336); per part of speech, over the total number of multi-word expressions in all the languages (718).

<sup>&</sup>lt;sup>38</sup> The sets are listed in alphabetical order.

<sup>&</sup>lt;sup>39</sup> As we will explain in the following Chapter (Chapter 4), Semantic Types are corpus-derived semantic classes (or categories) – see also Section 3.1. Semantic Types turned out to be extremely useful in the dataset implementation as a source of new possible sets. For each Semantic Type in T-PAS, a brief definition of its meaning is provided (they can be consulted at the following link: <u>https://tpas.sketchen-gine.eu/;</u> last access: 24/06/2022).

<sup>&</sup>lt;sup>40</sup> The definitions are freely adapted from Wikipedia pages (<u>https://en.wikipedia.org/wiki/Main\_Page</u>; last access: 24/06/2022) and WordReference (<u>https://www.wordreference.com/</u>; last access:

Finally, some sets are labelled with the same set name followed by "\_1" and "\_2" (see, for example, "Verbs\_Cooking\_1" and "Verbs\_Cooking\_2"): this means that the idea at the basis of the set is the same, but the items instantiating it, both as far as the list of the inliers and the outliers is concerned, differ (see Appendix 1 for the content of the lists). We adopted this strategy when the potential inliers were numerous – enough to form more than one set, and we believed it was worth including them all.

	set name	PoS	type	description
1	Art	noun	topic	any item of any nature that belongs to the
				visual arts field (watercolour, artist etc.)
2	Astronomical_Objects	noun	semantic	natural entities that populate the universe
			category	(stars, planets, etc.)
3	Biomes	noun	semantic	geographical units with distinct climate,
			category	plants and animals (savanna, desert, etc.)
4	Birds	noun	semantic	warm-blooded, egg-laying vertebrate animals
			category	with feathers (penguin, seagull etc.)
5	Bodies_of_Water	noun	semantic	natural entities consisting in accumulations of
			category	water on the surface of a planet (sea, river
				etc.)
6	Book_Genres	noun	semantic	types of literary products, determined by
			category	literary techniques, tone, content or length
				(short story, poem, sci-fi etc.)
7	Bugs	noun	semantic	small invertebrate animals, such as insects
			category	and arachnides (ant, spider etc.)
8	<b>Building_Materials</b>	noun	semantic	physical materials used for buildings
			category	construction, they can be both natural and
				man-made (wood, concrete, glass etc.)
9	Buildings	noun	semantic	man-made structures with a roof and walls,
			category	standing in one place and with specific
				functions (hospital, theatre etc.)
10	Car_Components	noun	semantic	man-made objects which are the constituents
			category	of this motor vehicle (wheel, airbag etc.)
11	Chemical_Elements	noun	semantic	substances consisting of atoms which can be
			category	found on the periodic table (oxygen, sodium,
				helium etc.)
12	Clothes	noun	semantic	man-made items that are worn on the body (t-
			category	shirt, dress etc.)
13	Colours	adj	semantic	properties of object that give it a certain
			category	appearance when light is reflected by it (red,
				blue etc.)

Table 10. List of the 128 sets of the current dataset (original sets + added sets), after the author's intervention

<sup>24/06/2022),</sup> in order to reflect what we intend specifically for each set. We will explain the role of Wikipedia in the dataset construction in the following Chapter (Chapter 4).

1.4	Commutan Commonate			above i and a set of the commuter is the advector
14	Computer_Components	noun	semantic category	physical parts of the computer, i.e., hardware (monitor, mouse etc.)
15	Containers			objects whose aim is to hold, contain, or
15	Containers	noun	semantic	
16	Cashing		category	protect other entities (box, bag etc.)
16	Cooking	noun	topic	any item of any nature that belongs to the
17				cuisine field (garlic, chef etc.)
17	Dairy_Products	noun	semantic	food products made from or containing milk
10	-		category	(butter, cream etc.)
18	Dances	noun	semantic	performances consisting in bodily
			category	movements following a rhythm or music
				(tango, polka etc.)
19	Dimensional_Features_1	adj	semantic	spatial properties of objects in length, width,
			category	thickness etc. (big, tight etc.)
20	Dimensional_Features_2	adj	semantic	spatial properties of objects in length, width,
			category	thickness etc. (big, tight etc.)
21	Dishes_and_Cutlery	noun	semantic	eating utensils including both food containers
			category	and tools used as hand implements to eat
				(fork, knife, bowl etc.)
22	Economics	noun	topic	any item of any nature that belongs to the
				economics field (bank, investment etc.)
23	Electronics	noun	semantic	devices powered by electricity which serve
			category	specific functions (television, mobile phone
				etc.)
24	External_Body_Parts	noun	semantic	any external part of the human body (leg, arm
			category	etc.)
25	Extreme_Natural_Events	noun	semantic	drastic natural events often determined by
			category	climate change that can cause a lot of damage
				to the surrounding environment (tornado,
				flood etc.)
26	Family_Members	noun	semantic	relationships that occur between people in a
			category	family (sister, mother etc.)
27	Fantasy_Characters	noun	semantic	fictional animates which are often magic
			category	creatures (witch, vampire etc.)
28	Farm_Animals	noun	semantic	domesticated animals raised in agricultural
			category	settings (cow, goat etc.)
29	Firearms	noun	semantic	any type of portable gun from which a
			category	projectile is fired by gunpowder (pistol,
				kalashnikov etc.)
30	Fish	noun	semantic	cold-blooeded, acquatic vertebrates having
			category	gills (tuna, shark etc.)
31	Flowers	noun	semantic	blossoms of plants or plants that bear
			category	blossoms (rose, daisy etc.)
32	Flying_Means_of_	noun	semantic	types of vehicles supported for flight in the
	Transport		category	air (helicopter, airplane etc.)
33	Food	noun	semantic	types of sustances eaten for nourishment
			category	(meat, grain etc.)
34	Food_Features	adj	semantic	sensory properties of food that determine
	_	5	category	their perception of smell and taste (salty,
			0- 1	sweet etc.)
L		1	l	

35	Free_Time_Activities	nour	semantic	any kind of amateur activity done for leisure
33	rice_rime_Activities	noun		
26	<b>F</b> . 4		category	(DIY, model building etc.)
36	Fruit	noun	semantic	products of plants that can be eaten by
			category	humans and have a sweet taste (orange, apple
				etc.)
37	Fruit_Trees	noun	semantic	types of trees that carry edible fruits (orange
			category	tree, apple tree etc.)
38	Furniture	noun	semantic	man-made objects that are used as ornament
			category	or with functional use in buildings (table,
				chair etc.)
39	Gemstones	noun	semantic	precious or semiprecious stones that can be
			category	used as jewellery (diamond, emerald etc.)
40	Grain	noun	semantic	edible seeds of cereal plants that provide
			category	carbohydrates to humans (rice, corn etc.)
41	Hair_Features	adj	semantic	properties of hair, referring to their colour,
			category	texture etc. (blonde, curly)
42	Herbs	noun	semantic	edible aromatic plants that are used in
			category	cooking or medicine (rosmary, parsley etc.)
43	Human_Features_	adj	semantic	negative properties of human personality
	Negativity		category	(selfish, dishonest etc.)
44	Human_Features_Positivity	adj	semantic	positive properties of human personality
			category	(kind, nice etc.)
45	Human_Moods	adj	semantic	properties referring to human emotional
			category	states (happy, sad etc.)
46	Human_Physical_Features	adj	semantic	properties referring to human physical shape
			category	(tall, fat etc.)
47	Illnesses	noun	semantic	diseases that affect humans or animals (fever,
			category	flu etc.)
48	Informatics	noun	topic	any item of any nature that belongs to the
			-	informatics field (
49	Internal_Body_Parts	noun	semantic	any internal part of the human body, more
			category	specifically its organs (stomach, bladder etc.)
50	Kitchenware	noun	semantic	tools used for food preparation (strainer,
			category	rolling pin etc.)
51	Landscape_Features	noun	semantic	natural entities which are specific features of
	·······		category	an area of land (hill, mountain etc.)
52	Languages	noun	semantic	human systems of communication identified
	00		category	by the nation(s) in which they are used
			2000 801 9	(English, Italian etc.)
53	Linguistics	noun	topic	any item of any nature that belongs to the
		noun	Pie	linguistics field (language, syllable etc.)
54	Liquid_Containers	noun	semantic	objects whose primary aim is to hold,
57	Liquiu_Containers	noull	category	contain, or protect liquids (glass, bottle etc.)
55	Materials	noun	semantic	natural or artificial substances out of which
55	1414011415	noun		
56	Mothe		category	objects are made (gold, leather etc.)
56	Maths	noun	topic	any item of any nature that belongs to the
57	Moone of Transmit			mathematics field (number, equation etc.)
57	Means_of_Transport	noun	semantic	vehicles primarily used to carry people (car,
			category	bike etc.)

58	Medicine	noun	topic	any item of any nature that belongs to the
50	Meuleme	noun	topic	medical field (pill, x-ray etc.)
59	Metals	noun	semantic	solid, shiny basic materials that can conduct
59	Wietais	noun		electricity (silver, copper etc.)
60	Music	noun	category topic	any item of any nature that belongs to the
00	Music	noun	topic	music field (microphone, sound etc.)
61	Music_Genres	noun	semantic	types of musical styles (jazz, rock etc.)
01	Music_Oemes	noun	category	types of musical styles (jazz, lock etc.)
62	Musical_Instruments	noun	semantic	devices created to make musical sounds
02	indolour_monuments	noun	category	(guitar, flute etc.)
63	Non-alcoholic_Drinks	noun	semantic	drinks that do not contain alcohol, such as
			category	soft drinks (lemonade, tea etc.)
64	Nuts	noun	semantic	dry edible fruit with a hard shell (walnut,
			category	pistachio etc.)
65	Office_Supplies	noun	semantic	equipment that can be found on an office
			category	desk or at school (pen, scissors etc.)
66	Parts_of_Head	noun	semantic	any external part of the human head (eye,
			category	mouth)
67	Parts_of_House	noun	semantic	constitutive parts of a building and its
			category	internal partitions (wall, floor etc.)
68	Parts_of_Skeleton	noun	semantic	any bone of the human skeleton (skull,
			category	coccyx etc.)
69	Parts_of_Speech	noun	semantic	grammatical classes of words (nouns,
			category	interjections etc.)
70	Politics	noun	topic	any item of any nature that belongs to the
				politics field (elections, president etc.)
71	Professions	noun	semantic	occupations requiring specific education or
		-	category	training (firefighter, police officer etc.)
72	Reptiles	noun	semantic	cold-blooded vertebrate animals covered with
			category	dry scales or horny plates (snake, crocodile
72	Dood Moone of			etc.)
73	Road_Means_of_	noun	semantic	types of vehicles that move on the road – or
74	Transport Rooms_in_the_House		category	on terrestrial locations – (car, bus etc.)
74	Rooms_m_me_nouse	noun	semantic category	any partition of a house which serves a specific function (kitchen, bedroom etc.)
75	Savanna_Animals	noun	semantic	various types of wild animals that populate
15	Savanna_Annnais	noun	category	this biome (giraffe, elephant etc.)
76	School_Subjects	noun	semantic	topics which are subject to teaching in classes
70	Sensor_Subjects	noun	category	at school (Maths, Science etc.)
77	Shapes	noun	semantic	geometric bi- or tri-dimensional figures
, ,	~	noun	category	(cube, circle etc.)
78	Shoes	noun	semantic	pieces of garment used to externally cover
-			category	and support the feet (boot, flip-flop etc.)
79	Shops	noun	semantic	places which are usually situated in buildings
	<u>^</u>		category	used for the retail sale of goods and services
				(grocery, bakery etc.)
80	Sources_of_Energy	noun	semantic	types of substances used to provide electricity
			category	and other powering (oil, wind power etc.)
81	Spices	noun	semantic	vegetable substances used for flavoring or
			category	coloring food (pepper, turmeric etc.)
		•	•	

00				
82	Spirits	noun	semantic	alcoholic drinks produced by distillation
	~		category	(beer, whiskey etc.)
83	Sport	noun	topic	any item of any nature that belongs to the
	~			sport field (tennis ball, referee etc.)
84	Sports	noun	semantic	competitive physical activities or gaims
			category	which require various kinds of skills
				(volleyball, tennis etc.)
85	Sweets	noun	semantic	sweet foods made with sugar (cake, candy
			category	etc.)
86	Temperature_Features	adj	semantic	properties of the objects or of the weather
			category	with respect to their temperature (hot, cold
				etc.)
87	Textile_Fibres	noun	semantic	natural or artificial materials used for creating
			category	fabrics and then clothes (wool, cotton etc.)
88	Touch_Features	adj	semantic	properties of objects that can be detected by
			category	touching them (hard, soft etc.)
89	Trees	noun	semantic	plants with woody trunk and branches (oak,
			category	pine etc.)
90	Units_of_Time	noun	semantic	intervals used to measure duration in time
			category	(year, minute etc.)
91	Vegetables	noun	semantic	plants or parts of plants (such as roots) that
	C .		category	are edible (potato, cabbage etc.)
92	Verbs_Animal_Sounds	verb	semantic	verbs that refer to the sounds typically made
			category	by animals (bark, meow etc.)
93	Verbs_Cognition	verb	semantic	verbs that refer to actions involving human
	_ 8		category	mind (understand, forget etc.)
94	Verbs_Communication_1	verb	semantic	verbs that refer to actions involving the
			category	human ability to communicate verbally (say,
			0,	repeat etc.)
95	Verbs_Communication_2	verb	semantic	verbs that refer to actions involving the
			category	human ability to communicate verbally (say,
			0,	repeat etc.)
96	Verbs_Cooking_1	verb	semantic	verbs that refer to actions done while
	· · · · · _ · · · · · · · · · · · · · ·		category	preparing and cooking food (roast, fry etc.)
97	Verbs_Cooking_2	verb	semantic	verbs that refer to actions done while
			category	preparing and cooking food (roast, fry etc.)
98	Verbs_Crime	verb	semantic	verbs that refer to human fraudulent or
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,			category	criminal actions against other humans
				(kidnap, threaten etc.)
99	Verbs_Destroy	verb	semantic	verbs that refer to the actions performed in
	· · · · · · · · · · · · · · · · · · ·	, 010	category	order to destroy something (demolish, break
			cutegory	etc.)
100	Verbs_Dog	verb	semantic	verbs that refer to actions typically made by
100	· · · · · · · · · · · · · · · · · · ·	,010	category	dogs (bark, growl etc.)
101	Verbs_Driving	verb	semantic	verbs that refer to actions that can be done by
101	verus_Driving	verb		humans while driving a vehicle (brake, park
			category	etc.)
102	Vorba Foting	uanh	comontia	verbs that refer to actions that humans
102	Verbs_Eating	verb	semantic	
			category	perform while eating (chew, swallow etc.)

103Verbs_Economicsverbsemanticverbs that refer to the economics category104Verbs_Farmingverbsemanticverbs that refer to the agriculture gardening field (cultivate, fertiliss105Verbs_Hairverbsemanticverbs that refer to actions that car category106Verbs_Human_Soundsverbsemanticverbs that refer to the sounds typi category107Verbs_Killingverbsemanticverbs that refer to violent actions results into somebody being kille somebody else (kill, drown etc.)108Verbs_Measureverbsemanticverbs that refer to changes in spat of physical objects (lengthen, red category109Verbs_Motionverbsemanticverbs that refer to actions that car category110Verbs_Mouthverbsemanticverbs that refer to changes in spat category110Verbs_Mouthverbsemanticverbs that refer to actions that car category111Verbs_Perceptionverbsemanticverbs that refer to the music field category113Verbs_Plantsverbsemanticverbs that refer to life or seasonal category114Varbs_Plantsverbsemanticverbs that refer to actions that car category113Verbs_Plantsverbsemanticverbs that refer to life or seasonal category114Varbs_Plantsverbsemanticverbs that refer to life or seasonal category114Varbs_Plantsverbsemanticverbs that ref	or e etc.) be , curl etc.)
104Verbs_Farmingverbsemantic categoryverbs that refer to the agriculture gardening field (cultivate, fertilise category105Verbs_Hairverbsemantic categoryverbs that refer to actions that car performed on humans hair (comb106Verbs_Human_Soundsverbsemantic categoryverbs that refer to the sounds typi category107Verbs_Killingverbsemantic categoryverbs that refer to violent actions results into somebody being kille 	e etc.) be , curl etc.)
Interpretationcategorygardening field (cultivate, fertilise105Verbs_Hairverbsemantic categoryverbs that refer to actions that car performed on humans hair (comb106Verbs_Human_Soundsverbsemantic categoryverbs that refer to the sounds typi by humans (laugh, sing etc.)107Verbs_Killingverbsemantic categoryverbs that refer to violent actions results into somebody being kille 	e etc.) be , curl etc.)
105Verbs_Hairverbsemantic categoryverbs that refer to actions that car performed on humans hair (comb106Verbs_Human_Soundsverbsemantic categoryverbs that refer to the sounds typi category107Verbs_Killingverbsemantic categoryverbs that refer to violent actions results into somebody being kille 	be , curl etc.)
Image: Normal state in the image	, curl etc.)
106Verbs_Human_Soundsverbsemantic categoryverbs that refer to the sounds typi by humans (laugh, sing etc.)107Verbs_Killingverbsemantic categoryverbs that refer to violent actions 	
107Verbs_Killingverbsemantic semantic categoryverbs that refer to violent actions results into somebody being kille somebody else (kill, drown etc.)108Verbs_Measureverbsemantic categoryverbs that refer to changes in spat of physical objects (lengthen, red of physical objects (lengthen, red109Verbs_Motionverbsemantic categoryverbs that refer to changes in spat of physical objects (lengthen, red space (go, return etc.)110Verbs_Mouthverbsemantic categoryverbs that refer to actions that car performed using the mouth (kiss,111Verbs_Musicverbsemantic categoryverbs that refer to actions that car performed using the mouth (kiss,112Verbs_Perceptionverbsemantic categoryverbs that refer to actions that car performed using the five senses ( etc.)113Verbs_Plantsverbsemantic categoryverbs that refer to life or seasonal plants (bloom, sprout etc.)	cally made
107Verbs_Killingverbsemantic categoryverbs that refer to violent actions results into somebody being kille somebody else (kill, drown etc.)108Verbs_Measureverbsemantic categoryverbs that refer to changes in spat category109Verbs_Motionverbsemantic categoryverbs that refer to changes in spat category110Verbs_Motionverbsemantic categoryverbs that refer to changes in dire category110Verbs_Mouthverbsemantic categoryverbs that refer to actions that car category111Verbs_Musicverbsemantic categoryverbs that refer to the music field category112Verbs_Perceptionverbsemantic categoryverbs that refer to actions that car category113Verbs_Plantsverbsemantic categoryverbs that refer to life or seasonal plants (bloom, sprout etc.)	carry made
Image: Constraint of the second sec	
108Verbs_Measureverbsemantic categoryverbs that refer to changes in spat category109Verbs_Motionverbsemantic categoryverbs that refer to changes in dire space (go, return etc.)110Verbs_Mouthverbsemantic categoryverbs that refer to actions that car performed using the mouth (kiss, sing etc.)111Verbs_Musicverbsemantic categoryverbs that refer to actions that car performed using the mouth (kiss, sing etc.)112Verbs_Perceptionverbsemantic categoryverbs that refer to actions that car performed using the five senses ( etc.)113Verbs_Plantsverbsemantic categoryverbs that refer to life or seasonal plants (bloom, sprout etc.)	
108Verbs_Measureverbsemantic categoryverbs that refer to changes in spat of physical objects (lengthen, red109Verbs_Motionverbsemantic categoryverbs that refer to changes in dire category110Verbs_Mouthverbsemantic categoryverbs that refer to changes in dire space (go, return etc.)111Verbs_Mouthverbsemantic categoryverbs that refer to actions that car performed using the mouth (kiss,111Verbs_Musicverbsemantic categoryverbs that refer to the music field sing etc.)112Verbs_Perceptionverbsemantic categoryverbs that refer to actions that car performed using the five senses ( etc.)113Verbs_Plantsverbsemantic categoryverbs that refer to life or seasonal plants (bloom, sprout etc.)	l by
Image: 109Verbs_Motionverbsemantic categoryof physical objects (lengthen, red verbs that refer to changes in direct space (go, return etc.)110Verbs_Mouthverbsemantic categoryverbs that refer to changes in direct space (go, return etc.)110Verbs_Mouthverbsemantic categoryverbs that refer to actions that car performed using the mouth (kiss, sing etc.)111Verbs_Musicverbsemantic categoryverbs that refer to the music field sing etc.)112Verbs_Perceptionverbsemantic categoryverbs that refer to actions that car performed using the five senses ( etc.)113Verbs_Plantsverbsemantic categoryverbs that refer to life or seasonal plants (bloom, sprout etc.)	
109Verbs_Motionverbsemantic categoryverbs that refer to changes in dire space (go, return etc.)110Verbs_Mouthverbsemantic categoryverbs that refer to actions that car performed using the mouth (kiss, sing etc.)111Verbs_Musicverbsemantic categoryverbs that refer to the music field category112Verbs_Perceptionverbsemantic categoryverbs that refer to actions that car performed using the five senses ( etc.)113Verbs_Plantsverbsemantic categoryverbs that refer to life or seasonal plants (bloom, sprout etc.)	-
110       Verbs_Mouth       verb       semantic       verbs that refer to actions that car         111       Verbs_Music       verb       semantic       verbs that refer to the music field         111       Verbs_Perception       verb       semantic       verbs that refer to actions that car         112       Verbs_Perception       verb       semantic       verbs that refer to actions that car         113       Verbs_Plants       verb       semantic       verbs that refer to life or seasonal         113       Verbs_Plants       verb       semantic       verbs that refer to life or seasonal	
110       Verbs_Mouth       verb       semantic category       verbs that refer to actions that car performed using the mouth (kiss, performed using the mouth (kiss, verbs that refer to the music field sing etc.)         111       Verbs_Music       verb       semantic category       verbs that refer to the music field sing etc.)         112       Verbs_Perception       verb       semantic category       verbs that refer to actions that car performed using the five senses (performed using the five	ction in the
111       Verbs_Music       verb       semantic       verbs that refer to the music field         111       Verbs_Perception       verb       semantic       verbs that refer to actions that car         112       Verbs_Perception       verb       semantic       verbs that refer to actions that car         113       Verbs_Plants       verb       semantic       verbs that refer to life or seasonal         113       Verbs_Plants       verb       semantic       verbs that refer to life or seasonal	
111       Verbs_Music       verb       semantic category       verbs that refer to the music field sing etc.)         112       Verbs_Perception       verb       semantic category       verbs that refer to actions that can performed using the five senses ( etc.)         113       Verbs_Plants       verb       semantic category       verbs that refer to life or seasonal category	
112       Verbs_Perception       verb       semantic category       verbs that refer to actions that car performed using the five senses (retc.)         113       Verbs_Plants       verb       semantic category       verbs that refer to life or seasonal category	-
112       Verbs_Perception       verb       semantic category       verbs that refer to actions that can performed using the five senses ( etc.)         113       Verbs_Plants       verb       semantic category       verbs that refer to life or seasonal plants (bloom, sprout etc.)	(record,
113     Verbs_Plants     verb     semantic category     performed using the five senses (retc.)       113     Verbs_Plants     verb     semantic category     plants (bloom, sprout etc.)	
113     Verbs_Plants     verb     semantic     verbs that refer to life or seasonal category       plants     verb     semantic     verbs that refer to life or seasonal category	be
113Verbs_Plantsverbsemantic categoryverbs that refer to life or seasonal plants (bloom, sprout etc.)	see, listen
category plants (bloom, sprout etc.)	
	phases of
114 Vanha Davah	
114Verbs_Psychverbsemanticverbs that refer to changes in hun	an
category emotional states (frighten, bore et	c.)
115 Verbs_Religion verb semantic verbs that refer to the religious field	eld (baptize,
category marry etc.)	
116 Verbs_School verb semantic verbs that refer to the actions that	students
category perform at school (learn, memori	ze etc.)
117 Verbs_Smell verb semantic verbs that refer to the smell sense	(sniff,
category perfume etc.)	
118         Verbs_Sport         verb         semantic         verbs that refer to actions perform	ned while
category doing sports (skate, ski etc.)	
119 Verbs_Telephone verb semantic verbs that refer to actions that car	be done
category while using a phone (call, text etc	.)
120 Verbs_Touch verb semantic verbs that refer to actions that inv	olve the
category touch sense (caress, grasp)	
121 Verbs_Weather verb semantic verbs that refer to the manifestation	ons of the
category states of the atmosphere (rain, the	nder etc.)
122 War noun topic any item of any nature that belong	gs to the war
field (battle, soldier etc.)	
123 Water_Means_of_ noun semantic types of vehicles that move on the	e water
Transportcategory(boat, ferry etc.)	
124 Weapons noun semantic man-made devices specifically but	:14 f
category attack or defense in fighting and	int for
sword etc.)	
125 Weather_Conditions adj semantic properties of the states of the atm	
category (rainy, foggy etc.)	wars (bomb,

126	Weather_Events	noun	semantic	states of the atmospheres determined by air
			category	pressure and temperature (rain, fog etc.)
127	Wild_Animals	noun	semantic	undomesticated animals that can be found in
			category	different natural ecosystems (deer, bear etc.)
128	Zodiac_Signs	noun	semantic	astrological signs that correspond to the
			category	twelve constellations of the ecliptic (taurus,
				gemini etc.)

The previous Table (Table 10) only presents the labels of the sets. The whole dataset, with the 16 items (inliers and outliers) per each set (per each language) can be examined in Appendix 1. Also, the dataset will be available under a permissible Creative Commons licence in a public repository, where the dataset will be stored in a machine-readable format.<sup>41</sup>

In the following Chapter (Chapter 4), we will discuss the methodology and the steps undertaken in order to build the dataset presented in this Section.

<sup>&</sup>lt;sup>41</sup> That is, each set is a folder containing 7 distinct .txt files, one for each language, with the 16 words in column, with the inliers separated from the outliers by an empty line (see Section 3.2.2). This kind of format is suitable for the scripts used in the experiment on the distributional thesauri.

# Chapter 4. Methodology: How to Build and Implement HAMOD

This is the core methodological Chapter of the thesis. In the following Sections we discuss the methodology we conceived and applied in order to increase and improve HAMOD dataset. We organized this Chapter in steps, which not only reflect the methodological steps undertaken, but also the chronology of our work in the dataset implementation. At each step, we discuss the methodological matters, and we recall and examine specific portions of the dataset which we shortly presented in Section 3.4.<sup>1</sup>

Here follows a brief outline of the Chapter.

In Section 4.1 we address the first step of the dataset implementation, that is, increasing the number of its sets and improving the distribution of the parts of speech contained in it.<sup>2</sup> Therefore, we specify the various sources (which were also useful in combination) we consulted and took inspiration from in order to add new sets. We also claim which sets come from which source(s), and eventually, how they were adjusted according to our dataset peculiarities.

In Section 4.2, we address some issues concerning the translation/adaptation of the dataset from the source language (English) to the other languages that are included in the project. We mention the sources we used in order to translate the new added sets, as well as the people – who are native speakers of some of the languages included – who helped this thesis' author in the translation/adaptation of the dataset.

In Section 4.3 we discuss the third step, that is, to verify if the words we chose as candidates for the sets (both inliers and outliers) actually belong to a 12-years-old person vocabulary knowledge. In order to do so, we had the chance to pre-test part of the dataset with a small group of primary school students; the results of this experiment helped us refining the dataset further, by substituting those words which were not understood by the group.

<sup>&</sup>lt;sup>1</sup> We will only discuss the 91 sets added by this thesis' author; the other 37, as we described in Chapter 3, were already defined and we only refined them as far as the translations and the formal layout is concerned.

 $<sup>^{2}</sup>$  As we pointed out in Section 3.3 and 3.4, the original dataset mainly contained sets for nouns; verbs and adjectives were underrepresented.

Finally, in Section 4.4 we present the guidelines<sup>3</sup> that summarise the criteria followed in the construction of the dataset, which have been a work in progress throughout the dataset construction and implementation. Included in the guidelines there is also a Section concerning the translation and adaptation of the dataset to other languages. These guidelines will be useful for whoever will be interested in further developing the outlier detection task or perform it on other languages that have not been included yet.

# 4.1 First Step: How to Create New Sets

After fixing the original dataset,<sup>4</sup> the first goal was to extend the number of sets in the dataset. In this Section we present the criteria we established in order to select new semantic categories and topics, and, in particular, we mention the sources which we took inspiration from in this phase. The idea of adding new sets came from the awareness that a wider range of vocabulary needed to be covered (37 sets covered 592 words)<sup>5</sup> and that the distribution of parts of speech was unbalanced (85% of the sets contained nouns).<sup>6</sup>

Concerning the second point, we can suppose that the reason for this is that nouns (which typically refer to entities) are easier to be grouped in semantic categories than verbs or adjectives, as well as to be recognized by humans as semantic categories. We will provide some insights regarding this assumption in the presentation and discussion of the results of the experiment on human evaluators in Chapter 6.

Also, in the creation of new sets, we found simpler to conceive semantic categories for nouns rather than for verbs and adjectives; this is the reason why nouns are still a preponderant portion in the new sets (59% of the sets).<sup>7</sup> However, we tried to increase the number of verb and adjective sets to a considerable extent, thus reaching a higher number for

<sup>&</sup>lt;sup>3</sup> As we will discuss later, the guidelines are adapted from those in Camacho-Collados & Navigli (2016), which we expanded and modified according to our project peculiarities.

<sup>&</sup>lt;sup>4</sup> See Section 3.3.

<sup>&</sup>lt;sup>5</sup> See Section 3.4.

<sup>&</sup>lt;sup>6</sup> See Section 3.4.

<sup>&</sup>lt;sup>7</sup> See Section 3.4.

both (30 overall sets for verbs and 12 overall sets for adjectives).<sup>8</sup> This is important as previous studies based on the outlier detection task did not include other parts of speech beyond nouns,<sup>9</sup> and we are interested in analysing the results of the experiments (Chapter 6) also as far as verb and adjective sets are concerned, comparing them with those of the noun sets.

Dataset enlargement through the addition of 91 new sets has been carried out manually by this thesis' author and occupied an important part of her Traineeship experience and of her thesis project in general. Finding and conceiving new sets was not an easy task, neither finding adequate inliers and outliers. Although some of them came to this thesis' author mind from her world knowledge, in order to reach a substantial number, we needed to take inspiration from other sources.

With these premises, we now introduce the sources we consulted for the new sets and in which terms they were useful for the dataset implementation.

### 4.1.1 T-PAS ontology and the Semantic Types

The first source we used was T-PAS ontology (also known as *System of Semantic Types*; Jezek, 2019), a hierarchy of corpus-derived semantic categories (called Semantic Types).

T-PAS (Jezek et al., 2014) is a corpus-derived resource consisting of an inventory of Typed Predicate-Argument Structures (T-PAS) for Italian verbs. The resource is being developed at the University of Pavia (Department of Humanities)<sup>10</sup> with the technical support of Lexical Computing and is intended to be used for linguistic analysis, language teaching, and computational applications. *Typed predicate-argument structures* are patterns which display the syntactic and semantic properties of verbs and their meanings. These patterns, derived from a corpus of Italian language,<sup>11</sup> are acquired through manual

<sup>&</sup>lt;sup>8</sup> See Section 3.4.

<sup>&</sup>lt;sup>9</sup> See Section 2.3.2.

<sup>&</sup>lt;sup>10</sup> This thesis' author has been involved for the last three years in T-PAS project, during which she has had the opportunity to develop skills in linguistic data annotation, semantic classification, and corpus analysis, as well as a deep knowledge of Sketch Engine and its tools, thanks to which she could apply for a Traineeship at Lexical Computing. From here the idea of using T-PAS Semantic Types in the dataset implementation, thus connecting two distinct projects which can mutually benefit from this merging.

<sup>&</sup>lt;sup>11</sup> The reference corpus for the resource is the web corpus ItWac (reduced), provided by Sketch Engine. It contains around 935 million tokens.

clustering and annotation of corpus instances, following the CPA methodology (Hanks, 2013).<sup>12</sup>

In T-PAS, a verb sense is determined by the arguments it combines with (subject, object, etc.) and their respective Semantic Types. Semantic Types are general semantic categories used to specify the semantics of arguments and they are obtained by manual clustering of the lexical items found in the argument positions with respect to the verbs in the corpus. They are organized in the System of Semantic types, which currently contains ca. 200 Semantic Types that are hierarchically structured on the basis of the "is a" (subsumption) relation (e.g., [Human] is an [Animate]). The System of Semantic Types is grounded in a long-standing project, which started from the Brandeis Semantic Ontology (Rumshisky et al. 2006; Pustejovsky et al., 2006), was first adapted to a sister resource for English predicate-argument structures – PDEV (Hanks, 2013)<sup>13</sup> – and has been absorbed and re-adapted again within T-PAS project, according to the specificities of the predicate-argument structures of Italian verbs (Jezek et al., 2014; Jezek, 2019).

In the following Figure (Figure 1) we can see some examples of Semantic Types in the resource, with their structure and definitions.<sup>14</sup>

<sup>&</sup>lt;sup>12</sup> Corpus Pattern Analysis (CPA) is a procedure in corpus linguistics which associates word meaning with word use by means of analysis of phraseological patterns and collocations (Hanks, 2004).

<sup>&</sup>lt;sup>13</sup> PDEV public interface: <u>https://pdev.org.uk/</u> (last access: 24/06/2022).

<sup>&</sup>lt;sup>14</sup> The Figure is taken and adapted from T-PAS web interface (<u>https://tpas.sketchengine.eu/;</u> last access: 24/06/2022). The System of Semantic Types can be explored at the same link.
System	of Semantic Types
Filter	
Anything -	
ST to use as	a last resort when [Eventuality], [Entity] and [Property] are equally likely
ENT	тү –
	ything] that exists independently of other things and has a distinct identity. [Anything] which is not an [Eventuality] nor a [Property] za, materiale, ambiente, economia, creatura, edificio
	Physical Entity +
	A tangible [Entity]
	ponte, faccia, tavolo, auto, fiore, uccello birra, merci, pietra, bambino, vulcano
	Abstract Entity +
	An intangible [Entity], such as a [Concept]
	idea, problema, concetto
	Energy +
	[Entity] produced by burning [Fuel] - thus becoming [Light] or [Heat] - or by movement - becoming [Sound]
	luce, energia, forza, calore
	Particle
	Microscopic [Entity]s, such as the subatomic parts of an atom
	elettrone, protone, fotone
EVE	NTUALITY -
lt c	an either be an [Event] involving movement, change or development or a fixed [State]
eve	ento, relazione, cambiamento, situazione
	Event +
	An [Eventuality] that involves movement, change, or development, unlike a [State]. An [Event] can either be a volitional [Activity] or a non-volitional [Process]
	incontro, morte, visita, matrimonio, trattamento, tempesta, guerra, richiesta
	State +
	A static [Eventuality] that does not involve activity, movement, or development
	pace, stabilità, situazione, equilibrio
Pro	PERTY +
Aq	uality or characteristic of [Anything]
pes	so, altezza, bellezza, forma, eleganza, reputazione

Figure 1. Screenshot of the System of Semantic Types in T-PAS resource

The System of Semantic Types resembles an ontology, in the sense that it is organizes the knowledge of the world in semantic categories (or classes) – from a bottom-up perspective, that is, using corpora in order to retrieve the categories. Thus, its aim it is not to be exhaustive in terms of types of entities, eventualities and properties classified in the hierarchy; instead, it reflects what is found in the corpora.

What is interesting is that these semantic categories are compatible with the definition of sets we have in this project, and therefore the Semantic Types turned out to be a valuable source for new sets. Moreover, some of the original 37 sets already shared some categories with the System, such as "Musical Instruments", "Fruit", "Means of Transport".

We used some of T-PAS Semantic Types as labels for our sets and their definitions in order to select the inliers included. Here follows a Table (Table 1) in which we present the list of the Semantic Types as candidates for the creation of new sets. In column 2 we report the name of the Semantic Type in T-PAS, in column 3 the adapted name in HAMOD.

	name of the Semantic Type in T-PAS	set name in HAMOD
1	Astronomical Object	Astronomical_Objects
2	Body of Water	Bodies_of_Water
3	Building	Buildings
4	Container	Containers
5	Fantasy Character	Fantasy_Characters
6	Firearm	Firearms
7	Fish	Fish
8	Flower	Flowers
9	Flying Vehicle	Flying_Means_of_Transport
10	Food	Food
11	Natural Landscape Features	Landscape_Features
12	Road Vehicle	Road_Means_of_Transport
13	Water Vehicle	Water_Means_of_Transport
14	Weapon	Weapons
15	Weather Event	Weather_Events

Table 1. Semantic Types from T-PAS used as labels in HAMOD dataset

The 15 sets we derived from T-PAS cover only noun as part of speech and mainly refer to specific kinds of physical entities. Looking at the hierarchy, we did not include general semantic categories (such as [Artifact] or [Abstract Entity]), but instead really specific ones, which can be found at deeper level in the hierarchy. This can be motivated by the fact that for the task of the outlier detection to be effective, we need to have a limited set of items to include, which can be easily identified as a cluster against the word that does not belong.

Later, after selecting the Semantic Types and corresponding labels in the dataset, we listed the 8 inliers + 8 outliers per each set which reflected the definition of the semantic category. The inliers were chosen both thanks to this thesis' author vocabulary and knowledge of the world, but also thanks to Wikipedia Lists, as we will address in the following Section (Section 4.1.2). As for the outliers, we selected them following the principles outlined in Section 3.2.1 regarding the formal layout of the sets in the dataset.

### 4.1.2 Wikipedia lists and categories

Wikipedia, the well-known multilingual free online encyclopedia, organizes its contents (that is, its pages) in categories.

> The central goal of the category system is to provide navigational links to Wikipedia pages in a hierarchy of *categories* which readers, knowing essential –

*defining* – characteristics of a topic, can browse and quickly find sets of pages on topics that are defined by those characteristics.<sup>15</sup>

Another means of organizing Wikipedia knowledge is through lists, which are often combined with categories and share similar principles of organization. Categories and lists are normally found at the bottom of an article page, but there are some general pages that collect lists. Above all, there is the "List of lists of lists" page, which gathers all the lists contained in Wikipedia, or, in their words, «an article that is a list of articles that are themselves lists of article lists».<sup>16</sup> In this page, the lists are themselves grouped by domain or field (such as Biology, Linguistics, Economy, Medicine, Art etc.) and each domain contains delimited lists of items which may direct to other lists or Wikipedia articles. See, as an example, the following Figure (Figure 2), in which there is a list of lists regarding the Biology domain. This structure clearly resembles a taxonomy or ontology, based on domains, and we used it as a source of new semantic categories or topics for the sets.

<sup>&</sup>lt;sup>15</sup> https://en.wikipedia.org/wiki/Wikipedia:Categorization (last access: 24/06/2022).

<sup>&</sup>lt;sup>16</sup> <u>https://en.wikipedia.org/wiki/List\_of\_lists\_of\_lists</u> (last access: 24/06/2022).



Figure 2. Screenshot from the "List of lists of lists" page in Wikipedia

Also, categories in Wikipedia were extremely useful as for the selection of the inliers, providing lists of pages, in alphabetical order, which were good instances of inliers within the semantic categories or the topics. See, as an example, the following Figure (Figure 3), with a list of pages related to the category "Kitchenware".<sup>17</sup> Most of the items contained in it (*can opener, egg timer, bowl, mixer*) can be viewed as "types of" Kitchenware, thus being hyponyms that can clearly form a semantic category taken together.

<sup>&</sup>lt;sup>17</sup> That is, tools, utensils, appliances, dishes, and cookware used in food preparation, or the serving of food. <u>https://en.wikipedia.org/wiki/Category:Kitchenware</u> (last access: 24/06/2022).

The following 53 pages ar	e in this category, out of 53 total. This list may not reflect recent
changes (learn more).	
12thala ann ann	
<ul> <li>Kitchenware</li> </ul>	
A	
Anti-griddle	
В	
<ul> <li>Beverage opener</li> </ul>	
Bowl	
Bread warmer	
c	
Can opener	
Carafe	
<ul> <li>Cherry pitter</li> </ul>	
<ul> <li>Combination plate</li> </ul>	
CorningWare	
D	
Daunglan	
<ul> <li>Decanter</li> </ul>	
Dipper well	
E	
Egg timer	

Figure 3. Screenshot from "Category:Kitchenware" page in Wikipedia

Another way to access this kind of information is to visit the relevant Wikipedia page, in which there may be a paragraph named "Types of…" and a list of even more items than those in the category page, which can increase the number of inliers in the sets. See, as an example, the following Figure (Figure 4), from "Kitchenware" page.<sup>18</sup>

<sup>&</sup>lt;sup>18</sup> <u>https://en.wikipedia.org/wiki/Kitchenware</u> (last access: 24/06/2022).



Figure 4. Screenshot of a paragraph from "Kitchenware" page in Wikipedia

As we only needed 8 inliers, the procedure we followed in selecting them within the list or categories was to go through them and choose those which we believed were not too domain or culture specific, thus belonging to some rather basic vocabulary.<sup>19</sup>

To sum up, Wikipedia has been a valuable source both for new sets labels and for instances of inliers in the dataset. In the following Table (Table 2) we present the Wikipedia pages or categories or lists (column 2) and the corresponding sets in the dataset (column 3). We also specify (column 4) which kind of contribution these pages gave (only the idea for the semantic category or also the inliers).<sup>20</sup>

	Wikipedia pages	set name in HAMOD	contribution
1	Lists of astronomical objects	Astronomical_Objects	inliers
2	Biome	Biomes	semantic category, inliers
3	Category:Bodies of water	Bodies_of_Water	inliers
4	Category:Building materials	Building_Materials	semantic category, inliers
5	Category:Auto parts	Car_Components	semantic category, inliers
6	Category:Chemical elements	Chemical_Elements	inliers
7	Computer hardware	Computer_Components	semantic category, inliers

Table 2. Wikipedia pages, categories or lists used in HAMOD dataset

<sup>19</sup> See the following Section (Section 4.3) for a discussion on what "basic vocabulary" means and how this can be tested.

<sup>20</sup> In this Table, we did not include a link for each page but using the words in column 2 as queries in the English Wikipedia will provide the corresponding page we consulted.

8	Category: Containers	Containers	inliers
9	Category: Dairy products	Dairy_Products	semantic category, inliers
10	List of dances	Dances	inliers
11	Category:Natural disasters	Extreme_Natural_Events	semantic category, inliers
12	Lists of legendary creatures	Fantasy_Characters	inliers
13	Category:Livestock	Farm_Animals	inliers
14	Category:Firearms	Firearms	inliers
15	Category:Fish common names	Fish	inliers
16	Category:Aircrafts	Flying_Means_of_Transport	inliers
17	Category:Hobbies	Free_Time_Activities	semantic category, inliers
18	Category:Gemstones	Gemstones	semantic category, inliers
19	Grain	Grain	semantic category, inliers
20	Category:Kitchenware	Kitchenware	semantic category, inliers
21	Category:Metal	Metals	semantic category, inliers
22	Category:Music genres	Music_Genres	inliers
23	Category:Soft drinks	NonAlcoholic_Drinks	inliers
24	Nut (fruit)	Nuts	semantic category, inliers
25	Portal:Reptiles	Reptiles	semantic category, inliers
26	Category:Land transport	Road_Means_of_Transport	inliers
27	Category:Confectionery	Sweets	semantic category, inliers
28	Textile	Textile_Fibres	semantic category, inliers
29	Category:Water transport	Water_Means_of_Transport	inliers
30	Weapon	Weapons	inliers

### 4.1.3 Verb sets: using Levin (1993) semantic classes for verb sets

Verb sets posed a different range of issues with respect to nouns sets. The original dataset only contained two verb sets, namely "Verbs\_Sport" and "Verbs\_Ani-mal\_Sounds". As we mentioned at the beginning of this Section, we wanted to increase the number of verb sets to study the semantic categorization of verbs and the human perception of verb clusters in more detail. To do so, we used Levin's *English Verb Classes and Alternations* (Levin, 1993) as a source, even though some of the 27 sets have been conceived independently from this work.<sup>21</sup>

In her work, Levin classifies over 3,000 English verbs according to shared meaning and behavior. She shows how identifying verbs with similar syntactic behavior provides an effective means of distinguishing semantically coherent verb classes. What turned out

<sup>&</sup>lt;sup>21</sup> These sets are sets which resemble more "topics", as they collect a domain-specific range of verbs. These sets are: "Verbs\_Crime", "Verbs\_Dog", "Verbs\_Economics", "Verbs\_Farming", "Verbs\_Music", "Verbs\_Religion", "Verbs\_School", "Verbs\_Telephone". Though, what the inliers of these sets share is not only the fact of being domain-specific, but also some argumental properties. For example, the verbs contained in "Verbs\_Dog" all select in the subject position a dog performing the actions expressed by the verbs. This issue will not be investigated herein, thus it remains an interesting point to develop in future perspectives.

to be useful was the second part of the book, in which the author lists these semantic classes of verbs, including in each relevant verbs instantiating the class, illustrative examples, comments on noteworthy properties, and bibliographic references. We used both the names of the classes and the verbs instantiating them as sources of new verb sets in the dataset. In the following Table (Table 3) we display the verb classes in Levin (1993) (column 2) and the corresponding set in HAMOD dataset (column 3), and some examples of verbs contained in both (column 4). The examples are selected from those which we included in the dataset from Levin's examples. In some cases, the same verb class in Levin (e.g., "Verbs of Ingesting") was a source of inliers for different sets (e.g., "Verbs\_Eating" and "Verbs\_Driving").

	Levin's semantic class	set name in HAMOD	verb instances
1	Declare Verbs (Levin, 1993:	Verbs_Cognition	think, believe, suppose, know,
	182); Conjecture Verbs		judge
	(Levin, 1993: 183)		
2	Verbs of Communication	Verbs_Communication_1 <sup>22</sup>	ask, narrate, tell, announce,
	(Levin, 1993: 202)		declare, claim
3	Verbs of Communication	Verbs_Communication_2	ask, narrate, tell, announce,
	(Levin, 1993: 202)		declare, claim
4	Cooking Verbs (Levin, 1993:	Verbs_Cooking_1	bake, broil, fry, roast, steam,
	243)		stew
5	Cooking Verbs (Levin, 1993:	Verbs_Cooking_2	bake, broil, fry, roast, steam,
	243)		stew
6	Destroy Verbs (Levin, 1993:	Verbs_Destroy	destroy, devastate,
	239)		exterminate, ruin, wreck
7	Verbs of Motion Using a	Verbs_Driving	drive, park, bend, steer
	Vehicle (Levin, 1993: 267)		
8	Verbs of Ingesting (Levin,	Verbs_Eating	eat, chew, crunch, sip,
	1993: 213)		swallow, ingest
9	Verbs of Caring for a Specific	Verbs_Hair	braid, comb, wave, curl, shave
	Body Part (Levin, 1993: 228)		
10	Verbs Involving the Body	Verbs_Human_Sounds	grumble, hiss, murmur,
	(Levin, 1993: 217), Verbs of		whisper, whistle, hiccup
	Manner of Speaking (Levin,		
	1993: 205)		
11	Verbs of Killing (Levin, 1993:	Verbs_Killing	assassinate, poison, murder,
	230)		kill, strangle
12	Measure Verbs (Levin, 1993:	Verbs_Measure	enlarge, increase, widen
	272)		

Table 3. Levin's (1993) verb classes used in HAMOD dataset

<sup>&</sup>lt;sup>22</sup> For the reason why we use "\_1" and "\_2" in some sets, see Section 3.4. As for "Verbs\_Communication\_1" – "Verbs\_Communication\_2" and "Verbs\_Cooking\_1" – "Verbs\_Cooking\_2" we used the same source in Levin (1993), which included several instances of inliers, enough to form more than one set.

13	Verbs of Motion (Levin, 1993:	Verbs_Motion	leave, enter, go, exit, arrive
	263)		
14	Verbs of Ingesting (Levin,	Verbs_Mouth	yawn, cough, spit, kiss, lick
	1993: 213); Verbs Involving		
	the Body (Levin, 1993: 217)		
15	Verbs of Perception (Levin,	Verbs_Perception	feel, hear, notice, see, smell,
	1993: 185)		taste
16	Psych-Verbs (Levin, 1993:	Verbs_Psych	annoy, terrify, depress, bother,
	188)		encourage
17	Verbs of Smell Emission	Verbs_Smell	exhale, inhale
	(Levin, 1993: 236); Exhale		
	Verbs (Levin, 1993: 218)		
18	Verbs of Exterting Force:	Verbs_Touch	caress, pull, press, grasp,
	Push/Pull Verbs (Levin, 1993:		touch
	138); Verbs of Contact: Touch		
	Verbs (Levin, 1993: 155)		
19	Weather Verbs (Levin, 1993:	Verbs_Weather	drizzle, rain, snow, hail,
	276)		thunder

# 4.2 Second Step: How to Translate and Adapt the Sets into other Languages

HAMOD dataset has been *multilingual* since its very beginning in 2019.<sup>23</sup> It currently covers 7 languages, which we recall here: Czech, German, English, Estonian, French, Italian and Slovak. One of the issues with the original portion of the dataset is that there was not a single reference-source language for the translation of the others, thus creating some inconsistencies and mismatches among the various languages. As we have already mentioned,<sup>24</sup> the original core was in Czech, then 8 sets were added in Estonian, and French and German translations were not good enough. In order to avoid further problems, we decided to align all the other languages to English, thus English becoming the source language for the translations.

Therefore, 91 new sets which we added to the dataset were originally conceived in English and simultaneously translated and adapted into Italian, this thesis' author native

<sup>&</sup>lt;sup>23</sup> See Section 3.1.

<sup>&</sup>lt;sup>24</sup> See Section 3.3.

language. Before explaining how we proceeded with the five remaining languages, we briefly focus on what we mean by *adaptation* and why it is not only about *translation*.

By *adaptation* we mean not to choose a straightforward translation, but instead to adapt the word into something similar that still fits among the inliers or the outliers. This can happen in the following cases we detected:

- 1. there is no exact correspondence in the translation
- 2. the corresponding translation is infrequent (according to a reference corpus) in the target language(s)
- 3. the corresponding translation is too polysemous and/or ambiguous
- 4. the word is too culture-specific (e.g., names of food, means of transports, animals) and therefore absent in the new language(s)

Before choosing to adapt, in these circumstances, a multiword expression can be used if it is attested in the corpus - i.e., it is not Out-Of-Vocabulary (OOV). For example, *pet* in English would be translated *animale\_domestico* in Italian.

In case this is not possible, the target word can be replaced with a completely different one - but still semantically related according to the guidelines for the selection of the inliers and the outliers. For example, as *custard* in English is culture-specific, it can be adapted to *tvaroh*<sup>25</sup> in Czech (thus always referring to a dairy product).

These adaptations make the dataset not strictly a parallel one but a comparable one (Jakubíček et al., 2021). In the following Table (Table 4) we report some more examples of these cases, with the languages in parallel (column 2 to 8). Multi-word expressions are encoded with underscore among each token (e.g., *tap\_dance*).

<sup>&</sup>lt;sup>25</sup> *Tvaroh* is a typical Czech cheese.

set name	EN	CS	DE	ET	FR	IT	SK
Dairy	custard	syrovátka	Quark	hapukoor	crème_	ricotta	srvátka
Products					fraîche		
Dances	tap_dance	břišní_	Polka	rahvatants	claquettes	tip_tap	brušný_
		tanec					tanec
Fish	fish_pie	rybí_	Fischstäbchen	suitsukala	bouillabaisse	fritto_	rybačka
		prsty				misto	
Non	smoothie	sodovka	Smoothie	smuuti	smoothie	aranciata	odvar
alcoholic							
Drinks							
Rooms	utility_	dětský_	Kinderzimmer	lastetuba	buanderie	ripostiglio	detská_
in the	room	pokoj					izba
House							
Shoes	high_	lodičky	Pumps	ketsid	talons_hauts	scarpe_	lodičky
	heels					col_tacco	
Sweets	cheesecake	bábovka	Käsekuchen	juustukook	dragée	confetto	koláč
Verbs	photosynthesize	vadnout	vertrocknen	fotosünteesima	flétrir	appassire	rozkvitnúť
Plants							

While most of the examples in the previous Table are cases of culture-specific items (especially as far as food is concerned: "Dairy\_Products", "Fish", "Non-alcoholic Drinks", "Sweets"), some other are cases in which the corresponding translation was not frequent enough within the reference corpus in Sketch Engine ("Rooms in the House", "Verbs Plants").

Henceforth we describe how we approached to the translation/adaptation to the other languages, which sources we used and who was involved. Taking English as a source language, we proceeded as follows.

First, we roughly translated all the sets through machine translation,<sup>26</sup> quickly checking major mistakes or typos (that is, if the algorithm had skipped a word or kept the English term). Then, we submitted this machine-translated version of the words (always grouped by sets and always accompanied by their labels, as well as the part of speech of each set) to a small group of collaborators. This group had to browse the lists and check and verify whether the automatic translation was suitable, or, in case not, change the word to a better translation or even adapt the word (following the criteria outlined above). We also asked them to pay attention to these potential problems and address them as follows:

<sup>&</sup>lt;sup>26</sup> We used DeepL Translator (<u>https://www.deepl.com/translator</u>; last access: 24/06/2022), which we deem more accurate than Google Translate, even though it covers much less languages (but this did not affect our translations, because all our 7 languages are included in the software).

- the translated word has to be commonly used in the language (in particular for what concerns multiword expressions) – in case of uncertainty, we asked them to check a corpus in the target language and choose accordingly
- 2. the translated word has to be in the same part of speech of the other words in the set,<sup>27</sup> this being English the source language, significantly affected by part of speech ambiguities
- 3. the translated word has to be encoded as its lemma (e.g., no plural forms are accepted unless that is the only form for that word)<sup>28</sup>
- 4. in case of semantic polysemy/ambiguity, the translated word has to be contextually consistent with the others in the set. For example, *oil* in the set "Cooking" must refer to the ingredient used to cook or fry (thus, *olio* in Italian, *huile* in French etc.). If it is in the set "Sources of Energy", then it refers to *petroleum* used to power engines (thus, *petrolio* in Italian, *pétrole* in French etc.)
- 5. In case of part of speech ambiguity, the translated word has to be consistent with the others in the set. For example, *fast* in the set "Verbs Eating" cannot refer to its homonymous adjective meaning "quick, rapid", but instead it would mean "to eat no food" (and thus translated as *paastuma* in Estonian, *fasten* in German etc.)

With these brief indications,<sup>29</sup> we translated the dataset. As for Czech language, three members of Lexical Computing helped us: Jan Kraus (Language Analyst for Sketch Engine), Vlasta Ohlídalová (Technical Support for Sketch Engine, who is also a Master Student in Computational Linguistics at Masaryk University) and Ota Mikušek (Software Developer Junior, who is also a Bachelor Student in Computer Science). As for Slovak, Michaela Denisová (a Ph.D. student in Computer Science at Masaryk University) gave her contribution. Estonian was translated by Kristina Koppel, who we have already mentioned in Section 3.3. Finally, as far as German and French are concerned, we collaborated with this thesis' author sister, Sofia Romani, who is a high school student and masters German, French and English.

Then, after the collaborators' contribution, we post-processed the dataset by manually checking the changes the collaborators had done with respect to the machine-translated version (which we asked them to keep track of). In order to do so, we used some online dictionaries for the various languages, namely:

<sup>&</sup>lt;sup>27</sup> As we have discussed in Section 3.2, there cannot be mixed sets with both adjectives, verbs, and nouns: we only have only-verbs, only-nouns and only-adjectives sets.

<sup>&</sup>lt;sup>28</sup> E.g., *trousers* in English.

<sup>&</sup>lt;sup>29</sup> Which are also part of the official Guidelines, see Section 4.4 and Appendix 2.

- 1. For Estonian: Estonian English Dictionary (<u>https://aare.pri.ee/diction-ary.html?switch=en</u>); Lingea Multilingual Online Dictionary (<u>https://dict.com/</u>)
- 2. For Czech: Czech English Dictionary (<u>https://slovnik.seznam.cz/preklad/an-glicky</u>); Lingea Multilingual Online Dictionary (<u>https://dict.com/</u>)
- 3. For German: English German WordReference (<u>https://www.wordrefer-ence.com/ende/</u>); Lingea Multilingual Online Dictionary (<u>https://dict.com/</u>)
- 4. For French: English French WordReference (<u>https://www.wordrefer-ence.com/enfr/</u>); Lingea Multilingual Online Dictionary (<u>https://dict.com/</u>)
- 5. For Slovak: Lingea Multilingual Online Dictionary (https://dict.com/)

Finally, as for the multi-word expressions which were not encoded within those online dictionaries, we used Wikipedia and the versions of the same pages in the other languages (if present), to double-check whether the expression actually existed. Furthermore, we checked each the frequency of each multi-word expression using the Concordance tool<sup>30</sup> in Sketch Engine and we established as a minimum threshold around 1000 occurrences of the term within the reference corpus for that language. If these conditions were not satisfied, we changed the word and adapt it to another similar one. This because we did not want to include infrequent words or terms in the dataset, which could compromise the results of the distributional thesauri evaluation.

## **4.3** Third Step: Do the Words Belong to a Basic Vocabulary?

The third goal was to verify whether this kind of sets and the kind of words they contain fulfil one of the requirements of the dataset, that is, the fact that the lexicon should be easily understood by a 12-years-old person.<sup>31</sup> Therefore, we present a small pilot experiment we conducted on a group of primary school students in order to assess this. With the outcome of the experiment, we could improve the dataset and further simplify the difficulty of its words, aiming at covering some basic vocabulary of the languages included.

<sup>&</sup>lt;sup>30</sup> Sketch Engine Concordance: <u>https://www.sketchengine.eu/guide/concordance-a-tool-to-search-a-corpus/</u> (last access: 24/06/2022).

<sup>&</sup>lt;sup>31</sup> See Section 3.2.

The experiment was conducted during period of the traineeship in Brno. Thanks to Ota Mikušek, we could involve a group of 7 Czech students who attend a pioneer group,<sup>32</sup> of which Ota is one of the leaders. These students' ages range from 10 to 14 years old, thus being good candidates for checking whether the words we included within the dataset were part of their vocabulary knowledge. As the student's native language was Czech, we could only evaluate the Czech part of the dataset. However, we assumed that the words that the students did not understand would have been hard even for same-age native speakers of the other languages, and thus, after the pilot experiment results, we fine-tuned the dataset crosswise.

Being the students underage, we asked their parent's authorisation for them to take part of the experiment and for this thesis' author to attend the meeting in which the experiment was held. We informed them that he data we collected would have been treated anonymously and all the parents accepted.<sup>33</sup>

In collaboration with Ota Mikušek, we set the experiment as follows. As the main goal of the experiment was not to test whether the students were able to perform the outlier detection task with success, but instead to test their knowledge of the word, we conceived a series of exercises and linguistic games that could help us detect which words they recognize and which they did not. We wanted to avoid asking them directly which words they knew and which they did not, so we managed to achieve this by testing their language knowledge indirectly. We built these four types of exercises:

- 1. one-letter short anagrams
- 2. syllable long anagrams
- 3. fill in the missing letter
- 4. 1 out of 5 (that is, a reduced version of the outlier detection task itself)

What follows here are some examples of the exercises, in Czech:

## One-letter short anagrams

auřzb ávak aohbn aksop nvápe ícelž mérxi akáv

<sup>&</sup>lt;sup>32</sup> Pioneers are associations similar to Scouts.

<sup>&</sup>lt;sup>33</sup> The relation with the parents has been managed by Ota Mikušek and supervised by Miloš Jakubíček via email.

Syllable long anagrams

na li zm rz sc og e én ra fi če do ru ní ab ba ob um av ka zk nc at ta ov ut ěž so ga te gi by č čí ta po ov ač nk to pi dl zr ca o ed t mě př na li zm rz

### Fill in the missing letter

```
ná _ j

_ ušek

m_š

ba _ nek

e nerg _ svě _ o

pap _ rá_o

chuť k _ d lu

k _ eč

ma _ na

ná _ j

_ ušek
```

1 out of 5

kovový skleněný dřevěný těžký látkový malý papírový kožený hliníkový zlatý motorka loď chodník auto tramvaj autobus vlak letadlo vrtulník cesta nota skladatel obraz kytara rock flétna zvuk mikrofon zpěvák socha These exercises were automatically generated and randomized through some scripts starting from the dataset by Ota Mikušek. They were printed on paper and given to the students during one of the pioneers meetings, compiled manually by them and recollected for the analysis. Ota Mikušek who was also fundamental as an intermediary between these thesis' author and the students during the experiment, as well as for the analysis of the results of the experiment.

In the following Table (Table 5) we report the quantitative results of the experiment, that is, the number of correct answers (and thus, of words understood), per children (line 2 to 8) and per exercise (column 2 to 5). The overall results (column 6 and line 9) seem to be promising, especially as far as the 1 out of 5 (which is the reduced version of outlier detection).

	one letter	syllable	fill in the	1 out of 5	overall
	anagrams	anagrams	missing letter		performance
student 1 (10 yo)	0.75	0.60	0.85	0.76	0.74
student 2 (10 yo)	1.00	0.93	0.90	0.76	0.90
student 3 (11 yo)	1.00	0.70	0.95	1.00	0.91
student 4 (12 yo)	1.00	0.95	0.85	0.80	0.90
student 5 (13 yo)	0.60	0.73	0.73	0.88	0.74
student 6 (14 yo)	1.00	0.95	0.95	0.92	0.96
student 7 (14 yo)	1.00	0.85	0.90	0.52	0.82
overall	0.90	0.81	0.88	0.80	0.85

Table 5. Quantitative results of the experiment with Czech students, in percentage

We can also argue, with reference to the Table above, that age is not really a relevant factor in the performance, even though we would have expected a better performance (thus, a higher number of words guessed) among the oldest students. Other factors may have influenced the results, one of which may have been the motivation for completing the exercises and how much attention they have given to them.

We now report the specific words we changed according to the pilot experiment results. We simplified the words in all 7 languages, choosing a different word that could still fit among the inliers or the outliers.<sup>34</sup> In the following Table (Table 6), we report the original word which was guessed wrongly by one or more student(s) (column 2) and how we simplified it in English (column 3).<sup>35</sup>

<sup>&</sup>lt;sup>34</sup> How simple these substitutions actually are still relies on our judgments. A new test could have been done in order to assess if these changes could have led to an improvement, but this was not possible due to the fact that the traineeship was about to end.

<sup>&</sup>lt;sup>35</sup> See the dataset in Appendix 1 for the other languages.

set name	original word	new word
Astronomical_Objects	nebula	satellite
Astronomical_Objects	space_craft	astronaut
Biomes	shrubland	prairie
Biomes	flatland	shrub
Buildings	sport_centre	library
Buildings	old_town	avenue
Computer_Components	software	app
Computer_Components	update	program
Containers	casket	chest of drawers
Containers	ceramics	straw
Cooking	dining_room	breakfast
Dances	yoga	step
Dances	artistic_gymnastic	musical
Dances	stage	dancer
Fantasy Characters	evil_eye	magic
Firearms	withdrawal	murderer
Fish	anchovy	cod
Food	diary_product	fish
Informatics	update	app
Kitchenware	food_processor	cutting_board
Landscape_Features	dune	glacier
Landscape_Features	lagoon	plain
Liquid_Containers	jerry_can	tin
Politics	propaganda	poll
Politics	accountant	boss
Road_Means_of_Transport	campervan	taxi
Savanna_Animals	baobab	poacher
Savanna_Animals	cactus	safari
Sources_of_Energy	biomass	metan
Sport	shin_guard	ball
Sport	offside	match
Sport	modelling	board_game
Textile Fibres	polyester	linen
Textile Fibres	weaving	sewing_machine
Verbs_Cognition	doubt	believe
War	recruit	battle
Weapons	withdrawal	gunshot
Weather_Conditions	changeable	muggy
Weather_Events	haze	cloud

Table 6. Qualitative results of the experiment and words that have been changed (inliers and outliers)

The final outcome of HAMOD dataset, which we introduced in Section 3.4 and can be consulted in its entirety in Appendix 1, is what follows this last step of refinement.

## **4.4** Fourth Step: How to Further Implement the Dataset, that is, the Guidelines

Some Guidelines for the outlier detection dataset building were missing.<sup>36</sup> We believe that, in case of future developments of this project that would require the dataset implementation, clear guidelines are fundamental for new contributors to be able to create new sets that are consistent with those already included in the dataset. Writing these guidelines has been a work in progress throughout the dataset construction and we believe that if the experiments using the dataset led to poor or unwanted results (e.g., high agreement between human evaluators is not maintained) the dataset and therefore its guidelines could be modified progressively and accordingly.

In this Section, we discuss the content of these guidelines, recalling the principles and the criteria for the creation of the sets already mentioned in the previous Section (Section 4.1). Before this, we present Camacho-Collados & Navigli (2016) *Guidelines for Clusters Creation*,<sup>37</sup> which is the basis for our guidelines and which we adapted to our specific requirements. In addition, we point out some problematic issues that emerge from their guidelines, that we tried to solve in our version. Finally, as our dataset is – and is conceived as – multilingual, a consistent part of the guidelines is dedicated to the translation and adaptation of the dataset to other languages.

## 4.2.1 Guidelines for Clusters Creation

In the first part of their guidelines, Camacho-Collados & Navigli (2016) provide some concise indications for the creation of the clusters (which we call *sets*).<sup>38</sup>

First, the authors define a cluster as «a group of 8 words/concepts which are semantically very similar and are all connected by a clear well-known relation», providing as a cluster example theMonths of the Year" (January, February, March etc.). Not only the example is not appropriate for our dataset (we do not want to include proper nouns or named entities),<sup>39</sup> but also the authors do not clarify neither what they mean by

<sup>&</sup>lt;sup>36</sup> People who worked in this project before were not given written guidelines, but only oral indications on how to proceed progressively.

<sup>&</sup>lt;sup>37</sup> The guidelines are not contained in their paper (Camacho-Collados & Navigli, 2016), but they can be downloaded from the following link: <u>http://lcl.uniroma1.it/outlier-detection/</u> (last access: 24/06/2022).

<sup>&</sup>lt;sup>38</sup> All the quotes in quotation marks contained in this Section are accurately taken from their guidelines.

<sup>&</sup>lt;sup>39</sup> See Section 3.2.

*semantically very similar* nor by *clear well-known relation*. As we have already addressed in Chapter 1, the notion of semantic similarity is problematic, and it is not clear which kind of (semantic) relation the authors intend, as well as who should know (well) this relation. In our guidelines, we try to explain what we mean by semantic categories (that are the basis for our clusters-sets) and to define the relation between the items contained in them.

Then, they suggest to «select topics as narrow and diverse as possible», mentioning as an example "Countries in North America" (as a *narrow* topic) vs. "Countries" in general (as a *broad* topic). In our dataset we avoid narrow or domain-specific semantic categories, as we aim at collecting sets based on some general knowledge.<sup>40</sup>

After explaining how to select the inliers (that is, how to form the cluster of 8 items), a second section in their guidelines is related to how to choose the outliers. While choosing elements belonging to a semantic category (cluster) seems an easy task, selecting the outliers may be trickier and clearer guidelines are needed. We agree with the authors in that the outliers «are not supposed to be part of the cluster [i.e., of the inliers]», and we also follow (as explained in Section 3.2) their partition of the outliers in four sub-groups (or sub-sets)<sup>41</sup> – as well as their order, although the criteria for this partition are not specified, neither exemplified. What follows (Table 7) is how the four sub-groups (column 1) are defined in their guidelines (column 2).

sub-group	definition		
C1	Two very similar elements to the ones in the cluster (Important to be certain that they are		
	NOT as similar as the elements within the cluster and that they cannot form a separate		
	cluster with seven elements of the main cluster).		
C2	Two similar and related elements (lower degree of similarity and relatedness in		
	comparison with the first category) to the ones in the cluster.		
C3	Two related but not similar elements to the ones in the cluster.		
C4	Two unrelated and not similar elements to the ones in the cluster.		

Table 7. Outliers partition in Camacho-Collados & Navigli (2016) in four sub-groups, with their definition

We can notice that, again, some important concepts are neither defined nor exemplified: not only it is unclear what they mean by *similarity* and *relatedness*, but also which is the difference between the two. We did not use these concepts in our guidelines, but

<sup>&</sup>lt;sup>40</sup> See Section 3.2.

<sup>&</sup>lt;sup>41</sup> The authors use the word "category", which we do not report here because it could be misleading, as it may recall "semantic category".

we widely discussed these issues in Chapter 1.<sup>42</sup> Instead, we try to better define the criteria for the selection of the outliers, and we provide practical examples.

Finally, Camacho-Collados & Navigli (2016) guidelines define what the files format should be and how content of the sets should be organized. In our version of the guidelines, we also specify some formal requirements that are needed in the encoding of the dataset, in order to make it usable in the experiment.

In the following Section (Section 4.2.2) we briefly present the structure of our project guidelines and summarize its content. The full version of the guidelines can be found in Appendix 2, and it is also available in a public repository together with the dataset.

## 4.2.2 Guidelines for HAMOD dataset Creation and Translation

In the first part of the guidelines, we focus on the dataset creation. We define the structure of the sets and the specific restrictions we established for the selection of the semantic categories, that is:

- 1. no named entities or proper names should be included
- 2. semantic categories and the items they include should belong to some general knowledge
- 3. words should belong to a 12-year-old person vocabulary
- 4. the only criterion for the identification of the sets should be semantic

Then, we better define what we mean by semantic category and topic, providing details and examples on how the items among the inliers can be selected. After, we specify better criteria for the selection of the outliers, following Camacho-Collados & Navigli (2016) structure. Finally, we add some further restrictions for what concerns the encoding of the words in the dataset, that is:

- 1. all the words in a set should have the same part of speech
- 2. multi-word expressions should be encoded with an underscore between each token<sup>43</sup>
- 3. all the words should be encoded in their lemma form

and we report how the file format should be in order to be used both for the human and the distributional thesauri evaluation.

The second part of the guidelines – that is, the one regarding the translation and the adaptation of the dataset – is a peculiarity of our project, as compared to Camacho-

<sup>&</sup>lt;sup>42</sup> See Section 1.3.2.

<sup>&</sup>lt;sup>43</sup> In order to be machine-readable.

Collados & Navigli (2016). Translation has been a big part of the dataset creation and we want to avoid problems caused by bad translations from the source dataset when leading the experiments. Hereafter, what future contributors should pay attention to:

- 1. literal translations should be avoided
- 2. an eye should be kept on the frequency of the words in the reference corpora<sup>44</sup>
- 3. the part of speech of a corresponding word should be kept in the translation, and should be reflected among the other items in the set<sup>45</sup>
- 4. the word should be encoded as its lemma (except for *pluralia tantum*, that is, words with typically in their plural form)
- 5. semantic ambiguity and part of speech ambiguity should be addressed in case of uncertain translation

As mentioned at the beginning of this Section, the *Guidelines for HAMOD dataset Creation and Translation* can be found in their integral version in Appendix 2.

<sup>&</sup>lt;sup>44</sup> Low-frequency words may affect the performance of the distributional thesauri or word embeddings when running the outlier detection task.

<sup>&</sup>lt;sup>45</sup> That is, no mixed-part-of-speech sets are allowed.

## Part 3. An Experiment with HAMOD Dataset

## Chapter 5. Evaluation of Distributional Thesauri and Word Embeddings through HAMOD Dataset

This is the experimental part of this thesis project: the set-up of an experiment using HAMOD dataset (Chapter 5) and its results (Chapter 6). In this Chapter we describe how we arranged the outlier detection task as an experiment to evaluate distributional models (namely, Sketch Engine Thesaurus and word embeddings), by comparing their performance to the human evaluation on the same task.

Here follows a brief outline of the Chapter.

In Section 5.1 we present the steps followed in the experiment in general, and some hypotheses on the expected results, both as far as the human evaluation and the distributional models' evaluation is concerned.

In Section 5.2 we focus on the human evaluation, which is preliminary to the models' evaluation, as we discuss therein. We describe how we organized the experiment in a controlled environment, how the participants were selected, and which metrics are used to assess their performance.

In Section 5.3 we focus on the distributional models' evaluation. We first describe the corpora on which the thesaurus and the word embeddings were computed; then, we recall the metrics used to assess the models' performance; finally, we briefly mention the computational procedure used in order to perform the evaluation, as well as the arrangement of the output results.

## 5.1 Experiment Setup and Hypotheses on the Results of the Experiment

### 5.1.1 The experiment setup: steps and requirements

As we have discussed in Section 2.1, an evaluation procedure requires rigorous steps to be effective, reliable, and reproducible.

Given the task we selected (the *outlier detection task*) and the dataset build in order to perform this specific task as an intrinsic method to evaluate distributional models (*HAMOD dataset*), the first step in the evaluation procedure is to assess the quality of the dataset for that specific task. This is done by means of human evaluation. In our specific case, the human evaluation consists in involving a group of evaluators who are asked to perform the outlier detection task itself (i.e., given a set of 8 words, select the word that does not semantically belong to that set).

As we mentioned in Section 2.1.1, for the human evaluation to be reliable, there are some requirements:

- 1. the task needs to be carried out by *multiple* participants (the more they are, the more reliable the human evaluation is)
- 2. participants should be selected from a *well-defined population*
- 3. the participants need to work *independently* on the same task (they cannot influence each other's results)
- 4. the participants need to follow some shared clear *indications* in order to conduct the task, which they have to receive and read before the experiment takes place
- 5. the participants cannot *reiterate* the task more than once, as further attempts may affect the results (the participants may have already seen some parts of the task in previous runs)

To these general requirements, we add another one related to our specific task:

6. participants need to perform the task in their *native language*(s); as we assume that the knowledge of a second language may not be excellent, decisions in the task may be motivated by different reasons from native speakers and the task could be too misleading

In Section 5.2 we will discuss these points, reporting them to our experiment. Compared to the other experiments we reviewed in Section 2.3.2, regarding which we pointed out some weaknesses in the procedure they followed for the human evaluation, we believe that our experiment on the human evaluation is much more reliable, for the following reasons:

1. we include a consistent<sup>1</sup> number of participants, which were not previously involved in the dataset construction<sup>2</sup>

<sup>&</sup>lt;sup>1</sup> Even if the desideratum is much higher (see Jakubíček et al., 2021), for reasons of time and within the scope of this thesis, we collected a restricted number of human evaluators. A higher number of evaluators will be collected in the future developments of this project.

<sup>&</sup>lt;sup>2</sup> This happened in Camacho-Collados & Navigli (2016) and in Gamallo (2018).

- 2. these participants are selected from a well-defined population (namely, they are Linguistics students)
- the participants conduct the experiment independently, in a controlled environment
- 4. the participants are given clear guidelines before the experiment performance
- 5. the participants did not reiterate the  $task^3$
- 6. finally, we evaluated the whole dataset, which is, to our knowledge, the biggest being fully evaluated by humans for this specific task<sup>4</sup>

Then, once the human evaluation task has taken place, we proceed with the second step: we calculate the agreement between the participants and, if it results in a high percentage (in our case, our threshold is < 90%), we can consider the dataset as a benchmark, or a gold standard, for the models' evaluation. In case the human evaluation led to poor results in terms of agreement, the dataset should be modified accordingly, and, consequently, the experiment reiterated involving new participants.

If the dataset results into a gold standard for the experiment, the third step can take place, that is, the evaluation of the models (in our case, the distributional models – *distributional thesauri* and *word embeddings*) by exploiting standard evaluation metrics. In our case, we also introduce a custom-made refined metrics which allows us to gain insights into the performances of the models. The only requirement for evaluation of distributional thesauri and word embeddings we have is that, in order to make the two models more comparable, they have to be computed on the same corpora, for all the languages involved.

As a final step, these results can be compared to the human's ones. In our case, this is possible because both the human evaluators engaged in the experiment and the models perform exactly the same task on the same dataset (the whole dataset); thus, their results are comparable. We undertake a quantitative and a qualitative analysis of the results, also in comparison, in Chapter 6.

## 5.1.2 Hypotheses on the results of the experiment

We briefly mention here some hypotheses regarding the experiment and the results we expect.

First, as far as the human evaluation is concerned, we expect a high agreement between the human evaluators (i.e., < 90%). This because it has already been proved in previous experiments. Beside Camacho-Collados & Navigli (2016) who obtained a strikingly high

<sup>&</sup>lt;sup>3</sup> This happened in Camacho-Collados & Navigli (2016).

<sup>&</sup>lt;sup>4</sup> Blair et al. (2017) and Andersen et al. (2020) only evaluated part of their datasets through human evaluation.

agreement,<sup>5</sup> some tests have been carried out on the original version of HAMOD dataset (namely, the one containing 37 sets). Czech and Estonian versions of the original dataset have been evaluated by human annotators, as reported in Rychlý (2019) and Jakubíček et al. (2021), resulting into 97% of raw agreement for Czech and 93% for Estonian.

While we expect such percentages in our experiment, we have to keep in mind that the current dataset is ca. 3.5 times bigger than the original dataset (37 vs. 128 sets). The size could affect the results for the fact that 128 sets require more time to be evaluated by participants. Furthermore, the evaluators could lose attention or interest in such a task – which may result repetitive to some of them. These are all variables we need to consider in the discussion of the human evaluation results.

Second, as far as the distributional models' evaluation is concerned, we expect some differences among the two kinds of models. We hypothesise that word embeddings, that is, a predictive kind of model, outperform the distributional thesaurus, a rather count-based distributional model. Actually, previous tests using the original dataset (Jakubíček et al., 2021) showed rather that none of the models outperform the other one. Keeping in mind that the models were computed on different corpora and with a different size of the dataset, we will compare those results and the new one we calculate on HAMOD dataset in Section 6.2.

Moreover, we expect different performances in the various languages, and that these may depend on the size of the corpus used to compute the models: the bigger the corpus is, the better the quality of the models – and therefore their performance in the task – is. However, we cannot foresee which language will do best in the outlier detection task; nevertheless, we expect similar results in genetically closer languages (e.g., Czech and Slovak, English and German, French and Italian).

Finally, both as far as the human evaluation and the models' evaluation are concerned, we expect some variation in the results according to the position of the outlier. To recall this point, we have explained in Section 3.2 that the outliers are ranked according to the degree of relatedness or similarity to the inliers (i.e., the items belonging to a semantic category or topic). The higher the outlier is in the ranking, the closer in meaning it is to the inliers, and vice versa. Given this, we suppose that the closer the outlier is to the inliers, the more challenging its detection among the inliers is: thus, we expect better results in the detection of the items which are farther from the inliers, and worse results in the detection of the items which are closer.

Furthermore, we suppose that sets based on adjectives and verbs can be more challenging than those based on nouns, both for the humans – as clusters of verbs and adjectives could be less intuitive – and for the models: we expect more mistakes in adjective and verb sets.

<sup>&</sup>lt;sup>5</sup> But we need to remember that the amount of sets evaluated and the number of participants involved and the way they were involved is not really reliable; see Section 2.3.1.

In Section 6.6 we will go back to these hypotheses and verify whether they can be confirmed after the analysis of the results.

## 5.2 Human Evaluation

## 5.2.1 The online interface for the experiment

For human evaluators to perform the outlier detection task, a web, user-friendly interface has been built.<sup>6</sup> It is developed by Lexical Computing sro. and it was already available in other tests on the outlier detection (as we described in Section 5.1.2).

We now describe the web interface, starting from its main page. In the main page (Figure 1), the evaluator is asked whether they wants to start a new test or continue an existing one.



Figure 1. First page of the web interface for the outlier detection task

After clicking on "start a new exercise", the evaluator is asked to type a username and to select the language in which they want to perform the task (Figure 2). As we previously stated, we encourage to undertake the task in the native language. Nevertheless, we are aware that in uncontrolled experimental settings it is currently not possible to verify

<sup>&</sup>lt;sup>6</sup> Currently, it is visible at <u>https://milos.sketchengine.co.uk/outlier\_detection/#</u> (last access: 24/06/2022).

whether the evaluator is performing the task in its native language, especially as far as English is concerned.<sup>7</sup> Currently, 7 languages are available, as can be seen from the Figure.

<b>9</b>	our nickname	
- s	elect vour language Choose your language	
ls your if you \ words.	Czech	
	English	
	Estonian	
	French	© 2020
	German	ct has received funding from the European
	Italian	s Horizon 2020 research and innovation nme under grant agreement No 731015.
	Slovak	

Figure 2. Second page of the web interface, username and language selection

After typing the username and selecting the language, a page of indications for performing the task is shown (Figure 3). There is an exercise ID (a unique alphanumeric code) which can be saved in case one wants to interrupt the evaluation test and continue it (see Figure 1). We will discuss the precise indications concerning the task performance in the following Section (Section 5.2.2).

Here we mention that evaluators are told the motivation for this experiment, that is, the evaluation of automatic thesauri aimed at their improvement in terms of methods for their automatic generation.

Furthermore, it is important to underline here that evaluators are informed on the privacy policy, that is, how their data will be treated after the experiment and the fact that

<sup>&</sup>lt;sup>7</sup> This may be the case if the experiment was conducted through crowdsourcing – that is, involving large amounts of web users as evaluators through some crowdsourcing platforms. In future perspectives, if we want to enlarge the number of evaluators, we need to address this issue and find a way to verify the evaluator native language before the task performance.

the statistics from their results can be aggregated and anonymously treated for the calculation of the agreement, as we will do in Chapter 6.

Welcome emma
You are about to start a new exercise. Your exercise ID is
1e7c598e43c81331e93538b75c8a9cef
and you can return to this exercise later through this link or from the main page by entering your exercise ID.
Your task
<ul> <li>You will be presented 9 words.</li> <li>Your task is to choose an <i>outlier</i>. An outlier is a word that does not fit based on its meaning or sense.</li> <li>Examples: blue, red, green, yellow, orange, black, brown, white, table. Obviously, the last word is the outlier: all the others are names of colours. bricklayer, lawyer, shop assistant, gentleman, waitress, metheorologist. Gentleman is an outlier because it is not a job.</li> </ul>
Why should I do it?
<ul> <li>Your answers will be used for the evaluation of automatically built thesauri in order to improve the methods for their automatic generation.</li> <li>Examples of distributional thesauri: a word-sketch based one and a word-embeddings based one.</li> <li>The key reason for taking the task manually is to detect issues in the dataset which will manifest themselves by low inter-annotator agreement.</li> </ul>
How long does it take?
About 10 minutes.
What happens with the data?
<ul> <li>The resulting dataset will be openly available under the CC-BY-SA 4.0 licence in the ELEXIS GitHub repository.</li> <li>We will also publish anonymized statistics on the inter-annotator agreement evaluation.</li> <li>By taking the exercise you agree that your answers can be stored and used to improve the outlier detection datasets.</li> </ul>
Can I take the exercise in a language that is not my mother tongue
• Please do not! While your knowledge of the second language might be excellent, you decisions might be motivated differently than those of native speakers.
Can I take mutliple turns?
Please do not because your next turn could be affected by the data you have seen already.
I HAVE READ THIS AND LAGREE

Figure 3. Third page of the web interface, indications for the task

Once the indications are read and accepted, the evaluator can start the task, which currently consists of 128 exercise pages (e.g., the one shown in Figure 4). Each page, as can be seen from the Figure, presents 9 words or multi-words on 9 corresponding buttons, one of which is the outlier that needs to be selected.



Figure 4. Example of a set exercise in the web interface

The system is built so that the evaluator cannot click on the answer immediately (Figure 5): the page is "frozen" for 5 seconds (a timer runs above the buttons) and then the evaluator is allowed to click. We implemented this feature because we want the evaluators to carefully read all the words before choosing the outlier, as some may stop at the first words of the set and choose the wrong one.



Figure 5. Set exercise "frozen" before allowing the evaluator to click on the answer

In case the evaluator clicks too fast, or changes their mind on the answer, there is the possibility to undo the last click. Also, as it can be seen from Figure 4, if the evaluator does not want to answer, there is the possibility to click on "I'm not sure" button.

Finally, once the test is completed and all the 128 were evaluated, the evaluator is displayed their overall success score in percentage against the dataset (Figure 6).



Figure 6. Final page of the web interface, after the task is completed, showing the success score against the dataset

### 5.2.2 Selection and training of participants

As stated in Jakubíček et al. (2021), the final goal for the human evaluation is to reach at least 80 evaluators per language. This, unfortunately, was not feasible within this thesis project, due to the limited amount of time that could be dedicated to this part.

We redirected the human evaluation to a different kind of experiment. Indeed, engaging a high number of participants would have required crowdsourcing techniques and the experiment would have been less controlled. With the collaboration of Professor Elisabetta Jezek, who is the supervisor of this thesis, we organized, instead, an experiment in a controlled environment. Thanks to her, we had the possibility to involve 22 students of the Master faculty in Theoretical and Applied Linguistics at the University of Pavia. These students attended one of Professor's Jezek courses, namely the course in Syntax and Semantics (advanced). As we discussed in Section 5.1.1, this group is certainly a well-defined population, and their background in Linguistics helped them in the performance of the task without – we believe – much effort.

The experiment was held at the end of one seminar, in which this thesis' author introduced her thesis' topic and goals. We provided them with an overview on semantic relations, distributional semantics, and thesauri, giving them the possibility to practice on Sketch Engine Thesaurus and Sketch Engine Embedding Viewer. Finally, we introduced the outlier detection task in general and prepared them for the experiment.

The indications given are those in the web interface (Figure 3), which we report here:

- 1. You will be presented 9 words
- 2. Your task is to choose an *outlier*. An outlier is a word that does not fit based on its meaning or sense
- 3. Examples:
  - a. *blue*, *red*, *green*, *yellow*, *orange*, *black*, *brown*, *white*, *table*. Obviously, the last word is the outlier: all the others are names of colours.
  - b. bricklayer, lawyer, shop assistant, gentleman, waitress, metheorologist. Gentleman is an outlier because it is not a job.

Moreover, we asked them to carefully read all the words before choosing the outliers and to take all the time they needed in order to complete the task. Each evaluator was assigned a random sequence of all the 128 sets. For each exercise, that is, for each set visualized, of the 8 inliers were the same for all the evaluators, whereas the outlier was randomly picked from the list of the 8 possible outliers per each set. The 8 inliers + the outlier formed the 9-words sets. On average, it took around 25 minutes to process all the 128 sets.

The reason why we called this a "controlled experiment" is that the participants were trained before the experiment, the environmental setting was the same for all the participants, and this thesis author was present throughout the experiment, thus, she was able to control that the evaluators were performing the task independently and to intervene in case of problems during the experiment.

We believe that these being the conditions, the results can be more reliable than those collected by crowdsourcing, and thus we can use them as a benchmark for the dataset and for further evaluations. Nevertheless, large experiments on human evaluation in the future will be surely done by crowdsourcing, and we are aware that creating such controlled conditions for the experiment are not always achievable.

Finally, we point out here that the experiment was carried out only on Italian language, as all the students were Italian native speakers. We are aware that all the other languages need to be evaluated in terms of human performance, but this was not feasible within this thesis project. It will be certainly done in the future. For the time being, we assume that good results in one language are a good point in order to evaluate all the distributional models for all the languages, as we will do in the following Chapter (Chapter 6).

## 5.2.3 Metrics for the human evaluation: raw agreement

Finally, once the task was completed by all the 22 evaluators engaged, the agreement between them was calculated.<sup>8</sup> As we have already mentioned in Section 2.1.1, we use *raw agreement* as metrics for assessing the Inter-Annotator Agreement.

<sup>&</sup>lt;sup>8</sup> This has been done automatically by Lexical Computing supervision, namely, thanks to Miloš Jakubíček, this thesis co-supervisor and Lexical Computing CEO.

The raw agreement is the measure of the number of correct<sup>9</sup> answers against the overall number of possible answers. The reason why we calculate raw agreement and not other complex metrics is that the agreement by chance does not play a big role, as this is not a binary the number of possibilities for each set are eight, thus the probability that there is agreement by chance is significantly limited (Jakubíček et al., 2021).

We calculate agreement according to various parameters:

- 1. the *overall* agreement, counting all the 22 evaluators
- 2. the agreement per each evaluator, against the dataset
- 3. the agreement per *set* (that is, per each of the 128 sets)

We will recall and better explain these parameters in Section 6.1, in which we discuss the quantitative results of the experiment as for the human evaluation.

## 5.3 Evaluation of Word Embeddings and Sketch Engine Distributional Thesaurus

Once the human evaluation is carried out and the agreement is calculated and assessed as a gold-standard for the model evaluation (as we will discuss in Section 6.1), we proceed to evaluate the distributional models. This part has been entirely managed by Lexical Computing, namely by Miloš Jakubíček, this thesis' co-supervisor and Lexical Computing CEO, and Ondřej Herman, software developer at Lexical Computing.

## 5.3.1 Data and models

In Chapter 1, Section 1.4, we have already presented distributional models: the Sketch Engine distributional thesaurus, on one side, and *word2vec* word embeddings on the other. We have also widely discussed the reasons for their evaluation (Section 2.2) and in which terms their outputs can be compared (Section 1.4.3).

For a better comparison of the outputs, both the thesaurus and the embeddings were computed exactly on the same corpora. For example, we compare Sketch Engine Thesaurus for itTenTen20 corpus and word embeddings calculated on itTenTen20 corpus.

<sup>&</sup>lt;sup>9</sup> "Correct" against the dataset we have built.

In the following Table (Table 1) we report the corpus we used (column 2) for each language (column 1), as well as the size of the corpus (column 3) in number of words.

language	corpus	corpus size
Czech (CS)	csTenTen19	~8 million words
German (DE)	deTenTen20	~21 million words
English (EN)	enTenTen20	~36.5 billion words
Estonian (ET)	estonian_nc21	~2.4 billion words
French (FR)	frTenTen20	~24 billion words
Italian (IT)	itTenTen20	~12 billion words
Slovak (SK	elexis_skTenTen21	~1.2 billion words

Table 1. Corpora used per each target language

All these corpora are available in Sketch Engine. Czech, German, French, English, and Italian are the latest TenTen corpora (2019-2020). TenTen corpora are crawled from web and "TenTen" refers to the target corpus size of 10+ billion words per language.<sup>10</sup> Estonian has its own National Corpus, which also includes data crawled from the web, as well as texts collected from other domains.<sup>11</sup> Finally, also Slovak is a web corpus, but its size – as well as the Estonian corpus size – is much smaller than the other languages.<sup>12</sup>

## 5.3.2 Metrics for the evaluation of distributional models: accuracy and OPP

We evaluate the distributional thesaurus and word embeddings through two evaluation metrics.

One is *accuracy*, which we have already presented in Section 2.1.2. Accuracy is the percentage of answers correctly identified by the model against the dataset; it is calculated by simply dividing the number of correct answers by the overall number of answers. This metrics is analogous to the *raw agreement* for the human evaluation, therefore making the two scores comparable, as we will see in Section 6.5.

<sup>&</sup>lt;sup>10</sup> For more information see: <u>https://www.sketchengine.eu/documentation/tenten-corpora/</u> (last access: 24/06/2022).

<sup>&</sup>lt;sup>11</sup> For more information see: <u>https://www.sketchengine.eu/estonian-national-corpus/</u> (last access: 24/06/2022).

<sup>&</sup>lt;sup>12</sup> It is clear that the size also depends on the number of native language speakers that produce the texts crawled in the corpora: compare the population of Estonia (1.3 million in 2020) and Slovakia (5.5 million in 2020) vs. the population of Italy (60 million in 2020) and Germany (83.2 million in 2020).
The second score is the *outlier position percentage* (OPP),<sup>13</sup> a custom-made metrics implemented by Camacho-Collados & Navigli (2016) which we also exploit for the evaluation. Putting it simply, OPP indicates how close the outliers are to being correctly classified (Rychlý, 2019; Andersen et al., 2020). This metrics can be formalized as follows (Camacho-Collados & Navigli, 2016):

$$OPP = \frac{\sum_{W \in D} \cdot \frac{OP(W)}{|D|}}{|D|} \cdot 100$$

In the formula, W is a word set (8 inliers + 1 outlier) and D is the dataset consisting of a number of W (sets). OP (outlier position) is the position of the outlier in the list of words (inliers + outlier) in the list of the words ordered by compactness. The *compactness* score of a word in a set is the compactness of the cluster (i.e., the 8 inliers), calculated by averaging the semantic similarity score (i.e., cosine) of the inliers combined in pairs.<sup>14</sup>

The outlier position (OP) is a number from 0 to the number of words in the set (9 in the original dataset and in ours as well). 0 means the worst guess, the maximum means the correct answer. Therefore 1.00 OPP (= 100%) means 100% accuracy.

The outlier position percentage provides a more fine-grained evaluation metrics with respect to the accuracy, as, for incorrect answers the position of the right answer is differentiated and ranked according to the position of the outlier in the list (the closer, the more related, and vice versa).

#### 5.3.3 Calculating the results

The outlier detection task was performed on the distributional models by calculating the accuracy and the OPP through an adapted version (see Rychlý, 2019) of the Python script for the outlier detection originally provided by Camacho-Collados & Navigli (2016).

In a few words, the script takes as input the two models and returns as output first a short summary of the results (namely, accuracy and OPP results), and, if needed, more detailed results divided per set. Therefore, our analysis will be carried out according to the following parameters:

- 1. overall accuracy and OPP per model, per each language
- 2. accuracy per set (that is, per each of the 128 sets)

<sup>&</sup>lt;sup>13</sup> We introduced it in Section 2.3.1.

<sup>&</sup>lt;sup>14</sup> For additional explanations, see Camacho-Collados & Navigli (2016), Rychlý (2019), Andersen et al. (2020).

The output results are processed from raw .txt into .xlsx tables and organized for the discussion of the results in Chapter 6.

# **Chapter 6. Quantitative and Qualitative Analysis of the Results of the Experiment and Further Improvements**

In this Chapter we present and analyse the results of the experiment we conducted with HAMOD dataset on the outlier detection task. Our discussion of the results takes two directions: a quantitative analysis and a qualitative analysis. Furthermore, we provide a systematic comparison of the human and models' performances, in order to highlight analogies and differences among the two and to gain a better understanding of distributional models' properties.

Here follows a brief outline of the Chapter.

In Section 6.1 we conduct a quantitative analysis of the results from the human evaluation benchmark. We comment on the overall performance of the Italian native speakers evaluators and then we focus on the distribution of the human errors in general, and as far as the difficulty of the outlier and of the part of speech of the set is concerned.

In Section 6.2 we proceed with the quantitative analysis of the results from the distributional models' evaluation. In parallel to the human evaluation analysis, we focus on the overall performances according to the accuracy and OPP rates, and then we move to the analysis of the errors.

In Section 6.3 we provide a qualitative linguistic analysis of the results from the human evaluation, as far as Italian is concerned. We consider the most common errors in a qualitative perspective, in an attempt to give explanatory insights on what may have led to fail the outlier detection. After analysing the most common errors, we structure the analysis in two directions: first, we analyse the results according to the type of set (i.e., semantic category-based or topic-based);<sup>1</sup> after, we comment further on the diverse distribution of errors in the sets according to their part of speech.

In Section 6.4 we perform a similar qualitative linguistic analysis on the results from the models' evaluation. In this case, we orient our analysis in a multi-lingual perspective, as the models have been evaluated for various languages (whereas the human evaluation was conducted only for Italian language, as we explained in Chapter 5).

In Section 6.5 we compare the human and the models' performance, with a focus on Italian language. In this Section we are interested in spotting some discrepancies and commonalities between humans and the models, keeping the sets as a unit of analysis.

<sup>&</sup>lt;sup>1</sup> For this distinction, we recall Section 3.2.1.

In Section 6.6, we comment on the results in general, referring to the initial hypothesis we had on the possible results of the experiment. Moreover, we provide some comments on possible improvements of the dataset, with an eye on the weakest points (that is, those in which human evaluators first, and the models subsequently were mostly mistaken), and of the experiment setup in general.

### 6.1 Quantitative Analysis: Human Evaluation

The experiment conducted on the 22 human evaluators on the Italian part of the dataset was, as we expected, successful in terms of the rate of agreement between the participants (92% raw agreement). The results are anonymised (each evaluator is labelled as "student\_*n*"). First, we report the results per evaluator (Table 1): the number of sets they successfully completed (i.e., they provided the correct answer finding the correct outlier in the set against the dataset) (column 2); the number of sets in which they failed to detect the outlier (an inlier was selected instead) (column 3); the number of sets which were skipped (we recall the "I'm not sure" possibility, which the evaluators could use in case they were not able to detect the outlier) (column 4); the total number of sets evaluated (which has to be the overall number of sets, 128 in the current version of the dataset) (column 5); and finally, the raw agreement per evaluator, which is calculated against the dataset as the percentage of correct answers over the total number of possible answers (as described in Section 5.2.3) (column 6).<sup>2</sup>

<sup>&</sup>lt;sup>2</sup> The results are arranged from the best to the worst performance, in terms of percentage of agreement.

evaluator	number of	number of	number of	total number	raw
	correct sets	mistaken sets	skipped sets	of evaluated	agreement per
				sets	evaluator
student_1	128	0	0	128	1.00
student_2	124	2	2	128	0.97
student_3	124	1	3	128	0.97
student_4	122	3	3	128	0.95
student_5	121	5	2	128	0.95
student_6	121	1	6	128	0.95
student_7	120	5	3	128	0.94
student_8	119	9	0	128	0.93
student_9	119	7	2	128	0.93
student_10	119	6	3	128	0.93
student_11	119	6	3	128	0.93
student_12	119	5	4	128	0.93
student_13	118	7	3	128	0.92
student_14	118	5	5	128	0.92
student_15	118	5	5	128	0.92
student_16	118	0	10	128	0.92
student_17	117	6	5	128	0.91
student_18	117	5	6	128	0.91
student_19	115	8	5	128	0.90
student_20	110	10	8	128	0.86
student_21	109	14	5	128	0.85
student_22	108	7	13	128	0.84
overall counts	2603	117	96	2816	0.92

Table 1. Experiment results on the human evaluation, per evaluator

In the following Table (Table 2) we report the overall results from the last raw in the previous Table, as the average of the single evaluators performance. Given 22 evaluators, the number of correct, mistaken, skipped and overall sets is the sum of each evaluator performance; the overall raw agreement is the average of the agreement for each participant (see Table 1, column 6). If we consider the number of sets, which is more than three times higher than those in the original dataset, we can state that the experiment was successful, and the results can be considered as a benchmark for the models' evaluation.

Table 2. Global results on the human evaluation task

number of evaluators	22
number of correct sets	2603
number of mistaken sets	117
number of skipped sets	96
overall number of sets	2816
raw agreement	0.92

We can conduct further quantitative analysis on the data, especially with a focus on the distribution of errors with respect to the dataset standard.

First, we can consider how many times evaluators did not answer correctly per each set (or, in other words, as each evaluator was shown each set in the dataset only once, how many evaluators got the wrong answer for a particular set). As we can see from Table 3, the maximum number of evaluators jointly missing the correct answer for a specific set is 6 out of 22, and this affects a really limited number of sets. The Table is structured as follows: in the first column, there is the number of evaluators making a mistake on a specific set; in the second column, the number of sets affected by mistakes (out of the 128 sets in the dataset); from column 3 to 6 we report the distribution of wrong answers; in column 7 the percentage of the distribution; in column 8 the overall number of mistakes. In general, we can notice that a low number of sets is affected by a high number of mistakes (see the lowest rows in the Table), and vice versa; 40 sets were correctly guessed by all the 22 evaluators.

number of	number of	number of	number of	number of	number of	percentage	overall
evaluators	sets	sets with	sets with	sets with	sets with		
missing	affected	1 wrong	2 wrong	3 wrong	4 wrong		
the outlier		answer	answers	answers	answers		
0	40	0	0	0	0	0.31	
1	25	13	0	0	0	0.20	
2	28	13	11	0	0	0.22	
3	17	6	6	3	0	0.13	
4	12	0	2	6	1	0.09	
5	3	1	1	1	0	0.02	
6	3	0	0	2	1	0.02	
overall	128	33	40	36	8	1.00	<b>117</b> <sup>3</sup>

Table 3. Distribution of mistaken sets among the evaluators

Moreover, we can calculate a similar statistic for the number of skipped sets, which are complementary to the incorrectly detected sets (the distribution in percentage is exactly the same). From Table 4, as above (the structure of the Table is analogous), we can see that the maximum number of evaluators jointly skipping a specific set is 6 out of 22, and this affects a really limited number of sets.

 $<sup>^{3}</sup>$  The counts in this raw are calculated by multiplying the unique number of errors per the number of actual errors (e.g., in column 4 the sum is 20, which has to be multiplied by 2).

Table 4. Distribution of the skipped sets among the evaluators

number of	number of	number of	number of	number of	number of	percentag	overall
evaluators	sets	sets with 1	sets with 2	sets with 3	sets with 4	е	
skipping	affected	skipped	skipped	skipped	skipped		
the set		answer	answer	answer	answer		
1	25	12	0	0	0	0.20	
2	28	13	4	0	0	0.22	
3	17	6	6	2	0	0.13	
4	12	6	2	0	3	0.09	
5	3	0	1	1	1	0.02	
6	3	0	1	2	0	0.02	
overall	128	37	28	15	4	1.00	96

Then, we can compute the distribution of errors according to the position of the outlier in the dataset (i.e., how close the outlier is to the inliers as we arranged them within the dataset). The closer the outlier is to the inliers, the more challenging its detection should be, as the evaluators may be consider an outlier as part of the inliers and wrongly detect an inlier as the outlier. In the following Table (Table 5), we report the number of errors (out of the 117 overall errors distributed among the 22 evaluators) (column 2) per position of the outlier in the dataset (column 1) and the distribution in percentage (column 3). The outliers are grouped in pairs, as each item in the pair is equally close to the inliers (1 & 2 are the closest, 7 & 8 the furthest, see Section 3.2. Indeed, as we can see from the Table, the distribution reflects the difficulty of the outlier: a highest number of errors affects the most related items in the ranking (i.e., outliers in the first positions), and vice versa. This proves that the arrangement of the outliers according to the indications we discussed in 3.2 works, as the evaluators were mainly misled by inliers-related words rather than what we included in the dataset as furthest words. In other words, we can see that the higher the outlier is in the ranking (first rows in the Table), the higher number of mistakes occurs.

position of the outlier	number of errors	percentage
1 & 2	53	0.45
3 & 4	39	0.33
5 & 6	21	0.18
7 & 8	4	0.03
overall	117	1.00

Table 5. Distribution of the errors according to the position of the outlier in the dataset

Another issue is the distribution of the errors according to the part of speech of the set. We recall that we have three types of sets, per part of speech: noun-only sets, adjectiveonly-sets, verb-only sets (different parts of speech cannot be combined in the same set). As we can see from Table 6, reporting the number of sets with at least one mistake, taken only once (column 3) on the overall number of sets per part of speech (column 2) and their percentage (column 4),<sup>4</sup> verb-based sets are more subject to errors (67% of the verb sets were mistaken at least one time) than adjective and nouns. Surprisingly, the adjective-based sets have the lowest rate of error, although this may also be due to the limited amount of this kind of sets in the dataset, with respect to the noun-based and the verb-based.

part of speech	overall number of sets	number of affected sets	percentage
	for the part of speech		
noun	86	42	0.49
verb	30	20	0.67
adjective	12	4	0.33
overall	128	66	1.00

Table 6. Distribution of the errors per part of speech

Finally, we report in this Section the detailed quantitative results per each set (Table 7), by listing all the 128 sets, ranked first by agreement rate (column 6) and number of evaluators who correctly guessed the set (column 3), then according to the number of wrong sets (column 4) and finally alphabetically (column 1). Therefore, the sets are ranked according to the agreement, with the highest items in the Table being the ones which were subject to less error, and vice versa. We will comment on this data in more detail in the qualitative analysis, in an attempt to provide an interpretation of the rates for each set (see Section 6.4).

Table 7. Distribution of the errors per each of the 128 sets

set name	number of	number of	number of	number of	agreement
	evaluators	evaluators	evaluators	evaluators	per set
		correctly	wrongly	skipping	
		detecting	detecting	the set	
		the outlier	the outlier		
Birds	22	22	0	0	1.00
Bugs	22	22	0	0	1.00
Car_Components	22	22	0	0	1.00
Chemical_Elements	22	22	0	0	1.00
Colours	22	22	0	0	1.00
Dances	22	22	0	0	1.00
Dishes_and_Cutlery	22	22	0	0	1.00
Family_Members	22	22	0	0	1.00

<sup>&</sup>lt;sup>4</sup> Percentages are calculated by dividing the number of affected sets by the overall number of sets for that particular part of speech.

Food	22	22	0	0	1.00
Fruit	22	22	0	0	1.00
Herbs/IT-Herbs.txt	22	22	0	0	1.00
Human_Features_Negativity	22	22	0	0	1.00
Human_Moods	22	22	0	0	1.00
Informatics	22	22	0	0	1.00
Internal_Body_Parts	22	22	0	0	1.00
Kitchenware	22	22	0	0	1.00
Languages	22	22	0	0	1.00
Materials	22	22	0	0	1.00
Maths	22	22	0	0	1.00
Means_of_Transport	22	22	0	0	1.00
Metals	22	22	0	0	1.00
Musical_Instruments	22	22	0	0	1.00
Parts_of_Head	22	22	0	0	1.00
Professions	22	22	0	0	1.00
Shapes	22	22	0	0	1.00
Shoes	22	22	0	0	1.00
Shops	22	22	0	0	1.00
Spices	22	22	0	0	1.00
Spirits	22	22	0	0	1.00
Sport	22	22	0	0	1.00
Temperature_Features	22	22	0	0	1.00
Vegetables	22	22	0	0	1.00
Verbs_Communication_1	22	22	0	0	1.00
Verbs_Farming	22	22	0	0	1.00
Verbs_Motion	22	22	0	0	1.00
Verbs_Plants	22	22	0	0	1.00
Verbs_Religion	22	22	0	0	1.00
Weapons	22	22	0	0	1.00
Weather_Conditions	22	22	0	0	1.00
Zodiac_Signs	22	22	0	0	1.00
Electronics	22	21	1	0	0.95
Free_Time_Activities	22	21	1	0	0.95
Linguistics	22	21	1	0	0.95
Liquid_Containers	22	21	1	0	0.95
Music_Genres	22	21	1	0	0.95
Non-alcoholic_Drinks	22	21	1	0	0.95
Road_Means_of_Transport	22	21	1	0	0.95
School_Subjects	22	21	1	0	0.95
Sweets	22	21	1	0	0.95
Verbs_Crime	22	21	1	0	0.95
Verbs_Driving	22	21	1	0	0.95
Verbs_Killing	22	21	1	0	0.95
Water_Means_of_Transport	22	21	1	0	0.95
Clothes	22	21	0	1	0.95
Dimensional_Features_1	22	21	0	1	0.95
Flowers	22	21	0	1	0.95
Gemstones	22	21	0	1	0.95

22	21	0	1	0.95
		_		0.95
			-	0.95
		_		0.95
				0.95
		_		0.95
				0.95
				0.95
			-	0.93
				0.91
				0.91
			-	0.91
			-	0.91
				0.91
				0.91
				0.91
				0.91
				0.91
			-	0.91
				0.91
				0.91
				0.91
				0.91
			-	0.91
				0.91
				0.91
				0.91
				0.91
				0.91
				0.91
			-	0.91
				0.91
				0.91
				0.91
				0.91
				0.91
				0.86
				0.86
				0.86
			1	0.87
				0.86
			1	0.86
22	19	2	1	0.86
22	19	2	1	0.86
				0.86
22	19	1		0.86
22	19	1	2	0.86
22	19	1	2	0.86
	19	1	2	0.86
	22 22 22 22 22	22         21           22         21           22         21           22         21           22         21           22         21           22         21           22         21           22         20           22         <	$\begin{array}{c c c c c c c c c c c c c c c c c c c $	22 $21$ $0$ $1$ $22$ $21$ $0$ $1$ $22$ $21$ $0$ $1$ $22$ $21$ $0$ $1$ $22$ $21$ $0$ $1$ $22$ $20$ $2$ $0$ $22$ $20$ $2$ $0$ $22$ $20$ $2$ $0$ $22$ $20$ $2$ $0$ $22$ $20$ $2$ $0$ $22$ $20$ $2$ $0$ $22$ $20$ $2$ $0$ $22$ $20$ $2$ $0$ $22$ $20$ $1$ $1$ $22$ $20$ $1$ $1$ $22$ $20$ $1$ $1$ $22$ $20$ $1$ $1$ $22$ $20$ $1$ $1$ $22$ $20$ $1$ $1$ $22$ $20$ <

Savanna_Animals	22	19	1	2	0.86
Verbs_Telephone	22	19	1	2	0.86
Buildings	22	19	0	3	0.86
Medicine	22	19	0	3	0.86
Verbs_Perception	22	18	4	0	0.82
Art	22	18	3	1	0.82
Fish	22	18	3	1	0.82
Parts_of_Speech	22	18	3	1	0.82
Reptiles	22	18	3	1	0.82
Verbs_Cognition	22	18	3	1	0.82
Verbs_Eating	22	18	3	1	0.82
Cooking	22	18	2	2	0.82
Verbs_Mouth	22	18	2	2	0.82
Building_Materials	22	18	0	4	0.82
Sources_of_Energy	22	18	0	4	0.82
Verbs_Communication_2	22	18	0	4	0.82
Bodies_of_Water	22	17	3	2	0.77
Verbs_Psych	22	17	2	3	0.77
Sports	22	17	1	4	0.77
Human_Features_Positivity	22	16	4	2	0.73
Computer_Components	22	16	3	3	0.73
Verbs_School	22	16	3	3	0.73

We now proceed with the quantitative analysis of the models' performance, and then, in Section 6.3, get back to the human performance and analyse the results qualitatively and in a linguistic perspective.

## 6.2 Quantitative Analysis: Word Embeddings and Sketch Engine Thesaurus Evaluation

We performed the outlier detection task on Sketch Engine Thesaurus (SkeThe in the Tables) and Word Embeddings (WE in the Tables) calculated on the corpora we mentioned in Section 5.3.1.<sup>5</sup> Word Embeddings were calculated both from "word" attribute (WE\_word in the Tables) (that is, embeddings calculated on raw corpus text, with no

<sup>&</sup>lt;sup>5</sup> The corpora are listed in the Table below.

annotations) and from "lemma" (WE\_lemma in the Tables) attribute (that is, embeddings calculated from lemmatized corpora).

We first provide a general overview of the results of the evaluation, comparing accuracy and  $OPP^6$  for each model and each language. In Table 8 we confront the results in terms of *accuracy*, i.e., the percentage of correct answers, for Sketch Engine Thesaurus (column 3), Word Embeddings with word attribute (column 4) and Word Embeddings with lemma attribute (column 5) for each language (column 1). In column 2, we report the number of unique exercises each model performed (the number is obtained by multiplying the number of sets in the dataset, 128, per the number of 8 inliers + 1 outlier combinations, being the outliers 8 per set). In column 6, we calculate the average performance for each language on the three models, in order to assess which language gave the best and the worst results in accuracy.

language and corpus	n. of	SkeThe	WE_word	WE_lemma	average
	exercises				
CS (csTenTen19)	1024	0.482	0.687	0.628	0.599
DE (deTenTen20)	1024	0.478	0.657	0.647	0.594
EN (enTenTen20)	1024	0.403	0.618	0.596	0.539
ET (estonian_nc21)	1024	0.562	0.685	0.659	0.635
FR (frTenTen20)	1024	0.400	0.621	0.574	0.531
IT (itTenTen20)	1024	0.419	0.670	0.551	0.547
SK (elexis_skTenTen21)	1024	0.442	0.673	0.680	0.598

Table 8. Global accuracy, per language, for Sketch Engine Thesaurus and Word Embedding models

What emerges from the previous Table is that Word Embeddings with both attributes, "word" and "lemma", outperform Sketch Engine Thesaurus. As we will soon see below, this was not the case in previous tests on the outlier detection. As for Sketch Engine Thesaurus, Estonian is the language with the highest accuracy (0.562) and French the one with the lowest (0.400). Word Embeddings with "word" attribute surprisingly give better results than those calculated from "lemma" attribute (except for Slovak). We would expect that lemmatization helped disambiguate the words in the dataset when the model performs the outlier detection, but this seems not to be the case. At this level of analysis, we cannot explore possible reasons for this behaviour, but one variable affecting the results may be the morphology of each specific language, the lemmatizer used on the corpora, or the frequency of the words in the dataset for that language. As for Word Embeddings with attribute "word", Czech is the language with the highest accuracy (0.687) and English with the lowest (0.403), and for "lemma" attribute, Slovak with the highest (0.680) and Italian with the lowest (0.551). Results for Word Embeddings "word"

<sup>&</sup>lt;sup>6</sup> See Section 5.3.2.

attribute are rather homogeneous with each other, whereas there is more variability in the other two models. Yet, we cannot detect any clear pattern in the accuracy, as each language performs differently according to the model. On average, Estonian seems to have the highest accuracy and French the lowest, on the outlier detection task.

In Table 9 we report the results in terms of OPP, i.e., how precise the model is at the detecting the outlier (the higher the score, the more precise). The structure of the Table is the same as in Table 8).

language and corpus	n. of	SkeThe	WE_word	WE_lemma	average
	exercises				
CS (csTenTen19)	1024	0.851	0.926	0.901	0.892
DE (deTenTen20)	1024	0.841	0.921	0.905	0.889
EN (enTenTen20)	1024	0.837	0.898	0.890	0.875
ET (estonian_nc21)	1024	0.847	0.924	0.905	0.892
FR (frTenTen20)	1024	0.821	0.910	0.887	0.873
IT (itTenTen20)	1024	0.829	0.923	0.885	0.879
SK (elexis_skTenTen21)	1024	0.824	0.913	0.919	0.885

Table 9. Global OPP, per language, for Sketch Engine Thesaurus and Word Embedding models

The results are analogous to those on accuracy rates. Indeed, Word Embedding models outperform Sketch Engine Thesaurus, and Word Embeddings calculated with "word" attribute outperform those calculated with "lemma", except for Slovak. Czech has the highest results in the thesaurus (0.851), and French the lowest (0.821); Czech, again, has the highest results in Word Embeddings with attribute "word" (0.926) and, again, French has the lowest (0.910); finally, Slovak (0.919) has the highest results in Word Embeddings with attribute "word" (0.926) and, again, French has the lowest (0.810); finally, Slovak (0.919) has the highest results in Word Embeddings with attribute "lemma" and Italian the lowest (0.885). There is a clear correlation between the two measures, as we can see from the following Table (Table 10), in which OPP and accuracy are put together. On average, as for OPP, Czech and Estonian seem to have the highest OPP and French the lowest.

language	exercises	SkeThe	SkeThe	WE_word	WE_word	WE_	WE_
		accuracy	OPP	accuracy	OPP	lemma	lemma
						accuracy	OPP
CS	1024	0.482	0.851	0.687	0.926	0.628	0.901
DE	1024	0.478	0.841	0.657	0.921	0.647	0.905
EN	1024	0.403	0.837	0.618	0.898	0.596	0.890
ET	1024	0.562	0.847	0.685	0.924	0.659	0.905
FR	1024	0.400	0.821	0.621	0.910	0.574	0.887
IT	1024	0.419	0.829	0.670	0.923	0.551	0.885
SK	1024	0.442	0.824	0.673	0.913	0.680	0.919

Table 10. Global results, accuracy and OPP for all the models

It is worth comparing these results with those from a previous experiment on the original dataset (37 sets) carried out on the same languages, but on other corpora, with a smaller size in terms of words (Jakubíček et al., 2021). What we are interested in is verifying whether there has been some improvement or worsening in the performance of each language on the outlier detection task, from one version of the dataset to the other, and whether we can detect the same pattern in terms of which language did the best or the worst.

First, we report the results from Jakubíček et al. (2021), in Table 11. The Table is structured as the previous ones.

language and	n. of	SkeThe	SkeThe OPP	WE accuracy	WE OPP
corpus	exercises	accuracy			
CS (czTenTen12)	232	0.573	0.898	0.655	0.871
DE (deTenTen13)	232	0.349	0.798	0.323	0.746
EN (enTenTen13)	296	0.456	0.847	0.655	0.873
ET (estonian_nc17)	296	0.564	0.832	0.547	0.784
FR (frTenTen12)	232	0.276	0.744	0.427	0.768
IT (itTenTen16)	296	0.453	0.856	0.581	0.869
SK (skTenTen11)	296	0.389	0.777	0.591	0.851

Table 11. Global results on a previous experiment on the outlier detection (Jakubiček et al., 2021), Sketch Engine Thesaurus and Word Embeddings are compared

What emerges from this Table is that, in general, the embeddings (computed only for "word" attribute) outperform the thesaurus in accuracy, except for German and Estonian. In terms of OPP, the thesaurus outperforms the embeddings in Czech, German and French.

A different scenario emerges if we compare our thesis' results with those in Jakubíček et al. (2021). We split the two measures, first analysing changes in accuracy, and then in OPP.

In Table 12, for each language (column 1), we report the accuracy of Sketch Engine Thesaurus on the original 37 sets dataset (column 2) and on HAMOD dataset (column 3); and the accuracy of Word Embeddings on the original dataset (column 4) and on the new dataset (column 5).

language	original dataset	HAMOD dataset –	original dataset –	HAMOD dataset –
	SkeThe accuracy	SkeThe accuracy	WE accuracy	WE accuracy
CS	0.573	0.482	0.655	0.687
DE	0.349	0.478	0.323	0.657
EN	0.456	0.403	0.655	0.618
ET	0.564	0.562	0.547	0.685
FR	0.276	0.400	0.427	0.621
IT	0.453	0.419	0.581	0.670
SK	0.389	0.442	0.591	0.673

Table 12. Global results in accuracy on the original vs. HAMOD dataset, for Sketch Engine Thesaurus and Word Embeddings

We can derive some conclusion on the comparison of the two datasets on the models. First, if we consider Sketch Engine Thesaurus, the performance is worse with the new dataset, except for German, French and Slovak which had a significant improvement. The reason for this discrepancy may be the different lemmatizers used to annotate the corpora.<sup>7</sup> Furthermore, German and French original datasets were not built properly, as they contained several typos, wrong words, misspellings, incorrect diacritics. Thus, the systematization we carried out on these part of the dataset was successful as the accuracy significantly increased. Another case is that of the Word Embeddings: accuracy on HAMOD dataset is always higher than on the original dataset, except for English, whose accuracy is higher in the original dataset.

In Table 13, we conduct a similar analysis on the OPP scores. The structure of the Table is the same as in Table 12.

Table 13. Global results in OPP on the original vs. HAMOD dataset, for Sketch Engine Thesaurus and Word Embeddings

language	original dataset	HAMOD dataset –	original dataset –	HAMOD dataset –
	SkeThe OPP	SkeThe OPP	WE OPP	WE OPP
CS	0.898	0.851	0.871	0.926
DE	0.798	0.841	0.746	0.921
EN	0.847	0.837	0.873	0.898
ET	0.832	0.847	0.784	0.924
FR	0.744	0.821	0.768	0.910
IT	0.856	0.829	0.869	0.923
SK	0.777	0.824	0.851	0.913

<sup>&</sup>lt;sup>7</sup> We have to keep in mind that these results are calculated on different corpora, and older one for the original dataset, and a new one for HAMOD.

What emerges from this comparison is similar to what we commented in the previous paragraph, except for the fact that with OPP also Estonian, with German, French and Slovak, has a higher OPP with HAMOD dataset (whereas for the other languages it is the other way round); furthermore, none of the language has a higher OPP with the original dataset than the OPP with HAMOD dataset (even English, in this case).

To sum up, we can outline some remarks on this quantitative analysis. With a few exceptions, Word Embeddings outperform Sketch Engine Thesaurus in the outlier detection, both in terms of accuracy and of OPP, as the two metrics are correlated. With respect to the previous experiment on the original dataset, there is a significant improvement, from the original dataset to HAMOD dataset (and from an older and smaller corpus, to a bigger one) as far as the Word Embeddings are concerned; instead, from Sketch Engine Thesaurus there is no clear pattern, as some of the languages worsened with HAMOD dataset (namely, Czech, English, Estonian, Italian). Finally, in the experiment with the HAMOD dataset the results seem to be more homogeneous, in terms of diversity from one language to another.

In the last part of this Section, we introduce the detailed quantitative results per each of the 128 sets. A Table in which the sets are listed, with accuracy and OPP results per language and model can be visualized in Appendix 3. We will comment on this data in more detail in the qualitative analysis, in an attempt to provide an interpretation of the individual percentages for each set (see Section 6.4).

### 6.3 Qualitative Analysis: Human Evaluation

In this Section we proceed to the qualitative linguistic analysis of the results obtained from the human evaluation, as far as Italian language is concerned. We consider the most common errors from a qualitative perspective, in an attempt to give explanatory insights on what may have led evaluators to fail the detection of the outlier. After analysing the most common errors, we structure the analysis in two directions: first, we analyse the results according to the type of set (i.e., semantic category-based or topic-based);<sup>8</sup> after, we comment further on the diverse distribution of errors in the sets according to their part of speech.

<sup>&</sup>lt;sup>8</sup> For this distinction, we recall Section 3.2.1.

One of the focuses of this Section is the analysis of the errors. Given the high agreement (92%) among the evaluators, it is worth analysing in detail which were the most common errors ad which sets were most frequently mistaken, also as a useful insight with a view to further implementations or changes to HAMOD dataset. We recall here that the evaluators were only given a set<sup>9</sup> of nine words (8 inliers + 1 outlier) without the set name; therefore, they had to infer it by guessing what the words had in common.

As we saw in Section 6.1, in 31% of the sets (Table 3) the outliers were correctly detected by all the 22 evaluators.<sup>10</sup> From Table 7, we can collect some examples: among these sets, there are some typical semantic categories – according to this thesis' author, who compiled the dataset, such as "Fruit", "Colours", "Professions", "Musical\_Instruments", as well as some topics, such as "Maths", "Informatics", "Sport". Less typical semantic categories are also included: "Verbs\_Motion", "Verbs\_Plants", "Verbs\_Religion", "Human\_Features\_Negativity", "Human\_Moods".<sup>11</sup>

Around 13% of the sets gained less than 0.85 agreement, with 4 or more evaluators out of 22 could not detect the outlier in the set of words. Here (Table 14) follows the list of these sets and their scores, taken from Table 3 (the structure of the Table is analogous).

set name	number of	number of	number of	number of	agreement
	evaluators	evaluators	evaluators	evaluators	per set
		correctly	wrongly	skipping	
		detecting	detecting	the set	
		the outlier	the outlier		
Verbs_Perception	22	18	4	0	0.82
Art	22	18	3	1	0.82
Fish	22	18	3	1	0.82
Parts_of_Speech	22	18	3	1	0.82
Reptiles	22	18	3	1	0.82
Verbs_Cognition	22	18	3	1	0.82
Verbs_Eating	22	18	3	1	0.82
Cooking	22	18	2	2	0.82
Verbs_Mouth	22	18	2	2	0.82
Building_Materials	22	18	0	4	0.82
Sources_of_Energy	22	18	0	4	0.82
Verbs_Communication_2	22	18	0	4	0.82

Table 14. List of the sets with less than 0.85 agreement, ranked according to the agreement per set

<sup>&</sup>lt;sup>9</sup> As we explained in Section 3.2.1, sets are composed by 8 semantically similar words (the *inliers*) and 8 words that are not related at various degrees (the *outliers*).

 $<sup>^{10}</sup>$  Each evaluator had a random outlier (therefore, the difficulty of the set was different) taken from the list of the 8 outliers; the inliers, instead, were the same for all the evaluators (indeed, 8 inliers + 1 outliers per exercise, with no reiterations).

<sup>&</sup>lt;sup>11</sup> For the content of these sets, see Appendix 1; for the definitions of each set, see Section 3.4.

Bodies_of_Water	22	17	3	2	0.77
Verbs_Psych	22	17	2	3	0.77
Sports	22	17	1	4	0.77
Human_Features_Positivity	22	16	4	2	0.73
Computer_Components	22	16	3	3	0.73
Verbs_School	22	16	3	3	0.73

Some of the included sets are typical semantic categories, such as "Fish" and "Reptiles" (they are truly taxonomic), and "Sports". Some topics are also included – "Art", "Cooking" – as well as some verb sets "Verbs\_Perception", "Verbs\_Eating", "Verbs\_Psych", "Verbs\_School", which we would expect to be harder to identify as semantic clusters.

Then, we gathered the errors, in terms of pairs of words, out of which one word is the one selected as the outlier by the evaluator (but which was an inlier instead) and the other is actual outlier for that set. In the following Table (Table 15), these pair of words are ranked according to the position of the outlier in the dataset, thus those being in the highest position (1, 2) being the most challenging to detect. The Table is structured as follows: in the first column we report the name of the set; in the second the position of the outlier (1 to 8); in the third column the outlier which was not detected; in the fourth column, the inlier that was selected as outlier instead of the inlier in the third column; in the fifth column, the number of times this error happened (that is, the number of evaluators which made this specific mistake). Notice that some sets (e.g., "Verbs\_Economics") can be affected by more than one mistaken inlier, and that more than one evaluator may have done the same mistake (the reason for this is that, being 8 the possible combinations per set and 22 the evaluators, the same combination of 8 inliers + 1 outliers was randomly assigned to 2 or 3 evaluators).<sup>12</sup>

set name	outlier	correct outlier	inlier mistaken for	n. of
	position		outlier	evaluators
Cooking	1	birreria	chef	1
Dimensional_Features_2	1	acuto	conico	1
External_Body_Parts	1	rene	fianco	1
Farm_Animals	1	volpe	gallina	1
Firearms	1	proiettile	kalashnikov	1
Fish	1	delfino	carpa	1

Table 15. Detailed report of the evaluators error, with pairs of correct outlier – mistaken inlier per set and ranked according to the position of the outlier

<sup>&</sup>lt;sup>12</sup> We cannot translate all the words in the Table, please refer to Appendix 1 for the corresponding words in all the languages of the dataset. We will only provide the translations for the pair of words we include in the comment in the following Sections.

Flying_Means_of_Transport	1	automobile	parapendio	1
Hair_Features	1	rugoso	pelato	1
Human_Features_Positivity	1	alto	collaborativo	2
Landscape_Features	1	città	cascata	1
Music	1	lettera	rock	1
Parts_of_House	1	linoleum	pavimento	1
Reptiles	1	rana	serpente	2
Sports	1	jogging	sci_di_fondo	1
Verbs_Dog	1	ruggire	scodinzolare	1
Verbs_Economics	1	rapinare	investire	1
	1	rapinare	indebitare	1
Verbs_Hair	1	stirare	tingere	1
	1	stirare	rasare	1
Verbs_Sport	1	giacere	pattinare	1
Book_Genres	2	film	racconto_breve	1
Extreme_Natural_Events	2	vento	incendio	1
Fish	2	balena	anguilla	1
	2	balena	carpa	1
Free_Time_Activities	2	calcio	cucina	1
Music_Genres	2	opera	hip_hop	1
Non-alcoholic_Drinks	2	vino	frappè	1
Nuts	2	uvetta	castagna	2
Parts_of_House	2	piastrella	pilastro	1
Reptiles	2	ippopotamo	tartaruga	1
Road_Means_of_Transport	2	aeroplano	bicicletta	1
Savanna_Animals	2	lepre	giraffa	1
School_Subjects	2	ceramica	lingua_straniera	1
Verbs_Cooking_1	2	servire	saltare	1
Verbs_Crime	2	condannare	mentire	1
Verbs_Destroying	2	scoppiare	rovinare	1
	2	scoppiare	sterminare	1
Verbs_Dog	2	squittire	fiutare	1
Verbs_Driving	2	deragliare	guidare	1
Verbs_Eating	2	digiunare	mordere	1
	2	digiunare	deglutire	1
Verbs_Human_Sounds	2	sibilare	ridacchiare	1
	2	sibilare	ruttare	1
Verbs_Mouth	2	espirare	sorridere	1
Verbs_Perception	2	silenziare	percepire	1
Verbs_School	2	interrogare	contare	2
Verbs_Telephone	2	strillare	messaggiare	1
Verbs_Touch	2	annusare	graffiare	1
Water_Means_of_Transport	2	mongolfiera	nave	1
Biomes	3	isola	macchia_mediterranea	1
Bodies_of_Water	3	piscina	baia	1
Book_Genres	3	libro	diario_di_viaggio	1
Dimensional_Features_2	3	geometrico	quadrato	1
Fantasy_Characters	3	bacchetta_magica	zombie	1
Fruit_Trees	3	giglio	arancio	1

Hair_Features	3	roseo	spettinato	1
Human_Physical_Features	3	sorridente	tarchiato	1
Parts_of_Speech	3	coniugazione	interiezione	1
	3	coniugazione	numerale	1
Rooms_in_the_House	3	sala_concerti	gabinetto	1
Verbs_Animal_Sounds	3	incornare	fare_le_fusa	1
Verbs_Cognition	3	ascoltare	dimenticare	1
Verbs_Perception	3	capire	percepire	1
	3	capire	guardare	1
Verbs_Touch	3	saziare	premere	1
War	3	litigio	trattato_di_pace	1
Weather_Events	3	ombrello	nuvola	1
Art	4	ricamo	marmo	1
Computer_Components	4	file	scheda_audio	1
	4	file	modem	1
Cooking	4	stomaco	chef	1
Economics	4	pericolo	carta_di_credito	1
Electronics	4	film	altoparlante	1
Flying_Means_of_Transport	4	aquilone	parapendio	2
Human_Features_Positivity	4	pigro	collaborativo	1
Office_Supplies	4	tazza	astuccio	1
Politics	4	competizione	primo_ministro	1
Rooms_in_the_House	4	reception	gabinetto	1
Verbs_Animal_Sounds	4	calpestare	fare_le_fusa	1
Verbs_Cognition	4	udire	dimenticare	1
	4	udire	sapere	1
Verbs_Cooking_2	4	macellare	marinare	1
	4	macellare	mescolare	1
Verbs_Eating	4	cenare	digerire	1
Verbs_Mouth	4	pronunciare	sorridere	1
Verbs_Perception	4	ignorare	percepire	1
Verbs_Psych	4	criticare	preoccupare	1
Astronomical_Objects	5	gravità	satellite	1
Extreme_Natural_Events	5	vulcano	incendio	1
Office_Supplies	5	pennello	astuccio	2
Parts_of_Skeleton	5	frattura	mascella	1
Parts_of_Speech	5	grammatica	interiezione	1
Sweets	5	pasticciere	zucchero_filato	1
Verbs_Psych	5	migliorare	incoraggiare	1
Art	6	scrittura	marmo	1
Bodies_of_Water	6	fango	fiordo	1
	6	fango	oceano	1
Computer_Components	6	gigabyte	stampante	1
Dairy_Products	6	latte_materno	burro	1
Food_Features	6	scaduto	saporito	1
Linguistics	6	proprietà	prefisso	1
Liquid_Containers	6	drink	fiala	1
Music	6	pittore	rock	1
Verbs_Killing	6	seppellire	affogare	1

Verbs_School	6	frequentare	contare	1
War	6	arbitro	trattato_di_pace	2
Textile_Fibres	7	dente	viscosa	1
Verbs_Cooking_1	7	copiare	saltare	1
Art	8	soffitto	marmo	1
Human_Features_Positivity	8	colorato	altruista	1

We recall here that items with the highest positions in the dataset (i.e., the items more related to the semantic category or topic, even though not belonging to it) are also those which are more subject to error (see Section 6.1), as being more challenging for the evaluators to be perceived separately from the cluster of inliers. While for some mistakes it is not clear what could have led the evaluators to choose an inlier instead of the correct outlier, for others we attempt to provide an explanation. We have to keep in mind the length of the task in terms of number of sets included (128), thus one of the main reasons for errors may be distraction/tiredness while performing the task.

First, there are several cases in which the inlier mistaken for outlier is a loan word from English: *kalashnikov* (inlier) – *proiettile* (outlier; Eng. 'projectile') in "Firearms"; *rock* (inlier) – *lettera* (outlier; Eng. 'letter') in "Music"; *hip\_hop* (inlier) – *opera* (outlier; Eng. 'opera') in "Music\_Genres"; *zombie* (inlier) – *bacchetta\_magica* (outlier; Eng. "wand"). A loan word, even though well established in Italian language, may be perceived as less semantically coherent to the other words in the set, thus motivating its selection as an outlier.

Second, there are some cases in which we intentionally followed a strict taxonomic approach to compile the set, but the native speakers' perception on the items in the sets may not have been taxonomy-driven. To explain this point, let us consider the following examples: *carpa* (inlier; Eng. 'carp') – *delfino* (outlier; Eng. 'dolphin'), *balena* (inlier; Eng. 'whale') – *anguilla* (outlier; Eng. 'eel') in "Fish"; *castagna* (inlier; Eng. 'chestnut') – *uvetta* (outlier; Eng. 'raisin') in "Nuts". In all these cases, the items mistaken for outliers, even though belonging to the same taxonomic category as the other inliers (namely, species of fish and types of nuts), may be perceived as less prototypical<sup>13</sup> with respect to the category. Indeed, we believe *carp* and *eel* are less prototypical in the category "Fish" for Italian native speakers, even though taxonomically they have the same status as other items such as *tuna* and *salmon*. Furthermore, we tend not to differentiate the species (in this case, mammal and fish) of animates living in the sea; therefore, even though *dolphin* and *whale* are taxonomically *mammals*, not *fish*, they tend to be perceived together with fish, rather than outliers in the set of words. The same can be said for nuts: *chestnut* is taxonomically a nut, as much as *hazelnut* and *almond*, and *raisin* is a processed fruit, but

<sup>&</sup>lt;sup>13</sup> This has to do with a theory in which membership in a category is gradient, rather than sharp, with some items being more typically associated with the category than others. On this topic, see Jezek (2016: 62, 70).

*chestnut* may have been perceived more as a fruit and *raisin* as a nut. Differently from the "Fish" example, one motivation for this error can be subjective (namely, how this kinds of food are eaten; maybe raisins are eaten more often with other nuts and thus more easily associated to them).<sup>14</sup>

In further implementations of the dataset, it may be worth exploring extensively the notion of prototypicality in the definition of semantic categories for the sets.

Third, a reason for some mistakes may be the specificity of some words we selected as inliers with respect to the outlier which was missed. Even though we tried to include only common-knowledge vocabulary (see Section 3.2), in some cases the number of members of a semantic category we chose was limited, thus we had to include some more specific (in terms of domain-specificity) items. Consider, for example: *parapendio* (inlier, Eng. 'paraglider') – *automobile* (outlier; Eng. 'car') in "Flying\_Means\_of\_Transport"; *macchia\_mediterranea* (inlier; Eng. 'mediterranean\_shrub') – *isola* (outlier; Eng. 'island') in "Biomes"; *fiordo* (inlier; Eng. 'fjord') – *fango* (outlier; Eng. 'mud') in "Bodies\_of\_Water"; *fiala* (inlier; Eng. 'phial') – *drink* (outlier; Eng. 'drink') in "Liquid\_Containers"; *viscosa* (inlier; Eng. 'viscose') – *dente* (outlier; Eng. 'tooth'). We may consider the possibility that the evaluator did not know the meaning or the context of use of these words, thus failing to perceive them as inliers in the sets.

Fourth, after analyzing sets based on semantic categories, let us focus on sets based on *topics*. Out of 11 topics (see Section 3.4), 3 were never mistaken (namely, "Informatics", "Maths", "Medicine", and "Sport"), whereas the other 8 were mistaken from 1 to 3 times (with "Art" being the most mistaken), as we can see from Table 16 (which is structured as the previous Table). Therefore, around 73% of the sets based on topics was mistaken at least by one evaluator. Being the definition of this sets looser, with items *related* rather than *similar* (see Section 3.2), we believe it could have been harder for the evaluators to perceive these groups of words as items related to a specific topic.

<sup>&</sup>lt;sup>14</sup> Consider, as another example (not included in the dataset), the perception of *tomato*, which taxonomically-speaking is a fruit, as a vegetable in Italian, due to its use in salty dishes and together with other proper vegetables.

Table 16. Evaluators' errors on topic-based sets

set name	outlier position	correct outlier	inlier mistaken for outlier	n. of evaluators
Cooking	1	birreria	chef	1
Music	1	lettera	rock	1
War	3	litigio	trattato_di_pace	1
Art	4	ricamo	marmo	1
Cooking	4	stomaco	chef	1
Economics	4	pericolo	carta_di_credito	1
Politics	4	competizione	primo_ministro	1
Art	6	scrittura	marmo	1
Linguistics	6	proprietà	prefisso	1
Music	6	pittore	rock	1
War	6	arbitro	trattato_di_pace	2
Art	8	soffitto	marmo	1

Finally, we can focus on the errors according to the part of speech of the set. What is interesting for us are sets based on verbs and adjectives, at they can be seen as less typical semantic categories and thus more challenging for the evaluators. As we saw in Section 6.1, around 33% of the adjective-based (5 sets out of 12) sets 67% of the verb-based sets were mistaken et least by one evaluator.

In the following Table (Table 17) we report the adjective-based sets with at least one error, and which is the pair of correct outlier – inlier mistaken for outlier for each set. These sets were mistaken from 1 to 4 times. As we have already mentioned at the beginning of this Chapter, one of the reasons for such a small error rate may be the limited number of adjective-based sets. Nevertheless, the ones with mistakes may provide a good insight for further improvements of the dataset.

set name	outlier	correct outlier	inlier mistaken for	n. of
	position		outlier	evaluators
Dimensional_Features_2	1	acuto	conico	1
Hair_Features	1	rugoso	pelato	1
Human_Features_Positivity	1	alto	collaborativo	2
Dimensional_Features_2	3	geometrico	quadrato	1
Hair_Features	3	roseo	spettinato	1
Human_Physical_Features	3	sorridente	tarchiato	1
Human_Features_Positivity	4	pigro	collaborativo	1
Food_Features	6	scaduto	saporito	1
Human_Features_Positivity	8	colorato	altruista	1

Table 17. Evaluators' errors on adjective-based sets

In this Table (Table 18), we report the verb-based sets with at least one error, and which is the pair of correct outlier – inlier mistaken for outlier for each set. Verb-based sets are those with the highest percentage of error, with 20 out of 30 sets with at least one evaluator failing to detect the outlier. For sure, this may demonstrate it is challenging for the evaluators to perceive verbs as semantic clusters, but also that the decisions we made in compiling the dataset may not be satisfactory and further investigation and improvements in this part of the dataset may be needed.

One possible motivation for errors is verb polysemy (verbs tend to be more polysemous than other parts of speech; see Jezek, 2016: 58), even though we expected that displaying the words in sets – and not completely out of context – could have helped evaluators to disambiguate polysemous instances. We highlight some polysemous verbs in Table 18, which may have caused misunderstandings. For example, *investire* means both 'to run over someone' (e.g., while driving a car), to 'financially invest money on something', and to 'assign, appoint a title or decoration to someone for some merit'.

set name	outlier	correct outlier	inlier mistaken for	n. of
	position		outlier	evaluators
Verbs_Dog	1	ruggire	scodinzolare	1
Verbs_Economics	1	rapinare	investire	1
	1	rapinare	indebitare	1
Verbs_Hair	1	stirare	tingere	1
	1	stirare	rasare	1
Verbs_Sport	1	giacere	pattinare	1
Verbs_Cooking_1	2	servire	saltare	1
Verbs_Crime	2	condannare	mentire	1
Verbs_Destroying	2	scoppiare	rovinare	1
	2	scoppiare	sterminare	1
Verbs_Dog	2	squittire	fiutare	1
Verbs_Driving	2	deragliare	guidare	1
Verbs_Eating	2	digiunare	mordere	1
	2	digiunare	deglutire	1
Verbs_Human_Sounds	2	sibilare	ridacchiare	1
	2	sibilare	ruttare	1
Verbs_Mouth	2	espirare	sorridere	1
Verbs_Perception	2	silenziare	percepire	1
Verbs_School	2	interrogare	contare	2
Verbs_Telephone	2	strillare	messaggiare	1
Verbs_Touch	2	annusare	graffiare	1
Verbs_Animal_Sounds	3	incornare	fare_le_fusa	1
Verbs_Cognition	3	ascoltare	dimenticare	1
Verbs_Perception	3	capire	percepire	1
	3	capire	guardare	1
Verbs_Touch	3	saziare	premere	1
Verbs_Animal_Sounds	4	calpestare	fare_le_fusa	1

Table 18. Evaluators' errors on verb-based sets

Verbs_Cognition	4	udire	dimenticare	1
	4	udire	sapere	1
Verbs_Cooking_2	4	macellare	marinare	1
	4	macellare	mescolare	1
Verbs_Eating	4	cenare	digerire	1
Verbs_Mouth	4	pronunciare	sorridere	1
Verbs_Perception	4	ignorare	percepire	1
Verbs_Psych	4	criticare	preoccupare	1
Verbs_Psych	5	migliorare	incoraggiare	1
Verbs_Killing	6	seppellire	affogare	1
Verbs_School	6	frequentare	contare	1
Verbs_Cooking_1	7	copiare	saltare	1

We will recall this qualitative analysis in Section 6.5, comparing some relevant results to those by the distributional models.

## 6.4 Qualitative Analysis: Word Embeddings and Sketch Engine Thesaurus Evaluation

In this Section we proceed with the qualitative analysis of the distributional models' performance. As we can infer from the results in terms of accuracy and OPP, the amount of mistakes is much higher in the models than in the human performance. A detailed analysis is therefore more challenging. We decided to limit this analysis to some general considerations on the performance of each set. To do so, we first calculate the average accuracy and OPP of all the models (Sketch Engine distributional Thesaurus, Word Embeddings with attribute "word", Word Embeddings with attribute "lemma) for all the 7 languages all together.<sup>15</sup> The results are reported in Table 19: in the first column we report the name of the set, in the second its type, in the third the part of speech, in the fourth the average accuracy and OPP (upper part of the Table) to those with the lowest. This Table will be exploited in the comparison between human and model performance, in Section 6.5, in other to verify whether there are common patterns in the kind of sets which

<sup>&</sup>lt;sup>15</sup> The average is calculated as the average of each accuracy score per each model per each language.

were most easily identified or were more challenging both for the models and the evaluators.

set name	type of set	part of speech	average accuracy	average OPP
Trees	semantic category	noun	0.940	0.992
Building_Materials	semantic category	noun	0.893	0.968
Sport	topic	noun	0.887	0.958
Colours	semantic category	adjective	0.881	0.965
External_Body_Parts	semantic category	noun	0.869	0.975
Birds	semantic category	noun	0.869	0.970
Family_Members	semantic category	noun	0.857	0.984
Verbs_Cooking_2	semantic category	verb	0.857	0.971
Languages	semantic category	noun	0.851	0.904
Dimensional_Features_1	semantic category	adjective	0.810	0.954
Spirits	semantic category	noun	0.810	0.952
Gemstones	semantic category	noun	0.798	0.964
Verbs_Cooking_1	semantic category	verb	0.798	0.962
Weather_Conditions	semantic category	adjective	0.798	0.957
Weather_Events	semantic category	noun	0.798	0.954
Human_Features_Negativity	semantic category	adjective	0.780	0.950
Furniture	semantic category	noun	0.780	0.944
Biomes	semantic category	noun	0.768	0.956
Firearms	semantic category	noun	0.768	0.956
Fish	semantic category	noun	0.768	0.949
Car_Components	semantic category	noun	0.750	0.963
Shops	semantic category	noun	0.744	0.954
Musical_Instruments	semantic category	noun	0.726	0.944
Politics	topic	noun	0.726	0.933
Landscape_Features	semantic category	noun	0.708	0.943
Informatics	topic	noun	0.708	0.888
School_Subjects	semantic category	noun	0.702	0.919
Illnesses	semantic category	noun	0.696	0.954
Grain	semantic category	noun	0.696	0.935
Economics	topic	noun	0.696	0.926
Herbs	semantic category	noun	0.685	0.948
Extreme_Natural_Events	semantic category	noun	0.679	0.940
Clothes	semantic category	noun	0.679	0.932
Savanna_Animals	semantic category	noun	0.679	0.906
Music_Genres	semantic category	noun	0.673	0.932
Vegetables	semantic category	noun	0.661	0.938
Internal_Body_Parts	semantic category	noun	0.655	0.916
Farm_Animals	semantic category	noun	0.649	0.911
Verbs_Motion	semantic category	verb	0.649	0.895
Medicine	topic	noun	0.649	0.878
Verbs_Farming	semantic category	verb	0.643	0.932

Table 19. Sets ranked per average accuracy and OPP in all the models and languages

Materials			0.643	0.925
	semantic category	noun		
Weapons	semantic category	noun	0.643	0.919
Chemical_Elements	semantic category	noun	0.643	0.914
Electronics	semantic category	noun	0.637	0.937
Dimensional_Features_2	semantic category	adjective	0.637	0.936
Rooms_in_the_House	semantic category	noun	0.631	0.921
Water_Means_of_Transport	semantic category	noun	0.631	0.894
Linguistics	topic	noun	0.631	0.872
Verbs_Sport	semantic category	verb	0.625	0.895
Metals	semantic category	noun	0.619	0.935
Fruit_Trees	semantic category	noun	0.613	0.938
Dishes_and_Cutlery	semantic category	noun	0.613	0.920
Sports	semantic category	noun	0.613	0.917
Verbs_Plants	semantic category	verb	0.613	0.888
Bugs	semantic category	noun	0.613	0.879
Bodies_of_Water	semantic category	noun	0.607	0.883
Dairy_Products	semantic category	noun	0.607	0.872
Reptiles	semantic category	noun	0.607	0.842
Verbs_Measure	semantic category	verb	0.601	0.915
Parts_of_Skeleton	semantic category	noun	0.601	0.877
Verbs_Communication_1	semantic category	verb	0.601	0.867
Flying_Means_of_Transport	semantic category	noun	0.589	0.893
Music	topic	noun	0.583	0.886
Spices	semantic category	noun	0.577	0.915
Liquid_Containers	semantic category	noun	0.577	0.890
Textile_Fibres	semantic category	noun	0.571	0.908
Shapes	semantic category	noun	0.571	0.901
Verbs_Economics	semantic category	verb	0.565	0.893
 Parts_of_Head	semantic category	noun	0.565	0.889
Art	topic	noun	0.565	0.854
Verbs_Perception	semantic category	verb	0.542	0.866
Wild Animals	semantic category	noun	0.536	0.888
 Human_Moods	semantic category	adjective	0.536	0.885
Containers	semantic category	noun	0.536	0.854
Food	semantic category	noun	0.530	0.905
Means_of_Transport	semantic category	noun	0.530	0.894
Road_Means_of_Transport	semantic category	noun	0.530	0.888
Professions	semantic category	noun	0.530	0.846
Cooking	topic	noun	0.524	0.916
Fantasy_Characters	semantic category	noun	0.524	0.880
Units_of_Time	semantic category	noun	0.518	0.892
Food_Features	semantic category	adjective	0.518	0.860
Maths	topic	noun	0.518	0.851
Computer_Components	semantic category	noun	0.518	0.818
Verbs_Mouth		verb	0.518	0.818
Human_Physical_Features	semantic category	adjective	0.512	0.837
	semantic category	, ř		
Sweets	semantic category	noun	0.506	0.888
Zodiac_Signs	semantic category	noun	0.506	0.884
Sources_of_Energy	semantic category	noun	0.506	0.812

Touch_Features	semantic category	adjective	0.500	0.876
Dances	semantic category	noun	0.500	0.855
Verbs_Weather	semantic category	verb	0.500	0.831
Verbs_Communication_2	semantic category	verb	0.494	0.873
Fruit	semantic category	noun	0.470	0.872
Buildings	semantic category	noun	0.464	0.846
War	topic	noun	0.464	0.832
Temperature_Features	semantic category	adjective	0.464	0.824
Verbs_Religion	semantic category	verb	0.446	0.885
Flowers	semantic category	noun	0.440	0.881
Parts_of_House	semantic category	noun	0.435	0.905
Verbs_Cognition	semantic category	verb	0.435	0.858
Non-alcoholic_Drinks	semantic category	noun	0.429	0.832
Verbs_Eating	semantic category	verb	0.423	0.837
Verbs_Hair	semantic category	verb	0.417	0.844
Free_Time_Activities	semantic category	noun	0.417	0.810
Book_Genres	semantic category	noun	0.411	0.787
Verbs_Driving	semantic category	verb	0.405	0.823
Verbs_Animal_Sounds	semantic category	verb	0.399	0.817
Parts_of_Speech	semantic category	noun	0.393	0.829
Verbs_Music	semantic category	verb	0.393	0.812
Kitchenware	semantic category	noun	0.387	0.823
Verbs_Psych	semantic category	verb	0.381	0.809
Verbs_Killing	semantic category	verb	0.375	0.814
Verbs_Dog	semantic category	verb	0.357	0.775
Verbs_Crime	semantic category	verb	0.351	0.782
Astronomical_Objects	semantic category	noun	0.345	0.804
Office_Supplies	semantic category	noun	0.345	0.801
Verbs_Destroy	semantic category	verb	0.345	0.798
Nuts	semantic category	noun	0.327	0.805
Verbs_Human_Sounds	semantic category	verb	0.327	0.730
Shoes	semantic category	noun	0.321	0.669
Human_Features_Positivity	semantic category	adjective	0.310	0.823
Verbs_Touch	semantic category	verb	0.304	0.782
Verbs_Smell	semantic category	verb	0.286	0.781
Hair_Features	semantic category	adjective	0.280	0.699
Verbs_School	semantic category	verb	0.268	0.788
Verbs_Telephone	semantic category	verb	0.185	0.735
overall average			0.578	0.884

First, we focus on topic-based sets, which are reported in Table 20 (taken from Table 19). Almost all the sets have an average accuracy which is over the general average accuracy (0.578, see Table 19), which makes us assume that even sets containing related (not similar) words tend to be perceived as clusters by the models.

set name	type of set	part of speech	average	average
			accuracy	OPP
Sport	topic	noun	0.887	0.958
Politics	topic	noun	0.726	0.933
Informatics	topic	noun	0.708	0.888
Economics	topic	noun	0.696	0.926
Medicine	topic	noun	0.649	0.878
Linguistics	topic	noun	0.631	0.872
Music	topic	noun	0.583	0.886
Art	topic	noun	0.565	0.854
Cooking	topic	noun	0.524	0.916
Maths	topic	noun	0.518	0.851
War	topic	noun	0.464	0.832

Table 20. Topic-based sets ranked per average accuracy and OPP in all the models and languages

Then, we focus on adjective-based sets (Table 21). Here the results are rather uneven: while some typical clusters, such as "Colours", "Weather\_Conditions" have really high accuracy, some less typical, such as "Hair\_Features", "Touch\_Features" do not give satisfactory results.

Table 21. Adjective-based sets ranked per average accuracy and OPP in all the models and languages

set name	type of set	part of speech	average	average
			accuracy	OPP
Colours	semantic category	adjective	0.881	0.965
Dimensional_Features_1	semantic category	adjective	0.810	0.954
Weather_Conditions	semantic category	adjective	0.798	0.957
Human_Features_Negativity	semantic category	adjective	0.780	0.950
Dimensional_Features_2	semantic category	adjective	0.637	0.936
Human_Moods	semantic category	adjective	0.536	0.885
Food_Features	semantic category	adjective	0.518	0.860
Human_Physical_Features	semantic category	adjective	0.512	0.821
Touch_Features	semantic category	adjective	0.500	0.876
Temperature_Features	semantic category	adjective	0.464	0.824
Human_Features_Positivity	semantic category	adjective	0.310	0.823
Hair_Features	semantic category	adjective	0.280	0.699

Finally, we perform a similar analysis with the verb-based sets, in Table 22. In general, most of verb-sets gained a low performance (below the average, 0.578). This is also the case of some sets based on Levin (1993) verb semantic classes, such as "Verbs\_Perception", "Verbs\_Psych", "Verbs\_Destroy".

set name	type of set	part of speech	average accuracy	average OPP
Verbs_Cooking_2	semantic category	verb	0.857	0.971
Verbs_Cooking_1	semantic category	verb	0.798	0.962
Verbs_Motion	semantic category	verb	0.649	0.895
Verbs_Farming	semantic category	verb	0.643	0.932
Verbs_Sport	semantic category	verb	0.625	0.895
Verbs_Plants	semantic category	verb	0.613	0.888
Verbs_Measure	semantic category	verb	0.601	0.915
Verbs_Communication_1	semantic category	verb	0.601	0.867
Verbs_Economics	semantic category	verb	0.565	0.893
Verbs_Perception	semantic category	verb	0.542	0.866
Verbs_Mouth	semantic category	verb	0.512	0.857
Verbs_Weather	semantic category	verb	0.500	0.831
Verbs_Communication_2	semantic category	verb	0.494	0.873
Verbs_Religion	semantic category	verb	0.446	0.885
Verbs_Cognition	semantic category	verb	0.435	0.858
Verbs_Eating	semantic category	verb	0.423	0.837
Verbs_Hair	semantic category	verb	0.417	0.844
Verbs_Driving	semantic category	verb	0.405	0.823
Verbs_Animal_Sounds	semantic category	verb	0.399	0.817
Verbs_Music	semantic category	verb	0.393	0.812
Verbs_Psych	semantic category	verb	0.381	0.809
Verbs_Killing	semantic category	verb	0.375	0.814
Verbs_Dog	semantic category	verb	0.357	0.775
Verbs_Crime	semantic category	verb	0.351	0.782
Verbs_Destroy	semantic category	verb	0.345	0.798
Verbs_Human_Sounds	semantic category	verb	0.327	0.730
Verbs_Touch	semantic category	verb	0.304	0.782
Verbs_Smell	semantic category	verb	0.286	0.781
Verbs_School	semantic category	verb	0.268	0.788
Verbs_Telephone	semantic category	verb	0.185	0.735

Table 22. Verb-based sets ranked per average accuracy and OPP in all the models and languages

## 6.5 Human vs. Distributional Models' Performance

In this Section we conduct a comparative analysis among the performances between human and distributional models. First, we focus on the results of the quantitative analysis, then we move to the qualitative analysis. As the human evaluation was performed only on Italian language, we mainly compare results from Italian models.

#### 6.5.1 Quantitative analysis

We compare first the raw agreement from the human evaluation and the accuracy from the distributional models. As they can be both defined as the rate of correct answers on the overall number of possible answers, the two metrics are comparable (see Chapter 5).

In the following Table (Table 23), we report the scores. As expected, the human performance has a significantly higher percentage of success, with respect to the various models for each language in the Table. We have to keep in mind that the evaluation was conducted on a limited number of people, and that more people involved may lead to worse results even in human evaluation. There is therefore room for improvement in the distributional models, even though we cannot expect the models to outperform the human evaluation: what is perceived as a set of semantically related or similar words by a human does not necessarily overlap with the actual distribution of these words belonging to a set in the word space. And when the model fails in recognizing the set and detecting the outlier, this does not necessarily mean that the model's quality is bad – simply, it may be that the model does not capture the same salient features as the human evaluator, and vice versa.

language	SkeThe	WE_word	WE_lemma	average
CS (csTenTen19)	0.482	0.687	0.628	0.599
DE (deTenTen20)	0.478	0.657	0.647	0.594
EN (enTenTen20)	0.403	0.618	0.596	0.539
ET (estonian_nc21)	0.562	0.685	0.659	0.635
FR (frTenTen20)	0.400	0.621	0.574	0.531
IT (itTenTen20)	0.419	0.670	0.551	0.547
SK (elexis_skTenTen21)	0.442	0.673	0.680	0.598
IT human benchmark				0.920

Table 23. Overall scores of the models in accuracy, compared to the raw agreement from the human evaluation

We can consider agreement per part of speech, comparing the human performance to the average accuracy of the Italian models (Table 24) – with the average computed on Sketch Engine Thesaurus, Word Embeddings "word" and Word Embeddings "lemma" (column 3); the human raw agreement (column 2) per each part of speech (column 1). Even if the distance of the scores is much more significant in the models, there is a clear pattern in both human evaluation and model evaluation: verb-based sets tend to gain the lowest agreement/accuracy and adjective-based higher.

Table 24. Agreement vs. accuracy per part of speech

part of speech	human agreement	models average accuracy
noun	0.93	0.61
verb	0.90	0.47
adjective	0.94	0.88

#### 6.5.2 Qualitative analysis

We now proceed to compare the metrics per set, in order to verify if there are similarities among the human evaluators and the models. In Table 25 we report the 128 sets of HAMOD dataset (column 1), ranked first per human agreement (column 2) and then per average model accuracy for Italian (column 2).

set name	human agreement	model accuracy
Family_Members	1.00	1.000
Human_Features_Negativity	1.00	1.000
Languages	1.00	1.000
Weather_Conditions	1.00	0.958
Car_Components	1.00	0.917
Verbs_Communication_1	1.00	0.875
Birds	1.00	0.833
Dishes_and_Cutlery	1.00	0.833
Spirits	1.00	0.833
Musical_Instruments	1.00	0.792
Informatics	1.00	0.750
Internal_Body_Parts	1.00	0.750
Sport	1.00	0.750
Verbs_Motion	1.00	0.708
Colours	1.00	0.667
Materials	1.00	0.667
Temperature_Features	1.00	0.667
Weapons	1.00	0.667
Food	1.00	0.583
Professions	1.00	0.583
Bugs	1.00	0.542
Maths	1.00	0.542
Means_of_Transport	1.00	0.500
Verbs_Farming	1.00	0.500
Herbs	1.00	0.458
Shops	1.00	0.458
Verbs_Religion	1.00	0.458
Fruit	1.00	0.417

Spices	1.00	0.417
Human Moods	1.00	0.375
Parts_of_Head	1.00	0.375
Metals	1.00	0.333
Shapes	1.00	0.333
Verbs_Plants	1.00	0.333
Dances	1.00	0.393
Vegetables	1.00	0.292
Chemical_Elements	1.00	0.252
Kitchenware	1.00	0.250
Shoes	1.00	0.250
Zodiac_Signs	1.00	0.208
Trees	0.95	0.750
Dimensional_Features_1	0.95	0.708
Electronics	0.95	0.708
Linguistics	0.95	0.708
Sweets	0.95	0.708
Gemstones	0.95	0.667
Water_Means_of_Transport	0.95	0.667
Clothes	0.95	0.625
Liquid_Containers	0.95	0.625
Road_Means_of_Transport	0.95	0.625
Verbs_Measure	0.95	0.625
Music_Genres	0.95	0.542
Verbs_Driving	0.95	0.542
Units_of_Time	0.95	0.500
Verbs_Crime	0.95	0.500
School_Subjects	0.95	0.417
Non-alcoholic_Drinks	0.95	0.375
Verbs_Smell	0.95	0.375
Flowers	0.95	0.333
Verbs_Killing	0.95	0.292
Verbs_Weather	0.95	0.292
Wild_Animals	0.95	0.292
Grain	0.95	0.250
Free_Time_Activities	0.95	0.208
Touch_Features	0.95	0.208
Dimensional_Features_2	0.91	0.958
Verbs_Cooking_1	0.91	0.917
Firearms	0.91	0.875
Furniture	0.91	0.875
Dairy_Products	0.91	0.792
Parts_of_Skeleton	0.91	0.792
Economics	0.91	0.750
External_Body_Parts	0.91	0.750
Illnesses	0.91	0.750
Extreme_Natural_Events	0.91	0.708
Verbs_Sport	0.91	0.667
Verbs_Economics	0.91	0.625

Weather_Events	0.91	0.625
Book Genres	0.91	0.500
Music	0.91	0.500
Verbs_Animal_Sounds	0.91	0.300
Containers	0.91	0.438
Farm_Animals	0.91	0.417
Food_Features	0.91	0.417
Verbs_Human_Sounds	0.91	0.417
Astronomical_Objects	0.91	0.417
Fantasy_Characters	0.91	0.333
Rooms_in_the_House Fruit_Trees	0.91	0.292
	0.91	0.208
Hair_Features	0.91	0.208
Nuts	0.91	0.208
Textile_Fibres	0.91	0.208
Verbs_Music	0.91	0.167
Parts_of_House	0.87	0.458
Verbs_Cooking_2	0.86	0.958
Landscape_Features	0.86	0.750
Medicine	0.86	0.750
Politics	0.86	0.750
Biomes	0.86	0.667
War	0.86	0.625
Flying_Means_of_Transport	0.86	0.542
Savanna_Animals	0.86	0.500
Verbs_Dog	0.86	0.500
Buildings	0.86	0.458
Human_Physical_Features	0.86	0.458
Verbs_Hair	0.86	0.458
Verbs_Touch	0.86	0.458
Verbs_Destroy	0.86	0.375
Office_Supplies	0.86	0.250
Verbs_Telephone	0.86	0.208
Building_Materials	0.82	0.833
Fish	0.82	0.833
Verbs_Communication_2	0.82	0.792
Reptiles	0.82	0.750
Cooking	0.82	0.708
Verbs_Cognition	0.82	0.667
Sources_of_Energy	0.82	0.583
Art	0.82	0.542
Verbs_Perception	0.82	0.542
Verbs_Mouth	0.82	0.458
Parts_of_Speech	0.82	0.333
Verbs_Eating	0.82	0.333
Bodies_of_Water	0.77	0.667
Verbs_Psych	0.77	0.500
Sports	0.77	0.250
Computer_Components	0.73	0.375
r	0.75	5.575
Verbs_School	0.73	0.333
---------------------------	------	-------
Human_Features_Positivity	0.73	0.167

There are some interesting patterns that emerge from this comparison. First, we highlight in bold those scores which were strikingly divergent (that is, the human agreement is really high and the model accuracy really low).<sup>16</sup> These are the sets affected: "Dances", "Vegetables", "Chemical\_Elements", "Kitchenware", "Shoes", "Zodiac\_Signs", "Verbs\_Killing", "Verbs\_Weather", "Wild\_Animals", "Grain", "Free\_Time\_Activities", "Touch\_Features", "Rooms\_in\_the\_House", "Fruit\_Trees", "Hair\_Features", "Nuts", "Textile\_Fibres", "Verbs\_Music". In this list we can find some verb clusters – which may be harder to be perceived as semantically coherent – but also, surprisingly, well defined semantic categories such as "Vegetables", "Fruit Trees", "Chemical Elements".

Another point is that there are a few cases in which, unexpectedly, human agreement is (even slightly) lower than model accuracy. This is the case of "Dimensional\_Features\_2", "Verbs\_Cooking\_2", "Building\_Materials" and "Fish".

Finally, we perform a latter comparison on the detailed pairs of correct outlier – inlier mistaken as outlier from the human evaluation and the detailed results of the Italian model evaluation. We report those cases in which the human evaluation and the model evaluation coincided in the word pairs. For the sake of simplicity, we simply focus on Word Embedding model with attribute "word", which is the one which gave the best performance. In the following Table (Table 26) we report these results in detail: in the first column, the name of the set; in the second and fourth, the correct outlier; in the third and fifth, the mistaken inlier. Notice that there may be more than one answer as it took more than one attempt to the model to find the outlier. We can see that there are points in common (those highlighted in bold), both as whole pairs, or only as correct outlier or mistaken inlier.

set name	human evaluation – correct outlier	human evaluation – inlier mistaken for outlier	model evaluation – correct outlier	model evaluation – inlier mistaken for outlier
Cooking	birreria	chef	digestione	chef
	stomaco	chef	appetito	chef
External_Body_Parts	rene	fianco	rene	fianco
			articolazione	fianco
			cartilagine	fianco

Table 26. Comparison between the detailed results of the human evaluation and the model evaluation in terms of word pairs "correct outlier – inlier mistaken as outlier"

<sup>&</sup>lt;sup>16</sup> We established the following thresholds: more than 90% agreement and less than 30% accuracy.

Farm_Animals	volpe	gallina	volpe	cavallo
 Flying_Means_of_Transport	aquilone	parapendio	aquilone	jet
	rugoso	pelato	rugoso	crespo, liscio,
	1	Permit	1	riccio, <b>pelato</b>
Human_Features_Positivity	alto	collaborativo	magro	collaborativo,
frumun_i outuros_i ositivity	unto	conuborativo	mugro	forte
	pigro	collaborativo	pigro	collaborativo
	colorato	altruista	colorato	collaborativo,
	colorato	unituistu	colorato	forte
Landscape_Features	città	cascata	strada	cascata
Lundscupe_r cutures	entu	cuscutu	pozzanghera	cascata
	pittore	rock	pittore	nota
Parts_of_House	linoleum	pavimento	linoleum	pilastro, porta,
Tarts_01_110use	moleum	pavimento	moleum	finestra
	piastrella	pilastro	piastrella	pilastro, porta,
	plasticilă	phastru	plasticila	finestra, scale
Reptiles	rana	serpente	ippopotamo	camaleonte,
Kepules	rana	serpente	трророганно	serpente,
				coccodrillo,
				iguana,
				tartaruga,
				alligatore, geco
Sports	jogging	sci_di_fondo	jogging	sci_di_fondo
Sports	Jogging	sci_ui_tonuo	escursionismo	sci_di_fondo
			atleta	sci_di_fondo
Verbs_Dog	ruggire	scodinzolare	ruggire	fiutare, guaire
Veros_Dog	squittire	fiutare	squittire	fiutare, guaire,
	squittire	nutare	squittire	azzannare,
				scodinzolare,
				mordere
			ferire	fiutare
			gridare	fiutare
Verbs_Economics	rapinare	investire	rapinare	addebitare
verbs_leonomies	rapinare	indebitare	Tapillare	addebitate
Verbs_Hair	stirare	tingere	stirare	intrecciare,
verus_riali	Surare	ungere	surare	tagliare,
				spettinare
Book_Genres	film	racconto_breve	film	diario_di_viagg
DOOK_OCIIICS		racconto_preve		io, poesia,
				biografia,
				poliziesco,
				racconto_
				breve
	libro	diario_di	libro	diario_di_viag
		viaggio		<b>gio</b> , poesia,
		viaggio		<b>gio</b> , poesia, biografia,
				poliziesco
Extrama Natural Exants	vento	incendio	vulcano	-
Extreme_Natural_Events	vento vulcano	incendio	vuicano	valanga, <b>incendio</b>
	vuicano	incentito		mcenuio

Free_Time_Activities	calcio	cucina	attrezzo	lettura, <b>cucina</b> ,
Tree_Time_Activities	calcio	cucina	attrezzo	escursionismo,
				pittura
Non-alcoholic_Drinks	vino	frappè	limone	frappè
Non-alcoholic_Dilliks	VIIIO	ITappe	bicchiere	
Nata				frappè
Nuts	uvetta	castagna	uvetta	castagna,
				mandorla,
				nocciola,
				arachide,
				pistacchio,
				anacardi
Savanna_Animals	lepre	giraffa	lepre	leopardo
School_Subjects	ceramica	lingua_	insegnante	storia,
		straniera		lingua_stranie
				ra, chimica
Verbs_Cooking_1	servire	saltare	servire	saltare
	copiare	saltare	apparecchiare	saltare
Verbs_Destroying	scoppiare	rovinare	minacciare	rovinare,
	scoppiare	sterminare		sterminare,
				demolire,
				rompere
Verbs_Driving	deragliare	guidare	deragliare	parcheggiare,
- 6	0	C	0	svoltare,
				accelerare
Verbs_Human_Sounds	sibilare	ridacchiare	sibilare	ruttare,
	sibilare	ruttare	-	cantare,
	5101111			singhiozzare
Verbs_Mouth	espirare	sorridere	annusare	soffiare,
	copilate	501114010		sorridere,
				leccare
Verbs_School	interrogare	contare	interrogare	contare
Verbs_Telephone	strillare	messaggiare	strillare	conversare,
veros_relephone	Sti mui c	messaggiare	Sti mur c	richiamare,
				riattaccare,
				rispondere
Verbs_Touch	annusare	graffiare	annusare	graffiare
verbs_rouch		8	-	0
Watan Maana of Transmort	saziare	premere	annusare	premere
Water_Means_of_Transport	mongolfiera	nave	mongolfiera	canoa
Biomes	isola	macchia_	habitat	macchia_
		mediterranea		mediterranea
Fantasy_Characters	bacchetta_	zombie	bacchetta_	sirena
	magica	· · · · · · · · · · · · · · · · · · ·	magica	
Fruit_Trees	giglio	arancio	faggio	arancio
			giglio	pero
Human_Physical_Features	sorridente	tarchiato	sorridente	basso, alto
Parts_of_Speech	coniugazione	interiezione	coniugazione	nome,
				congiunzione
	coniugazione	numerale	consonante	nome,
				numerale

	grammatica	interiezione	grammatica	nome,
			-	congiunzione
Rooms_in_the_House	sala_concerti	gabinetto	sala_concerti	gabinetto,
				cantina
			reception	gabinetto,
				cantina
Verbs_Animal_Sounds	incornare	fare_le_fusa	ridacchiare	fare_le_fusa
	calpestare	fare_le_fusa	calpestare	fare_le_fusa
Verbs_Perception	capire	percepire	capire	notare,
				percepire
Weather_Events	ombrello	nuvola	ombrello	temporale
Art	ricamo	marmo	ricamo	quadro
Computer_Components	file	scheda_audio	file	disco_rigido,
	file	modem		scheda_madre
	gigabyte	stampante	gigabyte	disco_rigido,
				scheda_madre
Verbs_Eating	cenare	digerire	vomitare	digerire
Verbs_Perception	ignorare	percepire	ignorare	percepire
Verbs_Psych	criticare	preoccupare	criticare	incoraggiare,
				rallegrare
Astronomical_Objects	gravità	satellite	eclissi	satellite
Office_Supplies	pennello	astuccio	pennello	astuccio
Parts_of_Skeleton	frattura	mascella	lussazione	mascella
Liquid_Containers	drink	fiala	birra	fiala
Verbs_Killing	seppellire	affogare	seppellire	affogare
Verbs_School	frequentare	contare	frequentare	contare

#### 6.6 Final Remarks

As anticipated in Section 5.1.2, we briefly recall here the hypotheses we formulated regarding the experiment and the results we expect.

First, as far as the human evaluation is concerned, we expected high agreement between the human evaluators (i.e., < 90%), and this was confirmed (we gained 92% of agreement), as we reported in Section 6.1.

Second, as far as the distributional models' evaluation is concerned, we expected some differences among the two kinds of models. We hypothesised that word embeddings, that is, a predictive kind of model, outperform the distributional thesaurus, a rather count-based distributional model. With really few exceptions, this was the case, as we discussed in Section 6.2.

Moreover, we expected different performances in the various models calculated on different languages: that was the case, but we cannot say that the bigger the corpus on which the models computed is, the better the quality of the models – and therefore their performance in the task. See for example the performances of English (Section 6.2), which has the biggest corpus, but rather bad results. Then, we expected similar results in genetically closer languages (e.g., Czech and Slovak, English and German, French and Italian), but we could not detect any clear pattern not even for this aspect.

Finally, both as far as the human evaluation and the models' evaluation are concerned, we expected some variation in the results according to the position of the outlier: we expected better results in the detection of the items which are farther from the inliers, and worse results in the detection of the items which are closer. This was also confirmed, as we can see in Section 6.1.

Furthermore, we supposed that sets based on adjectives and verbs could be more challenging than those based on nouns, both for the humans and for the models: indeed, we spotted more mistakes in adjective and verb sets (see Section 6.1, 6.3, 6.5).

## Conclusions

In this thesis we implemented a multilingual dataset for the outlier detection task (HAMOD dataset), which we used for an intrinsic evaluation of distributional models.

HAMOD dataset consists of several sets (currently, 128) of semantically related words and corresponding outliers. The dataset was first created in 2019 at Masaryk University (Brno) and we implemented its size and improved its quality. The addition of new sets in the dataset was conducted by exploiting the notion of semantic category (or semantic type, as in T-PAS) and domain, and we used sources such as T-PAS ontology and Wikipedia structures in order to retrieve potential topics and words to store in the sets. The dataset refinement was carried out by testing the difficulty of the words contained in it by performing a small experiment on a group of Czech young students. The dataset translation was a collective step, in which some native speakers of the languages in the set (namely, Czech, German, English, Estonian, French, Italian, Slovak) were involved and coordinated with our supervision to make the new sets multilingual.

We used HAMOD dataset in a preliminary experiment of intrinsic evaluation, which was divided into two phases. First, a human evaluation was performed on a benchmark of 22 Linguistics students from the University of Pavia; the experiment resulted into high agreement between the evaluators, and the insights on the most common disagreements could help us refine the dataset. Then, distributional models computed on the most recent Sketch Engine web corpora were evaluated, with the technical support of Lexical Computing. The models' performance proved to be significantly lower than the human; furthermore, Word Embedding models outperformed Sketch Engine Thesaurus in all the languages evaluated. It is therefore worth investigating further on these results with a view to improve the Sketch Engine Thesaurus.

This experiment gave an insight on the nature of distributional models, and on a specific property, that is, the ability to form semantic clusters (see also Camacho-Collados & Navigli, 2016). Further work needs to be done in intrinsic evaluation techniques.

As far as our project is concerned, we can consider some future developments. First, the dataset size could be increased, but we would rather explore automatic or semi-automatic techniques to retrieve semantic categories and topics, as the manual work we conducted is time-consuming. Moreover, we have to keep in mind that further enlargements may make the human evaluation harder, in terms of time required for the task and potential tiredness of the evaluators. One possible solution could be to partition the dataset in sub-datasets, or select a sample of sets to be evaluated, as representative of the whole dataset. Second, new languages could be included, and translation/adaptation should not be too time demanding. Also, more languages can give further insights in terms of inter-

linguistic variability. Third, human evaluation needs to be carried out more extensively and on the missing languages, involving a higher number of human evaluators. This could be done in less controlled environments, for example by crowdsourcing, although we should find a way to verify the evaluator native language before the task performance. Fourth, other distributional models (e.g., other word embedding models apart from *word2vec*) could be included in the analysis, as most of the relevant studies do. Fifth, before performing new experiments, the dataset could be modified following the feedback from the experiment results, posing a threshold below which a set can be modified or even deleted or substituted in the dataset.

Finally, HAMOD dataset may be useful outside outlier detection task, as it is built on various semantic sources, thus making it a potential object for other analyses or applications.

## **Bibliography**

- Aitchison, Jean, David Bawden & Alan Gilchrist. 2003. *Thesaurus construction and use: a practical manual*. London: Routledge.
- Almeida, Felipe & Geraldo Xexéo. 2019. Word embeddings: A survey. In *arXiv preprint arXiv:1901.09069*.
- Andersen, Jesper Brink, Mikkel Bak Bertelsen, Mikkel Hørby Schou, Manuel R. Ciosici & Ira Assent. 2020. One of these words is not like the other: a reproduction of outlier identification using non-contextual word representations. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*. pp. 120–130.
- Artstein, Ron. 2017. Inter–annotator agreement. In James Pustejovsky & Nancy Ide (eds.), *Handbook of Linguistic Annotation*. Dordrecht: Springer. pp. 297–313.
- Bakarov, Amir. 2018. A survey of word embeddings evaluation methods. In *arXiv preprint arXiv:1801.09536*.
- Baroni, Marco & Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. In *Computational Linguistics*, *36*(*4*). pp. 673-721.
- Baroni, Marco, Georgiana Dinu & Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Baltimore, Maryland, USA. pp. 238-247.
- Ben–Gal, Irad. 2005. Outlier detection. In Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers. Boston, MA: Springer. pp. 131–146.
- Blair, Philip, Yuval Merhav & Joel Barry. 2017. Automated generation of multilingual clusters for the evaluation of distributed representations. In *Workshop Track Proceedings of the 5<sup>th</sup> International Conference on Learning Representations, ICLR 2017*, Toulon, France.
- Boleda, G. 2020. Distributional Semantics and Linguistic Theory. In *Annual Review of Linguistics* 6. pp. 213–234.

- Boukerche, Azzedine, Lining Zheng & Omar Alfandi. 2020. Outlier Detection: Methods, Models, and Classification. In *ACM Computing Surveys (CSUR)* 53(3), article 55. pp. 1–37.
- Budanitsky, Alexander & Graeme Hirst. 2006. Evaluating wordnet-based measures of lexical semantic relatedness. In *Computational linguistics* 32(1). pp. 13-47.
- Camacho–Collados, José & Roberto Navigli. 2016. Find the word that does not belong: A framework for an intrinsic evaluation of word vector representations. In *Proceedings of the 1st Workshop on Evaluating Vector–Space Representations for NLP*. pp. 43–50.
- Clarke, Stella G. Dextre. 2019. The information retrieval thesaurus. In *KO Knowledge Organization 46(6)*. pp. 439-459.
- Claveau, Vincent & Ewa Kijak. 2016. Direct vs. indirect evaluation of distributional thesauri. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. Osaka, Japan. pp. 1837-1848.
- Cruse, D. Alan. 1986. Lexical semantics. Cambridge: Cambridge University Press.
- Cruse, D. Alan. 2000. *Meaning in language: An introduction to semantics and pragmatics*. Oxford: Oxford University Press.
- Faruqui, Manaal, Yulia Tsvetkov, Pushpendre Rastogi & Chris Dyer. 2016. Problems with evaluation of word embeddings using word similarity tasks. *arXiv preprint arXiv:1605.02276*.
- Fellbaum, Christiane (ed.). 1998. *WordNet: An Electronic Lexical Database*. Cambridge (Massachusetts): MIT Press.
- Ferret, Olivier. 2012. Combining bootstrapping and feature selection for improving a distributional thesaurus. In *ECAI*. pp. 336-341.
- Ferret, Olivier. 2017. Turning distributional thesauri into word vectors for synonym extraction and expansion. In *Proceedings of the 8th International Joint Conference on Natural Language Processing*. Taipei, Taiwan. Pp. 273-283.
- Firth, John Rupert. 1957. Modes of meaning. In *Papers in Linguistics 1934–1951*. London: Oxford University Press. pp. 190–215.
- Gamallo, Pablo. 2018. Evaluation of Distributional Models with the Outlier Detection Task. *In Proceedings of the 7th Symposium on Languages, Applications and Technologies (SLATE 2018).* Schloss Dagstuhl–Leibniz–Zentrum fuer Informatik.

- Geeraerts, Dirk. 2003. Meaning and definition. In Piet van Sterkenburg (ed.), *A practical Guide to Lexicography*. Amsterdam: John Benjamins Publishing Company. pp. 83-93.
- Gladkova, Anna & Aleksandr Drozd. 2016. Intrinsic evaluations of word embeddings: What can we do better?. In Proceedings of the 1st Workshop on Evaluating Vector– Space Representations for NLP. pp. 36–42.
- Gorunescu, Florin. 2011. Data Mining: Concepts, Models, and Techniques. Springer Science & Business Media.
- Grefenstette, Gregory. 1994. Explorations in Automatic Thesaurus Discovery. Springer.
- Hadj Taieb, Mohamed Ali, Torsten Zesch & Mohamed Ben Aouicha. 2020. A survey of semantic relatedness evaluation datasets and procedures. In *Artificial Intelligence Review 53*. pp. 4407–4448.
- Hanks, Patrick. 2004. Corpus pattern analysis. In *Euralex Proceedings. Vol. 1*. Lorient: Université de Bretagne-Sud. pp. 87–97.
- Hanks, Patrick. 2013. *Lexical Analysis: Norms and Exploitations*. Cambridge: The MIT Press.
- Harris, Zellig Sabbettai. 1954. Distributional structure. In Word 10(2-3). pp. 146-162.
- Hartmann, Reinhard R. K. 2006. Thesaurus. In Keith Brown (ed.), *Encyclopedia of lan*guage and linguistics, vol. 12. Boston-Oxford: Elsevier. pp. 668-676.
- Herman, Ondřej. 2021. Precomputed Word Embeddings for 15+ languages. In Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2021. pp. 41-46.
- Hill, Felix, Roi Reichart & Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. In *Computational Linguistics*, 41(4). pp. 665– 695.
- Hjelmslev, Louis Trolle. 1961. *Prolegomena to a Theory of Language* (2nd edn., 1st edn. 1943). Madison: University of Wisconsin Press.
- Hüllen, Werner. 2005. A history of Roget's Thesaurus. Oxford: Oxford University Press.
- Jakubíček, Miloš, Emma Romani, Pavel Rychlý & Ondřej Herman. 2021. Development of HAMOD: a High Agreement Multi–lingual Outlier Detection dataset. In Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2021. pp. 177–183.

- Jarmasz, Mario & Stan Szpakowicz. 2012. Roget's Thesaurus: A lexical resource to treasure. arXiv preprint arXiv:1204.0258.
- Jarmasz, Mario. 2003. *Roget's Thesaurus as a Lexical Resource for Natural Language Processing*. Ph.D. Thesis. University of Ottawa, Canada.
- Jezek, Elisabetta, Bernardo Magnini, Anna Feltracco, Alessia Bianchini & Octavian Popescu. 2014. T–PAS: A resource of corpus–derived Types Predicate–Argument Structures for linguistic analysis and semantic processing. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*. pp. 890–895.
- Jezek, Elisabetta. 2016. The Lexicon: An Introduction. Oxford: Oxford University Press.
- Jezek, Elisabetta. 2019. Sweetening Ontologies Cont'd: Aligning bottom-up with topdown ontologies. In Proceedings of the Joint Ontology Workshops 2019. Graz, Austria.
- Jurafsky, Daniel, & James H. Martin. 2021. Vector Semantics and Embeddings. In Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 3<sup>rd</sup> edition. Stanford University.
- Kennedy, Alistair & Stan Szpakowicz. 2008. Evaluating Roget's Thesauri. In Proceedings of ACL-08: HLT, Columbus, Ohio (USA). pp. 416–424.
- Kilgarriff, Adam & Colin Yallop. 2000. What's in a Thesaurus? In *Proceedings of LREC*. pp. 1371-1379.
- Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý & Vít Suchomel. 2014. The Sketch Engine: ten years on. In *Lexicography*, 1(1). pp. 7-36.
- Kolb, Peter. 2009. Experiments on the difference between semantic similarity and relatedness. In Kristiina Jokinen & Eckhard Bick (eds.), NODALIDA 2009 Conference Proceedings. pp. 81-88.
- Lassi, Monica. 2002. Automatic thesaurus construction. In *The GSLT course Linguistic Resources*. pp. 1-10.
- Levin, Beth. 1993. *English verb classes and alternations: A preliminary investigation*. Chicago: University of Chicago Press.
- Lin, Dekang. 1998. Automatic retrieval and clustering of similar words. In *COLING-ACL*. pp. 768–774.

Lyons, John. 1977. Semantics. Cambridge: Cambridge University Press.

- Mikolov, Tomáš, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *arXiv preprint arXiv:1301.3781*.
- Mikolov, Tomáš, Wen-tau Yih & Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*. pp. 746-751.
- Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross & Katherine J. Miller. 1990. Introduction to WordNet: An On-line Lexical Database. In *International Journal of Lexicography 3(4)*. Oxford: Oxford University Press. pp. 235-244.
- Murphy, M. Lynne. 2003. Semantic relations and the lexicon: Antonymy, synonymy and other paradigms. Cambridge: Cambridge University Press.
- Pinar Saygin, Ayse, Ilyas Cicekli & Varol Akman. 2000. Turing test: 50 years later. In *Minds and machines 10(4)*. pp. 463–518.
- Pustejovsky, James & Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning: A guide to corpus–building for applications*. O'Reilly Media, Inch.
- Pustejovsky, James, Catherine Havasi, Jessica Littman, Anna Rumshisky & Marc Verhagen. 2006. Towards a Generative Lexical resource: The Brandeis Semantic Ontology. In *Language Resources and Evaluation Conference, LREC*. Genoa, Italy. pp. 1702– 1705.
- Richards, Jack C. 1976. The role of vocabulary teaching. In *TESOl Quarterly 10(1)*. pp. 77–89.
- Riedl, Martin & Chris Biemann.2013. Scaling to Large<sup>3</sup> data: an effective method to compute Distributional Thesauri. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington. pp. 884–890.
- Roget, Peter Mark. 1852. *Thesaurus of English Words and Phrases Classified and Arranged so to Assist in Literary Composition*. London: London: Longman, Brown, Green, and Longmans.
- Rubenstein, Herbert & John B. Goodenough. 1965. Contextual correlates of synonymy. In *Communications of the ACM 8(10)*. pp. 627-633.

- Rumshisky, Anna, Patrick Hanks, Catherine Havasi & James Pustejovsky. 2006. Constructing a Corpus–based Ontology Using Model Bias. In *Proceedings of FLAIRS Conference*. pp. 327–332.
- Rychlý, Pavel & Adam Kilgarriff. 2007. An efficient algorithm for building a distributional thesaurus (and other Sketch Engine developments). In *Proceedings of the ACL* 2007 Demo and Poster Sessions. Prague, Czechia. pp. 41-44.
- Rychlý, Pavel. 2019. Evaluation of Czech Distributional Thesauri. In *Proceedings of Recent Advances in Slavonic Natural Language Processing RASLAN*. pp. 137-142.
- Sahlgren, Magnus. 2008. The distributional hypothesis. In *Italian Journal of Disability Studies 20*. pp. 33-53.
- Salton, Gerard & Donna Harman. 2003. Information Retrieval. In *Encyclopedia of Computer Science*. John Wiley and Sons Ltd., GBR. pp. 858–863.
- Saussure, Ferdinand de. 1916. Cours de linguistique générale. Paris: Edition Payot.
- Schnabel, Tobias, Igor Labutov, David Mimno & Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 Conference* on Empirical Methods in Natural Language Processing. Lisbon, Portugal: Association for Computational Linguistics. pp. 298–307.
- Torregrossa, François, Robin Allesiardo, Vincent Claveau, Nihel Kooli & Guillaume Gravier. 2021. A survey on training and evaluation of word embeddings. In *International Journal of Data Science and Analytics* 11(2). pp. 85–103.
- Turing, Alan M. 1950. Computing Machinery and Intelligence. In Mind, 59(236). Oxford: Oxford University Press. pp. 433–460.
- Turney, Peter D. & Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. In *Journal of artificial intelligence research*, *37*. pp. 141-188.
- Wang, Bin, Angela Wang, Fenxiao Chen, F., Yuncheng Wang & C. C. Jay Kuo. 2019. Evaluating word embedding models: Methods and experimental results. In APSIPA Transactions on Signal and Information Processing, 8.
- Wang, Yuxuan, Yutai Hou, Wanxiang Che & Ting Liu. From static to dynamic word representations: a survey. In *International Journal of Machine Learning and Cybernetics 11*. pp. 1611-1630.

Yang, Dongqiang & David M. Powers. 2008. Automatic thesaurus construction. In *Proceedings of the thirty-first Australasian conference on Computer Science-Volume 74.* pp. 147-156.

## Websites

- Czech English Dictionary Online: (<u>https://slovnik.seznam.cz/preklad/anglicky</u> (last access: 24/06/2022)
- DeepL Translator: <u>https://www.deepl.com/translator</u> (last access: 24/06/2022)
- English French WordReference: <u>https://www.wordreference.com/enfr/</u> (last access: 24/06/2022)
- English German WordReference: <u>https://www.wordreference.com/ende/</u> (last access: 24/06/2022)
- Estonian English Dictionary Online: <u>https://aare.pri.ee/dictionary.html?switch=en</u> (last access: 24/06/2022)
- Faloppa, Federico. 2011. Thesaurus. In Enciclopedia Treccani dell'Italiano: <u>https://www.treccani.it/enciclopedia/thesaurus\_%28Enciclopedia-dell%27Ita-</u> <u>liano%29/</u> (last access: 24/06/2022)
- GitLab: https://about.gitlab.com/ (last access: 24/06/2022)
- Lexical Computing: <u>https://www.lexicalcomputing.com/lexical-computing/</u> (last access: 24/06/2022)
- Lingea Multilingual Online Dictionary: <u>https://dict.com/</u> (last access: 24/06/2022)
- Merriam Webster Dictionary: <u>https://www.merriam-webster.com/</u> (last access: 24/06/2022)
- Merriam Webster Thesaurus: <u>https://www.merriam-webster.com/thesaurus</u> (last access: 24/06/2022)
- Outlier Detection Online Interface: <u>https://milos.sketchengine.co.uk/outlier\_detection/#</u> (last access: 24/06/2022)
- Outlier Detection Task Documentation (Camacho-Collados & Navigli, 2016): <u>http://lcl.uniroma1.it/outlier-detection</u> (last access: 24/06/2022)

PDEV resource: <u>https://pdev.org.uk/</u> (last access: 24/06/2022)

- Roget's Thesaurus Free Version: <u>https://archive.org/details/Rogets-Thesau-</u> <u>rus/page/n23/mode/2up</u> (last access: 24/06/2022)
- Sketch Engine, Automatic Thesaurus: <u>https://www.sketchengine.eu/blog/automatic-the-</u> <u>saurus-synonyms-for-all-words/</u> (last access: 24/06/2022)
- Sketch Engine, Concordance Documentation: <u>https://www.sketchengine.eu/guide/con-</u> cordance-a-tool-to-search-a-corpus/ (last access: 24/06/2022)
- Sketch Engine, CQL: <u>https://www.sketchengine.eu/documentation/corpus-querying/</u> (last access: 24/06/2022)
- Sketch Engine, Embedding Viewer: <u>https://embeddings.sketchengine.eu/</u>, last access: 24/06/2022)
- Sketch Engine, Estonian National Corpus: <u>https://www.sketchengine.eu/estonian-na-</u> <u>tional-corpus/</u> (last access: 24/06/2022)
- Sketch Engine, Regular Expressions: <u>https://www.sketchengine.eu/guide/regular-expres</u>sions/ (last access: 24/06/2022)
- Sketch Engine, Statistics Used: <u>https://www.sketchengine.eu/documentation/statistics-used-in-sketch-engine/</u> (last access: 24/06/2022)
- Sketch Engine, TenTen Corpora Family: <u>https://www.sketchengine.eu/documenta-tion/tenten-corpora/</u> (last access: 24/06/2022)
- Sketch Engine, TenTen English Corpora: <u>https://www.sketchengine.eu/ententen-english-</u> <u>corpus/</u> (last access: 24/06/2022)
- Sketch Engine, Word Sketch: <u>https://www.sketchengine.eu/guide/word-sketch-colloca-tions-and-word-combinations/</u> (last access: 24/06/2022)

Sketch Engine: <u>https://www.sketchengine.eu/</u> (last access: 24/06/2022)

- T-PAS resource: <u>https://tpas.sketchengine.eu/</u> (last access: 24/06/2022)
- Wikipedia Categorization, Documentation: <u>https://en.wikipedia.org/wiki/Wikipedia:Cat-</u> egorization (last access: 24/06/2022)
- Wikipedia List of Lists of Lists: <u>https://en.wikipedia.org/wiki/List\_of\_lists\_of\_lists</u> (last access: 24/06/2022)

- Wikipedia, Category:Kitchenware: <u>https://en.wikipedia.org/wiki/Category:Kitchenware</u> (last access: 24/06/2022)
- Wikipedia, Kitchenware: <u>https://en.wikipedia.org/wiki/Kitchenware</u> (last access: 24/06/2022)
- Wikipedia: https://en.wikipedia.org/wiki/Main\_Page (last access: 24/06/2022)
- WordNet 3.1 Online: <u>http://wordnetweb.princeton.edu/perl/webwn</u> (last access: 24/06/2022)

WordReference: https://www.wordreference.com/ (last access: 24/06/2022)

# **Appendix 1. HAMOD Dataset**

In this Appendix we report the whole HAMOD dataset, with its 7 languages (CS – Czech, DE – German, EN – English, ET – Estonian, FR – French, IT – Italian, SK – Slovak). Each set has its part of speech specified (*nouns*, *verbs*, *adjectives*), its *label* ("Art", "Verbs\_Sport", "Colours" etc.), the 8 inliers (first 8 elements) and the 8 outliers.

#### Table A.1. HAMOD Dataset

	CS	DE	EN	ET	FR	IT	SK
	nouns	nouns	nouns	nouns	nouns	nouns	nouns
1	Art	Art	Art	Art	Art	Art	Art
	výstava	Ausstellung	exhibition	näitus	exposition	mostra	výstava
	malba	Gemälde	painting	maal	peinture	quadro	mal'ba
	pastel	Buntstift	crayon	värvipliiats	pastel	pastello	pastel
	mozaika	Fresko	fresco	fresko	fresque	affresco	mozaika
	umělec	Künstler	artist	kunstnik	artiste	artista	umelec
	vodovky	Aquarell	watercolours	akvarell	aquarelle	acquerello	akvarel
	portrét	Porträt	portrait	portree	portrait	ritratto	portrét
	mramor	Marmor	marble	marmor	marbre	marmo	mramor
	malíř_pokojů	Dekorateur	house_painter	maaler	peintre_en_bâtiment	imbianchino	maliar_izieb
	lžíce	Löffel	spoon	lusikas	cuillère	cucchiaio	lyžica
	hudebník	Musiker	musician	muusik	musicien	musicista	hudobník
	výšivka	Stickerei	embroidery	tikand	broderie	ricamo	výšivka
	bajka	Märchen	tale	lugu	fable	favola	bájka
	psaní	Schreiben	writing	kirjutamine	écriture	scrittura	písanie
	ohřívač	Heizung	heater	küttekeha	chauffage	calorifero	ohrievač
	strop	Decke	ceiling	lagi	plafond	soffitto	strop
	nouns	nouns	nouns	nouns	nouns	nouns	nouns
2	Astronomical_	Astronomical_	Astronomical_	Astronomical_	Astronomical_	Astronomical_	Astronomical_
	Objects	Objects	Objects	Objects	Objects	Objects	Objects
	hvězda	Stern	star	täht	étoile	stella	hviezda
	planeta	Planet	planet	planeet	planète	pianeta	planéta
	černá_díra	schwarzes_Loch	black_hole	must_auk	trou_noir	buco_nero	čierna_diera
	satelit	Satellit	satellite	satelliit	satellite	satellite	satelit
	galaxie	Galaxie	galaxy	galaktika	galaxie	galassia	galaxia
	asteroid	Asteroid	asteroid	asteroid	astéroïde	asteroide	asteroid
	meteorit	Meteorit	meteorite	meteoriit	météorite	meteorite	meteorit
	kometa	Komet	comet	komeet	comète	cometa	kométa

	oběžná_dráha	Orbit	orbit	orbiit	orbite	orbita	obežná_dráha
	zatmění	Mondfinsternis	eclipse	varjutus	éclipse	eclissi	zatmenie
	astronaut	Astronaut	astronaut	astronaut	astronaute	astronauta	astronaut
	teleskop	Teleskop	telescope	teleskoop	télescope	telescopio	teleskop
	gravitace	Schwerkraft	gravity	gravitatsioon	gravitation	gravità	gravitácia
	světelný_rok	Lichtjahr	light_year	valgusaasta	annéelumière	anno_luce	svetelný_rok
	kašna	Brunnen	fountain	purskkaev	fontaine	fontana	fontána
	mír	Frieden	peace	rahu	paix	pace	mier
	nouns	nouns	nouns	nouns	nouns	nouns	nouns
3	Biomes	Biomes	Biomes	Biomes	Biomes	Biomes	Biomes
	deštný_prales	Regenwald	rainforest	vihmamets	forêt_tropicale	foresta_pluviale	dažďový_prales
	savana	Savanne	savanna	savann	savane	savana	savana
	džungle	Dschungel	jungle	džungel	jungle	giungla	džungľa
	step	Steppe	steppe	stepp	steppe	steppa	step
	poušť	Wüste	desert	kõrb	désert	deserto	púšť
	prérie	Prärie	prairie	preeria	prairie	macchia_mediterranea	préria
	tundra	Tundra	tundra	tundra	toundra	tundra	tundra
	tajga	Taiga	boreal_forest	taiga	forêt_boréale	foresta_boreale	tajga
	ekosystém	Ökosystem	ecosystem	ökosüsteem	écosystème	ecosistema	ekosystém
	životní prostředí	Habitat	habitat	elupaik	habitat	habitat	prostredie
	ostrov	Insel	island	saar	île	isola	ostrov
	keř	Strauch	shrub	põõsastik	arbuste	arbusto	ker
	zeměpisná_šířka	Breitengrad	latitude	laiuskraad	latitude	latitudine	zemepisná_šírka
	zeměpisná_délka	Längengrad	longitude	pikkuskraad	longitude	longitudine	zemepisná_dĺžka
	vztek	Wut	anger	viha	colère	rabbia	hnev
	polévka	Suppe	soup	supp	soupe	zuppa	polievka
	nouns	nouns	nouns	nouns	nouns	nouns	nouns
4	Birds	Birds	Birds	Birds	Birds	Birds	Birds
	labuť	Schwan	swan	luik	cygne	cigno	labuť
	kachna	Ente	duck	part	canard	anatra	kačka

racek	Möwe	seagull	kajakas	mouette	gabbiano	čajka
orel	Adler	eagle	kotkas	aigle	aquila	orol
holub	Taube	dove	tuvi	colombe	colomba	holubica
vrána	Rabe	crow	vares	corbeau	corvo	vrana
čáp	Storch	stork	kurg	cigogne	cicogna	bocian
husa	Gans	goose	hani	oie	oca	hus
opice	Affe	monkey	ahv	singe	scimmia	opica
losos	Lachs	salmon	lõhe	saumon	salmone	losos
kobylka	Heuschrecke	grasshopper	ritsikas	sauterelle	cavalletta	kobylka
moucha	Fliege	fly	kärbes	mouche	mosca	mucha
vejce	Ei	egg	muna	œuf	uovo	vajce
letadlo	Flugzeug	plane	lennuk	avion	aeroplano	lietadlo
žena	Frau	woman	naine	femme	donna	žena
útes	Riff	cliff	kalju	falaise	scogliera	útes
nouns	nouns	nouns	nouns	nouns	nouns	nouns
5 Bodies_of_Water	Bodies_of_Water	Bodies_of_Water	Bodies_of_Water	Bodies_of_Water	Bodies_of_Water	Bodies_of_Water
jezero	See	lake	järv	lac	lago	jazero
záliv	Bucht	bay	laht	baie	baia	zátoka
bažina	Sumpf	swamp	soo	marais	palude	močiar
potok	Bach	brook	oja	ruisseau	ruscello	potok
fjord	Fjord	fjord	fjord	fjord	fiordo	fjord
oceán	Özean	ocean	ookean	océan	oceano	oceán
moře	Meer	sea	meri	mer	mare	more
řeka	Fluss	river	jõgi	rivière	fiume	rieka
hora	Berg	mountain	mägi	montagne	montagna	hora
údolí	Tal	valley	org	vallée	valle	údolie
bazén	Schwimmbad	swimming_pool	bassein	piscine	piscina	bazén
				baignoire	vasca_da_bagno	vaňa
vana	Badewanne	bathtub	vann	Daighone	vasca_ua_bagno	
	Badewanne Wasser	bathtub water	vann vesi	eau		voda
vana				-	acqua fango	

	tučňák	Pinguin	penguin	pingviin	pingouin	pinguino	tučniak
	nouns	nouns	nouns	nouns	nouns	nouns	nouns
6	Book_Genres	Book_Genres	Book_Genres	Book_Genres	Book_Genres	Book_Genres	Book_Genres
	horor	Horror	horror	õudusromaan	horreur	horror	horor
	cestopis	Reisebericht	travelogue	reisikiri	carnet_de_voyage	diario_di_viaggio	cestopis
	povídka	Kurzgeschichte	short_story	lühijutt	nouvelle	racconto_breve	poviedka
	poezie	Poesie	poetry	luule	poésie	poesia	poézia
	fantasy	Fantasy	fantasy	armastusromaan	fantasy	fantasy	fantasy
	scifi	ScienceFiction	scifi	ulmekirjandus	sciencefiction	fantascienza	scifi
	životopis	Biographie	biography	elulugu	biographie	biografia	životopis
	detektivka	Krimi	detective_fiction	detektiivromaan	roman_policier	poliziesco	detektívka
	seriál	Serie	serial	seriaal	série	serie	seriál
	film	Film	film	film	film	film	film
	kniha	Buch	book	raamat	livre	libro	kniha
	přirovnání	Vergleich	simile	võrdlus	comparaison	similitudine	prirovnanie
	čtenář	Leser	reader	lugeja	lecteur	lettore	čitateľ
	spisovatel	Schriftsteller	writer	kirjanik	écrivain	scrittore	spisovateľ
	bagr	Bagger	digger	ekskavaator	pelleteuse	escavatrice	bager
	ubrousek	Serviette	napkin	salvrätik	serviette	tovagliolo	servítka
	nouns	nouns	nouns	nouns	nouns	nouns	nouns
	Bugs	Bugs	Bugs	Bugs	Bugs	Bugs	Bugs
	šváb	Schabe	roach	tarakan	blatte	scarafaggio	šváb
	komár	Mücke	mosquito	sääsk	moustique	zanzara	komár
	vosa	Wespe	wasp	herilane	guêpe	vespa	osa
	včela	Biene	bee	mesilane	abeille	ape	včela
	mravenec	Ameise	ant	sipelgas	fourmi	formica	mravec
	motýl	Schmetterling	butterfly	liblikas	papillon	farfalla	motýľ
	beruška	Marienkäfer	ladybug	lepatriinu	coccinelle	coccinella	lienka
	pavouk	Spinne	spider	ämblik	araignée	ragno	pavúk
	pavučina	Spinnennetz	spider_web	ämblikuvõrk	toile_d'_araignée	ragnatela	pavučina

	žihadlo	Stachel	stinger	astel	dard	pungiglione	žihadlo
	žába	Frosch	frog	konn	grenouille	rana	žaba
	zmije	Viper	viper	rästik	vipère	vipera	vretenica
	myš	Maus	mouse	hiir	souris	topo	myš
	veverka	Eichhörnchen	squirrel	orav	écureuil	scoiattolo	veverica
	pirát	Pirat	pirate	piraat	pirate	pirata	pirát
	smutek	Trauer	sadness	nukrus	tristesse	tristezza	smútok
	nouns						
8	Building_Materials						
	beton	Beton	concrete	betoon	béton_armé	cemento_armato	betón
	dřevo	Holz	wood	puit	bois	legno	drevo
	omítka	Gips	plaster	kips	plâtre	gesso	omietka
	jíl	Lehm	clay	savi	argile	laterizio	hlina
	mramor	Marmor	marble	marmor	marbre	marmo	mramor
	kámen	Stein	stone	kivi	pierre	pietra	kameň
	sklo	Glas	glass	klaas	verre	vetro	sklo
	cement	Zement	cement	tsement	ciment	calcestruzzo	cement
	střešní_taška	Dachziegel	roof_tile	katusekivi	tuile	tegola	škridla
	lešení	Gerüst	scaffolding	tellingud	échafaudage	ponteggio	lešenie
	buldozer	Bulldozer	bulldozer	buldooser	bulldozer	ruspa	buldozér
	jeřáb	Kran	crane	kraana	grue	gru	žeriav
	odolnost	Robustheit	robustness	robustsus	solidité	solidità	pevnosť
	tvrdost	Härte	hardness	kõvadus	dureté	durezza	tvrdosť
	křivka	Kurve	bend	painutus	courbe	curva	krivka
	přítel	Freund	friend	sõber	ami	amico	priateľ
	nouns						
9	Buildings						
	nemocnice	Krankenhaus	hospital	haigla	hôpital	ospedale	nemocnice
	továrna	Fabrik	factory	tehas	usine	fabbrica	továreň
	divadlo	Theater	theatre	teater	théâtre	teatro	divadlo
	kino	Kino	cinema	kino	cinéma	cinema	kino

	nákupní_centrum	Einkaufszentrum	shopping_mall	kaubanduskeskus	centre_commercial	centro_commerciale	nákupné_centrum
	stanice	Bahnhof	station	jaam	gare	stazione	stanica
	knihovna	Bibliothek	library	raamatukogu	bibliothèque	biblioteca	knižnica
	hotel	Hotel	hotel	hotell	hôtel	hotel	hotel
	mrakodrap	Wolkenkratzer	skyscraper	pilvelõhkuja	gratteciel	grattacielo	mrakodrap
	vila	Stadthaus	townhouse	ridaelamu	villa	villa	vila
	alej	Allee	avenue	puiestee	avenue	viale	aleja
	náměstí	Platz	square	väljak	place	piazza	námestie
	vstup	Eingang	entrance	sissepääs	entrée	ingresso	vstup
	balkon	Terrasse	terrace	terrass	terrasse	balcone	balkón
	zápěstí	Handgelenk	wrist	ranne	poignet	polso	zápästie
	stan	Zelt	tent	telk	tente	tenda	stan
	nouns	nouns	nouns	nouns	nouns	nouns	nouns
0	Car_Components	Car_Components	Car_Components	Car_Components	Car_Components	Car_Components	Car_Components
v	stěrač	Scheibenwischer	windscreen wiper	klaasipuhasti	essuieglace	tergicristallo	stierač
	čelní sklo	Windschutzscheibe	windshield	tuuleklaas	parebrise	parabrezza	čelné sklo
	blatník	Schutzblech	mudguard	poritiib	gardeboue	parafango	blatník
			0	1	•		
		Airbag	airbag	turvapadi	airbag	airbag	airbag
	airbag volant	Airbag Lenkrad	airbag steering wheel	turvapadi rooliratas	airbag volant	airbag volante	airbag volant
	airbag volant	÷.	steering_wheel	turvapadi rooliratas esituli	volant	airbag volante fanale	volant
	airbag	Lenkrad Scheinwerfer	<u> </u>	rooliratas esituli		volante	-
	airbag volant přední_světlo	Lenkrad	steering_wheel headlight	rooliratas	volant phare	volante fanale	volant predné_svetlo
	airbag volant přední_světlo klakson výfuk	Lenkrad Scheinwerfer Hupe Auspuff	steering_wheel headlight horn exhaust_pipe	rooliratas esituli autopasun väljalasketoru	volant phare klaxon échappement	volante fanale clacson marmitta	volant predné_svetlo klaksón výfuk
	airbag volant přední_světlo klakson výfuk auto	Lenkrad Scheinwerfer Hupe Auspuff Auto	steering_wheel headlight horn exhaust_pipe car	rooliratas esituli autopasun väljalasketoru auto	volant phare klaxon échappement voiture	volante fanale clacson marmitta automobile	volant predné_svetlo klaksón výfuk auto
	airbag volant přední_světlo klakson výfuk auto kamion	Lenkrad Scheinwerfer Hupe Auspuff Auto Lastwagen	steering_wheel headlight horn exhaust_pipe car lorry	rooliratas esituli autopasun väljalasketoru auto veoauto	volant phare klaxon échappement voiture camion	volante fanale clacson marmitta	volant predné_svetlo klaksón výfuk auto nákladné_vozidlo
	airbag volant přední_světlo klakson výfuk auto kamion řidičský_průkaz	Lenkrad Scheinwerfer Hupe Auspuff Auto Lastwagen Führerschein	steering_wheel headlight horn exhaust_pipe car	rooliratas esituli autopasun väljalasketoru auto veoauto juhiluba	volant phare klaxon échappement voiture	volante fanale clacson marmitta automobile	volant predné_svetlo klaksón výfuk auto nákladné_vozidlo vodičský_preukaz
	airbag volant přední_světlo klakson výfuk auto kamion	Lenkrad Scheinwerfer Hupe Auspuff Auto Lastwagen Führerschein Fahrgast	steering_wheel headlight horn exhaust_pipe car lorry driving_licence passenger	rooliratas esituli autopasun väljalasketoru auto veoauto juhiluba sõitja	volant phare klaxon échappement voiture camion permis_de_conduire passager	volante fanale clacson marmitta automobile camion patente passeggero	volant predné_svetlo klaksón výfuk auto nákladné_vozidlo vodičský_preukaz cestujúci
	airbag volant přední_světlo klakson výfuk auto kamion řidičský_průkaz cestující provoz	Lenkrad Scheinwerfer Hupe Auspuff Auto Lastwagen Führerschein Fahrgast Verkehr	steering_wheel headlight horn exhaust_pipe car lorry driving_licence passenger traffic	rooliratas esituli autopasun väljalasketoru auto veoauto juhiluba sõitja liiklus	volant phare klaxon échappement voiture camion permis_de_conduire passager trafic	volante fanale clacson marmitta automobile camion patente passeggero traffico	volant predné_svetlo klaksón výfuk auto nákladné_vozidlo vodičský_preukaz cestujúci premávka
	airbag volant přední_světlo klakson výfuk auto kamion řidičský_průkaz cestující provoz semafor	Lenkrad Scheinwerfer Hupe Auspuff Auto Lastwagen Führerschein Fahrgast Verkehr Ampel	steering_wheel headlight horn exhaust_pipe car lorry driving_licence passenger traffic traffic_light	rooliratas esituli autopasun väljalasketoru auto veoauto juhiluba sõitja liiklus valgusfoor	volant phare klaxon échappement voiture camion permis_de_conduire passager trafic feu_tricolore	volante fanale clacson marmitta automobile camion patente passeggero traffico semaforo	volant predné_svetlo klaksón výfuk auto nákladné_vozidlo vodičský_preukaz cestujúci premávka semafor
	airbag volant přední_světlo klakson výfuk auto kamion řidičský_průkaz cestující provoz	Lenkrad Scheinwerfer Hupe Auspuff Auto Lastwagen Führerschein Fahrgast Verkehr	steering_wheel headlight horn exhaust_pipe car lorry driving_licence passenger traffic	rooliratas esituli autopasun väljalasketoru auto veoauto juhiluba sõitja liiklus	volant phare klaxon échappement voiture camion permis_de_conduire passager trafic	volante fanale clacson marmitta automobile camion patente passeggero traffico	volant predné_svetlo klaksón výfuk auto nákladné_vozidlo vodičský_preukaz cestujúci premávka

	nouns	nouns	nouns	nouns	nouns	nouns	nouns
11	Chemical_Elements	Chemical_Elements	Chemical_Elements	Chemical_Elements	Chemical_Elements	Chemical_Elements	Chemical_Elements
	vodík	Wasserstoff	hydrogen	vesinik	hydrogène	idrogeno	vodík
	helium	Helium	helium	heelium	hélium	elio	hélium
	kyslík	Sauerstoff	oxygen	hapnik	oxygène	ossigeno	kyslík
	uhlík	Kohlenstoff	carbon	süsinik	carbone	carbonio	uhlík
	sodík	Natrium	sodium	naatrium	sodium	sodio	sodík
	vápník	Kalzium	calcium	kaltsium	calcium	calcio	vápnik
	síra	Schwefel	sulfur	väävel	soufre	zolfo	síra
	jód	Jod	iodine	jood	iode	iodio	jód
	molekula	Molekül	molecule	molekul	molécule	molecola	molekula
	atom	Atom	atom	aatom	atome	atomo	atóm
	chemická_reakce	chemische_Reaktion	chemical_reaction	keemiline_reaktsioon	réaction_chimique	reazione_chimica	chemická_reakcia
	oxidace	Oxidation	oxidation	oksüdatsioon	oxydation	ossidazione	oxidácia
	zkumavka	Reagenzglas	phial	katseklaas	éprouvette	provetta	skúmavka
	chemik	Chemiker	chemist	keemik	chimiste	chimico	chemik
	kostel	Kirche	church	kirik	église	chiesa	kostol
	zrcadlo	Spiegel	mirror	peegel	miroir	specchio	zrkadlo
	nouns	nouns	nouns	nouns	nouns	nouns	nouns
12	Clothes	Clothes	Clothes	Clothes	Clothes	Clothes	Clothes
	tričko	TShirt	tshirt	tsärk	tshirt	maglietta	tričko
	šaty	Kleid	dress	kleit	robe	vestito	šaty
	kalhoty	Hose	trousers	püksid	pantalon	pantaloni	nohavice
	kraťasy	Shorts	shorts	lühikesed_püksid	short	pantaloncini	kraťasy
	svetr	Pullover	jumper	džemper	pull	maglione	sveter
	sukně	Rock	skirt	seelik	jupe	gonna	sukňa
	košile	Hemd	shirt	särk	chemise	camicia	košeľa
	kabát	Mantel	coat	mantel	manteau	cappotto	kabát
	prostěradlo	Bettwäsche	sheet	lina	drap	lenzuolo	plachta
	deka	Decke	blanket	tekk	couverture	coperta	deka
	brýle	Brille	glasses	prillid	lunettes	occhiali	okuliare

	sponka	Stirnband	hair_clip	juukseklamber	barrette	fermacapelli	sponka
	bavlna	Baumwolle	cotton	puuvill	coton	cotone	bavlna
	vlna	Wolle	wool	vill	laine	lana	vlna
	klíčenka	Schlüsselanhänger	keychain	võtmehoidja	porteclés	portachiavi	kľúčenka
	hranice	Grenze	border	piir	frontière	confine	hranica
	adjectives						
13	Colours						
10	červený	rot	red	punane	rouge	rosso	červený
	modrý	blau	blue	sinine	bleu	blu	modrý
	zelený	grün	green	roheline	vert	verde	zelený
	žlutý	gelb	yellow	kollane	jaune	giallo	žltý
	fialový	violett	purple	lilla	violet	viola	fialový
	růžový	rosa	pink	roosa	rose	rosa	ružový
	oranžový	orange	orange	oranž	orange	arancione	oranžový
	hnědý	braun	brown	pruun	brun	marrone	hnedý
				F			
	temný	dunkel	dark	tume	sombre	scuro	tmavý
	jasný	hell	bright	hele	clair	chiaro	jasný
	dřevěný	hölzern	wooden	värviline	de_bois	di_legno	drevený
	skleněný	gläsern	glass	mustvalge	en_verre	di_vetro	sklenený
	pruhovaný	gestreift	striped	triibuline	rayé	a_righe	pruhovaný
	puntíkovaný	gefleckt	dotted	täpiline	à_pois	a_pois	bodkovaný
	smutný	traurig	sad	kurb	triste	triste	smutný
	nízký	niedrig	low	madal	faible	basso	nízky
	nouns						
14	Computer_ Components						
	monitor	Bildschirm	screen	ekraan	écran	monitor	monitor
	myš	Maus	mouse	hiir	souris	mouse	myš
	kabel	Modem	modem	modem	modem	modem	kábel
	tiskárna	Drucker	printer	printer	imprimante	stampante	tlačiareň
	klávesnice	Tastatur	keyboard	klaviatuur	clavier	tastiera	klávesnica

	i	İ	nouns	nouns	nouns	nouns	nouns
	40145	risentuen		lautina	nappe	tovagna	Solus
	ubrus	Tischtuch	tablecloth	laudlina	nappe	tovaglia	obrus
	život	Leben	life	elu	vie	consegna vita	život
	doručení	Lieferung	delivery	saadetis	livraison		darcek
	dárek	Geschenk	strawgift	kingitus	paille cadeau	paglia regalo	slama darček
	lepenka sláma	Stroh		papp kõrs	carton	cartone	
		Karton	cup cardboard	tass	tasse		kartón
	skříň šálek	Kleiderschrank Tasse	wardrobe	garderoob	armoire	armadio tazzina	skriňa šálka
	1 ~/~	771 1 1 1	1 1	1 1		1'	1 **
	koš	Korb	basket	korv	panier	cesto	kôš
	dóza	Einmachglas	jar	purk	pot .	barattolo	debna
	zásuvka	Schublade	drawer	sahtel	tiroir	cassetto	zásuvka
	váza	Vase	vase	vaas	vase	vaso	váza
	bedna	Truhe	chest_of_drawers	kummut	coffre	baule	truhlica
	pytel	Sack	sack	karp	sac	sacco	vrece
	taška	Beutel	bag	kott	sachet	sacchetto	taška
	krabice	Schachtel	box	kast	boîte	scatola	krabica
5	Containers	Containers	Containers	Containers	Containers	Containers	Containers
	nouns	nouns	nouns	nouns	nouns	nouns	nouns
	daněk	Rehe	deer	hirv	daim	daino	vysoká
	prach	Staub	dust	tolm	poussière	polvere	prach
	gigabyte	Gigabyte	gigabyte	gigabyte	gigaoctet	gigabyte	gigabyte
	program	Programm	program	programm	programme	programma	program
	soubor	Datei	file	fail	fichier	file	súbor
	internet	Internet	internet	internet	internet	internet	internet
	dálkové_ovládání	Fernbedienung	remote_control	kaugjuhtimispult	télécommande	telecomando	diaľkové_ovládan
	telefon	Telefon	telephone	telefon	téléphone	telefono	telefón
	zvuková karta	Soundkarte	sound_card	helikaart	carte_son	scheda audio	zvuková karta
	pevný_disk základní deska	Festplatte Hauptplatine	hard_disk motherboard	kõvaketas emaplaat	disque_dur carte_mère	disco_rigido scheda madre	hard_disk základná doska

16	Cooking	Cooking	Cooking	Cooking	Cooking	Cooking	Cooking
	hrnec	Kasserolle	pot	pott	casserole	pentola	hrniec
	mouka	Mehl	flour	jahu	farine	farina	múka
	naběračka	Schöpfkelle	ladle	kulp	louche	mestolo	naberačka
	šéfkuchař	Chefkoch	chef	kokk	chef	chef	šéfkuchár
	trouba	Backofen	oven	ahi	four	forno	rúra
	olej	Öl	oil	õli	huile	olio	olej
	ingredience	Zutat	ingredient	koostisaine	ingrédient	ingrediente	ingrediencia
	česnek	Knoblauch	garlic	küüslauk	ail	aglio	cesnak
	pivnice	Kneipe	pub	pubi	brasserie	birreria	krčma
	stůl	Tisch	table	laud	table	tavolo	stôl
	snídaně	Frühstück	breakfast	hommikusöök	petit_déjeuner	colazione	raňajky
	žaludek	Magen	stomach	kõht	estomac	stomaco	žalúdok
	trávení	Verdauung	digestion	seedimine	digestion	digestione	trávenie
	chuť_k_jídlu	Appetit	appetite	söögiisu	appétit	appetito	chuť_do_jedla
	pupek	Nabel	navel	naba	nombril	ombelico	pupok
	zločin	Verbrechen	crime	kuritegevus	crime	reato	zločin
	nouns	nouns	nouns	nouns	nouns	nouns	nouns
17	Dairy_Products	Dairy_Products	Dairy_Products	Dairy_Products	Dairy_Products	Dairy_Products	Dairy_Products
	sýr	Käse	cheese	juust	fromage	formaggio	syr
	syrovátka	Quark	custard	hapukoor	crème_fraîche	ricotta	srvátka
	máslo	Butter	butter	või	beurre	burro	maslo
	jogurt	Joghurt	yogurt	jogurt	yaourt	yogurt	jogurt
	smetana	Sahne	cream	koor	crème	panna	smotana
	mléko	Milch	milk	piim	lait	latte	mlieko
	tvaroh	Kondensmilch	condensed_milk	kondenspiim	crème_chantilly	latte_condensato	tvaroh
	zmrzlina	Eis	ice_cream	jäätis	glace	gelato	zmrzlina
	sójové_mléko	Sojamilch	soy_milk	sojapiim	lait_de_soja	latte_di_soia	sójové_mlieko
	arašídové_máslo	Erdnussbutter	peanut_butter	maapähklivõi	beurre_de_cacahuète	burro_di_arachidi	arašidové_maslo
	kráva	Kuh	cow	lehm	vache	mucca	krava
	dojení	Melken	milking	lüpsmine	traite	mungitura	dojenie

	kyselina_mléčná	Milchsäure	lactic_acid	piimhape	acide_lactique	acido_lattico	kyselina_mliečna
	mateřské_mléko	Muttermilch	breast_milk	rinnapiim	lait_maternel	latte_materno	materské_mlieko
	věda	Wissenschaft	science	teadus	science	scienza	veda
	netopýr	Fledermaus	bat	nahkhiir	chauvesouris	pipistrello	netopier
	nouns						
18	Dances						
	tango						
	breakdance	Breakdance	breakdance	polka	breakdance	breakdance	breakdance
	břišní_tanec	Polka	tap_dance	rahvatants	claquettes	tip_tap	brušný_tanec
	mazurka	Salsa	salsa	salsa	salsa	salsa	salsa
	valčík	Walzer	waltz	valss	valse	valzer	valčík
	polka	Samba	samba	samba	samba	samba	polka
	balet	Ballett	classical_dance	ballett	danse_classique	danza_classica	balet
	společenský_tanec	Gesellschaftstanz	ballroom_dance	seltskonnatants	danse_de_salon	ballo_liscio	spoločenský_tanec
	rap	Rap	rap	räpp	rap	rap	rap
	soul						
	krok	Schritt	step	samm	pas	passo	krok
	muzikál	Musical	musical	muusikal	comédie_musicale	musical	muzikál
	divadlo	Theater	theatre	teater	théâtre	teatro	divadlo
	tanečník	Tänzer	dancer	tantsija	danseur	ballerino	tanečník
	truhlář	Tischler	woodworker	puusepp	menuisier	falegname	tesár
	úsměv	Lächeln	smile	naeratus	sourire	sorriso	úsmev
	adjectives						
19	Dimensional_ Features_1						
	tlustý	dick	thick	paks	épais	spesso	hrubý
	široký	breit	wide	lai	large	largo	široký
	krátký	kurz	short	lühike	court	corto	krátky
	dlouhý	lang	long	pikk	long	lungo	dlhý
	úzký	eng	tight	kitsas	étroit	stretto	úzky
	malý	klein	small	väike	petit	piccolo	malý

	velký	groß	big	suur	grand	grande	veľký
	tenký	dünn	thin	peenike	mince	sottile	tenký
	pyramidální	pyramidenförmig	pyramidal	püramiidikujuline	pyramidal	piramidale	pyramídový
	trojrozměrný	dreidimensional	threedimensional	kolmemõõtmeline	tridimensionnel	tridimensionale	trojrozmerný
	tvrdý	hart	hard	kõva	dur	duro	tvrdý
	pružný	elastisch	elastic	elastne	élastique	elastico	pružný
	bílý	weiß	white	valge	blanc	bianco	biely
	černý	schwarz	black	must	noir	nero	čierny
	vědecký	wissenschaftlich	scientific	teaduslik	scientifique	scientifico	vedecký
	jedovatý	giftig	poisonous	mürgine	venimeux	velenoso	jedovatý
	adjectives	adjectives	adjectives	adjectives	adjectives	adjectives	adjectives
0	Dimensional	Dimensional	Dimensional	Dimensional	Dimensional	Dimensional	Dimensional
U	Features 2	Features 2	Features_2	Features 2	Features 2	Features 2	Features 2
	čtvercový	quadratisch	squared	ruudukujuline	carré	quadrato	štvorcový
	kulatý	rund	rounded	ümmargune	arrondi	rotondo	okrúhly
	kuželový	konisch	conical	koonusekujuline	conique	conico	kužeľové
	oválný	oval	oval	ovaalne	ovale	ovale	oválny
	válcový	zylindrisch	cylindrical	silindriline	cylindrique	cilindrico	valcový
	trojúhelníkový	dreieckig	triangular	kolmnurkne	triangulaire	triangolare	trojuholníkový
	obdélníkový	rechteckig	rectangular	ristkülikukujuline	rectangulaire	rettangolare	obdĺžnikový
	kulový	kugelförmig	spherical	kerakujuline	sphérique	sferico	guľový
	a aturía	h rf	-h	4			a studi
	ostrý	scharf flach	sharp flat	terav	pointu	acuto	ostrý
	plochý			lame	plat	piatto	plochý
	geometrický	geometrisch	geometrical	geomeetriline	géométrique	geometrico	geometrický
	matematický	mathematisch	mathematical	matemaatiline	mathématique	matematico	matematický
	těžký	schwer	heavy	raske	pesant	pesante	ťažký
	lehký	leicht	light	kerge	léger aléatoire	leggero	ľahký
	nahodilý	zufällig	random	juhuslik		casuale	náhodný
	stejný	identisch	identical	identne	identique	uguale	identický
	nouns	nouns	nouns	nouns	nouns	nouns	nouns

21	Dishes_and_ Cutlery						
	talíř	Teller	plate	taldrik	assiette	piatto	tanier
	miska	Schüssel	bowl	kauss	bol	ciotola	miska
	sklenice	Glas	glass	klaas	verre	bicchiere	pohár
	hrnek	Tasse	mug	kruus	mug	tazza	hrnček
	šálek	Kelch	cup	tass	tasse	calice	šálka
	vidlička	Gabel	fork	kahvel	fourchette	forchetta	vidlička
	nůž	Messer	knife	nuga	couteau	coltello	nôž
	lžíce	Löffel	spoon	lusikas	cuillère	cucchiaio	lyžička
	kladivo	Hammer	hammer	vasar	marteau	martello	kladivo
	sekera	Axt	axe	kirves	hache	ascia	sekera
	sendvič	Sandwich	sandwich	võileib	sandwich	sandwich	sendvič
	omeleta	Omelett	omelette	omlett	omelette	omelette	omeleta
	doušek	Schluck	sip	lonks	gorgée	sorso	dúšok
	pití	Trinkgelage	drink	jook	beuverie	bevuta	pitie
	díra	Loch	hole	auk	trou	buco	dierka
	balónek	Luftballon	balloon	õhupall	ballon	palloncino	balón
	nouns						
22	Economics						
	banka	Bank	bank	pank	banque	banca	banka
	peníze	Geld	money	raha	argent	denaro	peniaze
	rozpočet	Jahresabschluss	financial_statement	finantsaruanne	bilan	bilancio	rozpočet
	daňový_ráj	Steueroase	tax_haven	maksuparadiis	paradis_fiscal	paradiso_fiscale	daňový_raj
	dluh	Schuld	debt	võlg	dette	debito	dlh
	kreditní_karta	Kreditkarte	credit_card	krediitkaart	carte_de_crédit	carta_di_credito	kreditná_karta
	inflace	Inflation	inflation	inflatsioon	inflation	inflazione	inflácia
	krach	Konkurs	bankrupt	pankrot	banqueroute	bancarotta	bankrot
	vězení	Gefängnis	prison	vangla	prison	prigione	väzenie
	magistrát	Magistrat	magistrate	magistraat	magistrat	magistrato	magistrát
	epidemie	Epidemie	epidemic	epideemia	épidémie	epidemia	epidémia
	nebezpečí	Gefahr	danger	oht	danger	pericolo	nebezpečenstvo
----	----------------	----------------	-----------------	----------------	---------------------	--------------------	-------------------
	ráj	Paradies	paradise	paradiis	paradis	paradiso_terrestre	raj
	pečicí_papír	Backpapier	parchment_paper	tervituskaart	papier_sulfurisé	carta_da_forno	papier_na_pečenie
	housenka	Raupe	caterpillar	röövik	chenille	bruco	húsenica
	cirkus	Zirkus	circus	tsirkus	cirque	circo	cirkus
	nouns	nouns	nouns	nouns	nouns	nouns	nouns
23	Electronics	Electronics	Electronics	Electronics	Electronics	Electronics	Electronics
	televize	Fernsehen	television	televiisor	télévision	televisione	televízor
	reproduktor	Lautsprecher	speaker	kõlar	hautparleur	altoparlante	reproduktor
	notebook	Laptop	laptop	sülearvuti	ordinateur_portable	laptop	notebook
	tablet	Tablet	tablet	tahvelarvuti	tablette_tactile	tablet	tablet
	počítač	Computer	computer	arvuti	ordinateur	computer	počítač
	mobil	Handy	mobile_phone	mobiiltelefon	téléphone_portable	cellulare	mobil
	rádio	Radio	radio	raadio	radio	radio	rádio
	playstation	Playstation	playstation	playstation	playstation	playstation	playstation
	blok	Notizbuch	notebook	märkmik	cahier	taccuino	blok
	sešit	Arbeitsheft	workbook	vihik	classeur	eserciziario	zošit
	kniha	Buch	book	raamat	livre	libro	kniha
	film	Film	movie	film	film	film	film
	energie	Energie	energy	energia	puissance	energia	energia
	světlo	Licht	light	valgus	lumière	luce	svetlo
	oslava	Feier	party	pidu	fête	festa	oslava
	bublina	Blase	bubble	mull	bulle	bolla	bublina
	nouns	nouns	nouns	nouns	nouns	nouns	nouns
24	External_Body_	External_Body_	External_Body_	External_Body_	External_Body_Pa	External_Body_	External_Body_
	Parts	Parts	Parts	Parts	rts	Parts	Parts
	břicho	Bauch	abdomen	kõht	abdomen	addome	brucho
	záda	Rücken	back	selg	dos	schiena	chrbát
	paže	Arm	arm	käsi	bras	braccio	rameno
	noha	Bein	leg	jalg	jambe	gamba	noha
	bok	Hüfte	hip	puus	hanche	fianco	bok

	koleno	Knie	knee	põlv	genou	ginocchio	koleno
	rameno	Schulter	shoulder	õlg	épaule	spalla	plece
	lýtko	Wade	calf	säär	mollet	polpaccio	lýtko
	ledvina	Niere	kidney	neer	rein	rene	oblička
	plíce	Lunge	lung	kops	poumon	polmone	pľúca
	kloub	Gelenk	joint	liiges	articulation	articolazione	kĺb
	chrupavka	Knorpel	cartilage	kõhr	cartilage	cartilagine	chrupka
	chuť_k_jídlu	Appetit	appetite	söögiisu	appétit	appetito	chuť_do_jedla
	pot	Schweiß	sweat	higi	sueur	sudore	pot
	dáma	Dame	lady	daam	madame	signora	dáma
	kontinent	Kontinent	continent	maailmajagu	continent	continente	kontinent
	nouns	nouns	nouns	nouns	nouns	nouns	nouns
25	Extreme_Natural_	Extreme_Natural_	Extreme_Natural_	Extreme_Natural_	Extreme_Natural_	Extreme_Natural_	Extreme_Natural_
	Events	Events	Events	Events	Events	Events	Events
	požár	Brand	fire	tulekahju	incendie	incendio	požiar
	povodně	Überschwemmung	flood	üleujutus	inondation	alluvione	záplavy
	hurikán	Wirbelsturm	hurricane	orkaan	ouragan	uragano	hurikán
	tsunami	Tsunami	tsunami	tsunami	tsunami	tsunami	tsunami
	tornádo	Tornado	tornado	tornaado	tornade	tornado	tornádo
	zemětřesení	Erdbeben	earthquake	maavärin	tremblement_de_terre	terremoto	zemetrasenie
	erupce	Eruption	volcanic_eruption	vulkaanipurse	éruption_volcanique	eruzione_vulcanica	erupcia
	lavina	Lawine	avalanche	laviin	avalanche	valanga	lavína
	déšť	Regen	rain	vihm	pluie	pioggia	dážď
					-		
	vítr	Wind	wind	tuul	vent	vento	vietor
	bombardování	Wind Bombardierung	bombardment	pommitamine	bombardement	bombardamento	bombardovanie
		Wind Bombardierung Genozid	bombardment genocide	pommitamine genotsiid	bombardement génocide	bombardamento genocidio	
	bombardování genocida sopka	Wind Bombardierung Genozid Vulkan	bombardment	pommitamine genotsiid vulkaan	bombardement génocide volcan	bombardamento genocidio vulcano	bombardovanie genocída sopka
	bombardování genocida sopka řeka	Wind Bombardierung Genozid	bombardment genocide	pommitamine genotsiid	bombardement génocide	bombardamento genocidio	bombardovanie genocída
	bombardování genocida sopka	Wind Bombardierung Genozid Vulkan	bombardment genocide volcano	pommitamine genotsiid vulkaan	bombardement génocide volcan	bombardamento genocidio vulcano	bombardovanie genocída sopka

	nouns						
26	Family_Members						
	babička	Großmutter	grandmother	vanaema	grandmère	nonna	babka
	dědeček	Großvater	grandfather	vanaisa	grandpère	nonno	dedko
	sestra	Schwester	sister	õde	sœur	sorella	sestra
	matka	Mutter	mother	ema	mère	madre	matka
	teta	Tante	aunt	tädi	tante	zia	teta
	strýc	Onkel	uncle	onu	oncle	zio	strýko
	bratr	Bruder	brother	vend	frère	fratello	brat
	otec	Vater	father	isa	père	padre	otec
	člověk	Mensch	human	inimene	humain	essere_umano	človek
	osoba	Person	person	isik	personne	persona	osoba
	stařec	Greis	old_man	vanamees	vieil_homme	anziano	starec
	puberťák	Teenager	teenager	teismeline	adolescent	adolescente	adolescent
	domácí_zvíře	Haustier	pet	lemmikloom	animal_de_compagni	animale_domestico	maznáčik
	chůva	Babysitter	babysitter	lapsehoidja	nourrice	babysitter	pestúnka
	svíčka	Kerze	candle	küünal	bougie	candela	sviečka
	tetování	Tätowierung	tattoo	tätoveering	tatouage	tatuaggio	tetovanie
	nouns						
27	Fantasy_Characters						
	upír	Vampir	vampire	vampiir	vampire	vampiro	upír
	čarodějnice	Hexe	witch	nõid	sorcière	strega	čarodejnica
	skřítek	Kobold	elf	haldjas	elfe	elfo	škriatok
	drak	Drache	dragon	draakon	dragon	drago	drak
	jednorožec	Einhorn	unicorn	ükssarvik	licorne	unicorno	jednorožec
	mořská panna	Meerjungfrau	mermaid	merineitsi	sirène	sirena	morská_panna
	zombie	Zombie	zombie	zombi	zombie	zombie	zombie
	vlkodlak	Werwolf	werewolf	libahunt	loupgarou	lupo_mannaro	vlkolak
	had	Schlange	snake	madu	serpent	serpente	had
	člověk	Mensch	human	inimene	humain	essere umano	človek

	kouzelná_hůlka	Zauberstab	wand	võlurikepp	baguette_magique	bacchetta_magica	kúzelnícka_palička
	lektvar	Zaubertrank	potion	võlujook	potion	pozione	elixír
	magie	Magie	magic	maagia	magie	magia	mágia
	zaklínadlo	Zauberspruch	magic_spell	loits	sortilège	incantesimo	zaklínadlo
	příroda	Natur	nature	loodus	nature	natura	príroda
	poledne	Mittag	noon	keskpäev	midi	mezzogiorno	poludnie
	nouns	nouns	nouns	nouns	nouns	nouns	nouns
28	Farm_Animals	Farm_Animals	Farm_Animals	Farm_Animals	Farm_Animals	Farm_Animals	Farm_Animals
	kráva	Kuh	cow	lehm	vache	mucca	krava
	prase	Schwein	pig	siga	cochon	maiale	prasa
	kůň	Pferd	horse	hobune	cheval	cavallo	kôň
	králík	Kaninchen	rabbit	küülik	lapin	coniglio	králik
	koza	Ziege	goat	kits	chèvre	capra	koza
	slepice	Huhn	hen	kana	poule	gallina	sliepka
	ovce	Schaf	sheep	lammas	brebis	pecora	ovce
	osel	Esel	donkey	eesel	âne	asino	osol
	liška	Fuchs	fox	rebane	renard	volpe	líška
	vlk	Wolf	wolf	hunt	loup	lupo	vlk
	šunka	Schinken	ham	sink	jambon	prosciutto	šunka
	pečené_kuře	Schweinebraten	roast_chicken	ahjukana	poulet_rôti	pollo_arrosto	pečené kurča
	dojení	Melken	milking	lüpsmine	traite	mungitura	dojenie
	chov	Zucht	livestock	kariloomad	bétail	allevamento	chov
	podzim	Herbst	autumn	sügis	automne	autunno	jeseň
	plenka	Windel	diaper	mähkmed	couche	pannolino	plienka
	nouns	nouns	nouns	nouns	nouns	nouns	nouns
29	Firearms	Firearms	Firearms	Firearms	Firearms	Firearms	Firearms
	puška	Gewehr	rifle	vintpüss	fusil	fucile	puška
	pistole	Feuerwaffe	gun	relv	pistolet	pistola	pištoľ
	kulomet	Maschinengewehr	machine_gun	kuulipilduja	mitrailleuse	mitragliatrice	gul'omet
	samopal	Karabiner	carbine	karabiin	carabine	carabina	karabína
	brokovnice	Kalaschnikow	kalashnikov	kalašnikov	kalachnikov	kalashnikov	brokovnica

	revolver	Revolver	revolver	revolver	revolver	rivoltella	revolver
	granátomet	Pistole	pistol	püstol	lancegrenades	lanciagranate	samopal
	mušketa	Muskete	musket	musket	mousquet	moschetto	mušketa
	střela	Projektil	projectile	mürsk	projectile	proiettile	strela
	meč	Schwert	sword	mõõk	épée	spada	meč
	tank	Tank	tank	tank	char	carro_armato	tank
	voják	Soldat	soldier	sõdur	soldat	soldato	vojak
	válka	Krieg	war	sõda	guerre	guerra	vojna
	vrah	Mörder	murderer	mõrvar	tueur	assassino	vrah
	slovník	Wörterbuch	dictionary	sõnastik	dictionnaire	dizionario	slovník
	literatura	Literatur	literature	kirjandus	littérature	letteratura	literatúra
	nouns	nouns	nouns	nouns	nouns	nouns	nouns
30	Fish	Fish	Fish	Fish	Fish	Fish	Fish
	tuňák	Thunfisch	tuna	tuunikala	thon	tonno	tuniak
	žralok	Hai	shark	hai	requin	squalo	žralok
	úhoř	Aal	eel	angerjas	anguille	anguilla	úhor
	losos	Lachs	salmon	lõhe	saumon	salmone	losos
	pstruh	Forelle	trout	forell	truite	trota	pstruh
	sardinka	Sardine	sardine	sardiin	sardine	sardina	sardinka
	treska	Kabeljau	cod	tursk	morue	merluzzo	treska
	kapr	Karpfen	carp	karpkala	carpe	carpa	kapor
	delfín	Delphin	dolphin	delfiin	dauphin	delfino	delfín
	velryba	Wal	whale	vaal	baleine	balena	veľryba
	korál	Koralle	coral	korall	corail	corallo	koral
	mořské_řasy	Algen	seaweed	vetikad	algue	alga	riasy
	sushi	Sushi	sushi	sushi	sushi	sushi	sushi
	rybí_prsty	Fischstäbchen	fish_pie	suitsukala	bouillabaisse	fritto_misto	rybačka
	plamen	Flamme	flame	leek	flamme	fiamma	plameň
	klaun	Clown	clown	kloun	clown	pagliaccio	klaun
	nouns	nouns	nouns	nouns	nouns	nouns	nouns

31	Flowers						
	růže	Rose	rose	roos	rose	rosa	ruža
	sedmikráska	Gänseblümchen	daisy	kirikakar	marguerite	margherita	orchidea
	sněženka	Narzisse	daffodil	nartsiss	glycine	glicine	snežienka
	tulipán	Tulpe	tulip	tulp	tulipe	tulipano	tulipán
	mák	Mohn	рорру	moon	coquelicot	papavero	mak
	pelargonie	Geranie	geranium	kanarbik	géranium	geranio	fialka
	slunečnice	Sonnenblume	sunflower	päevalill	tournesol	girasole	slnečnica
	lilie	Lilie	lily	liilia	lys	giglio	l'alia
	stonek	Stängel	corolla	õiekroon	corolle	corolla	stonka
	okvětní lístek	Blütenblatt	petal	kroonleht	pétale	petalo	okvetný_lístok
	vrba	Weide	willow	paju	saule	salice	vŕba
	kaktus	Kaktus	cactus	kaktus	cactus	cactus	kaktus
	zahrada	Garten	garden	aed	jardin	giardino	záhrada
	konev	Gießkanne	watering_can	kastekann	arrosoir	annaffiatoio	krhla
	opasek	Gürtel	belt	vöö	ceinture	cintura	opasok
	kohoutek	Wasserhahn	tap	kraan	robinet	rubinetto	kohútik
	nouns						
32	Flying_Means_of_ Transport						
	letadlo	Flugzeug	airplane	lennuk	avion	aeroplano	lietadlo
	horkovzdušný_balón	Heißluftballon	hot_air_balloon	kuumaõhupall	montgolfière	mongolfiera	teplovzdušný_balón
	rogalo	Gleiter	hang_glider	deltaplaan	deltaplane	deltaplano	rogalo
	vrtulník	Hubschrauber	helicopter	helikopter	hélicoptère	elicottero	vrtuľník
	kluzák	Segelflugzeug	sailplane	purilennuk	planeur	aliante	klzák
	stíhačka	Gleitschirm	paraglider	paraplaan	parapente	parapendio	stíhačka
	tryskáč	Jet	jet	reaktiivlennuk	jet	jet	vetroň
	vzducholoď	Luftschiff	airship	õhulaev	dirigeable	dirigibile	vzducholoď
	auto	Auto	car	auto	voiture	automobile	auto
	vlak	Zug	train	rong	train	treno	vlak
	pták	Vogel	bird	lind	oiseau	uccello	vták

	papírový_drak	Kite	kite	tuulelohe	cerfvolant	aquilone	šarkan
	přistání	Landung	landing	maandumine	atterrissage	atterraggio	pristátie
	letuška	Flugbegleiter	flight_attendant	stjuuardess	hôtesse	assistente_di_volo	letuška
	demolice	Abbruch	demolition	lammutamine	démolition	distruzione	demolácia
	mýdlo	Seife	soap	seep	savon	sapone	mydlo
	nouns	nouns	nouns	nouns	nouns	nouns	nouns
33	Food	Food	Food	Food	Food	Food	Food
	zelenina	Gemüse	vegetable	köögiviljad	légumes	verdura	zelenina
	ovoce	Obst	fruit	puuviljad	fruits	frutta	ovocie
	chléb	Brot	bread	leib	pain	pane	chlieb
	ryba	Fisch	fish	kala	poisson	pesce	ryba
	maso	Fleisch	meat	liha	viande	carne	mäso
	vejce	Ei	egg	muna	œuf	uova	vajce
	luštěnina	Hülsenfrücht	legume	kaunviljad	légumineuse	legume	strukovina
	obilovina	Getreide	grain	teraviljad	céréale	cereale	obilnina
	lžíce	Löffel	spoon	lusikas	cuillère	cucchiaio	lyžica
	talíř	Teller	dish	roog	plat	piatto	tanier
	jídlo	Mahlzeit	meal	eine	repas	pasto	jedlo
	večeře	Abendessen	dinner	õhtusöök	dîner	cena	večera
	chuť	Geschmack	flavour	maitse	saveur	sapore	chuť
	smažení	Braten	frying	praadimine	friture	frittura	vyprážanie
	duha	Regenbogen	rainbow	vikerkaar	arcenciel	arcobaleno	dúha
	sanitka	Krankenwagen	ambulance	kiirabi	ambulance	ambulanza	sanitka
	adjectives	adjectives	adjectives	adjectives	adjectives	adjectives	adjectives
34	Food Features	Food Features	Food Features	Food_Features	Food Features	Food Features	Food Features
	sladký	süß	sweet	magus	sucré	dolce	sladký
	slaný	salzig	salty	soolane	salé	salato	slaný
	hořký	bitter	bitter	mõru	amer	amaro	horký
	kořeněný	würzig	spicy	vürtsikas	épicé	piccante	pikantný
	kyselý	sauer	sour	hapu	aigre	acido	kyslý
	mdlý	fade	insipid	mage	insipide	insipido	mdlý

	trpký	säuerlich	tart	mõrkjas	âpre	aspro	trpký
	chutný	schmackhaft	tasty	maitsev	savoureux	saporito	chutný
	nadýchaný	locker	fluffy	kohev	moelleux	soffice	nadýchaný
	sušený	getrocknet	dried	kuiv	sec	secco	sušený
	smradlavý	stinkend	stinky	haisev	puant	puzzolente	smradľavý
	vonný	aromatisch	aromatic	aromaatne	aromatique	aromatico	voňavý
	syrový	roh	raw	toores	cru	crudo	surový
	prošlý	abgelaufen	expired	aegunud	expiré	scaduto	zhnitý
	atletický	sportlich	athletic	sportlik	athlétique	atletico	atletický
	hornatý	gebirgig	mountainous	mägine	montagneux	montuoso	hornatý
	nouns	nouns	nouns	nouns	nouns	nouns	nouns
5	Free_Time_	Free_Time_	Free_Time_	Free_Time_	Free_Time_	Free_Time_	Free_Time_
-	Activities	Activities	Activities	Activities	Activities	Activities	Activities
	vaření	Kochen	cooking	toiduvalmistamine	cuisine	cucina	varenie
	čtení	Lesen	reading	lugemine	lecture	lettura	čítanie
	zahradničení	Gartenarbeit	gardening	aiandus	jardinage	giardinaggio	záhradníctvo
	kutilství	Selbermachen	DIY	isetegemine	bricolage	bricolage	majstrovanie
	modelářství	Modellbau	model_building	kalapüük	modelage	modellismo	modelovanie
	šití	Nähen	needlework	õmblemine	couture	cucito	šitie
	turistika	Wandern	hiking	matkamine	randonnée_pédestre	escursionismo	turistika
	malování	Malen	painting	maalimine	peinture	pittura	mal'ovanie
	hokej	Eishockey	hockey	hoki	hockey	hockey	hokej
	fotbal	Fußball	football	jalgpall	football	calcio	futbal
	volný čas	Freizeit	free_time	vaba_aeg	temps_libre	tempo_libero	voľný čas
	víkend	Wochenende	weekend	nädalavahetus	weekend	weekend	víkend
	nástroj	Werkzeug	tool	tööriist	outil	attrezzo	nástroj
	kniha	Buch	book	raamat	livre	libro	kniha
	buňka	Zelle	cell	rakk	cellule	cellula	bunka
	hračka	Spielzeug	toy	mänguasi	jouet	giocattolo	hračka
	nouns	nouns	nouns	nouns	nouns	nouns	nouns

36	Fruit	Fruit	Fruit	Fruit	Fruit	Fruit	Fruit
	pomeranč	Orange	orange	apelsin	orange	arancia	pomaranč
	malina	Himbeere	raspberry	vaarikas	framboise	lampone	malina
	borůvka	Heidelbeere	blueberry	mustikas	myrtille	mirtillo	čučoriedka
	ananas	Ananas	pineapple	ananass	ananas	ananas	ananás
	hroznové_víno	Traube	grape	viinamari	raisin	uva	hrozno
	meruňka	Aprikose	apricot	aprikoos	abricot	albicocca	marhul'a
	mango	Mango	mango	mango	mangue	mango	mango
	kiwi	Kiwi	kiwi	kiivi	kiwi	kiwi	kiwi
	okurka	Gurke	cucumber	kurk	concombre	cetriolo	uhorka
	dýně	Kürbis	pumpkin	kõrvits	citrouille	zucca	tekvica
	džus	Saft	juice	mahl	jus	succo	džús
	džem	Marmelade	jam	moos	confiture	marmellata	džem
	vejce	Ei	egg	muna	œuf	uovo	vajce
	cukr	Zucker	sugar	suhkur	lait	zucchero	cukor
	ponorka	UBoot	submarine	allveelaev	sousmarin	sottomarino	ponorka
	buben	Trommel	drum	trumm	tambour	tamburo	bubon
	nouns	nouns	nouns	nouns	nouns	nouns	nouns
37	Fruit_Trees	Fruit_Trees	Fruit_Trees	Fruit_Trees	Fruit_Trees	Fruit_Trees	Fruit_Trees
	jabloň	Apfelbaum	apple_tree	õunapuu	pommier	melo	jabloň
	hrušeň	Birnbaum	pear_tree	pirnipuu	poirier	pero	hruška
	třešeň	Kirschbaum	cherry_tree	kirsipuu	cerisier	ciliegio	čerešňa
	švestka	Pflaumenbaum	plum_tree	ploomipuu	prunier	pruno	slivka
	broskvoň	Pfirsichbaum	peach_tree	virsikupuu	pêcher	pesco	broskyňa
	pomerančovník	Orangenbaum	orange_tree	apelsinipuu	oranger	arancio	pomarančovník
	meruňka	Aprikosenbaum	apricot_tree	aprikoosipuu	abricotier	albicocco	marhul'a
	morušovník	Feigenbaum	mulberry_tree	mooruspuu	bananier	banano	moruša
	tis	Eibe	yew	jugapuu	if	tasso	tis
	buk	Buche	beech	pöök	hêtre	faggio	buk
	lilie	Lilie	lily	liilia	lis	giglio	ľalia
	sedmikráska	Gänseblümchen	daisy	karikakar	marguerite	margherita	sedmokráska

	větev	Zweig	branch	oks	branche	ramo	vetva
	kořen	Wurzel	root	juur	racine	radice	koreň
	štěstí	Glück	fortune	õnn	chance	fortuna	šťastie
	dolar	Dollar	dollar	dollar	dollar	dollaro	dolár
	nouns	nouns	nouns	nouns	nouns	nouns	nouns
38	Furniture	Furniture	Furniture	Furniture	Furniture	Furniture	Furniture
	křeslo	Sessel	armchair	tugitool	fauteuil	poltrona	kreslo
	pohovka	Sofa	sofa	diivan	canapé	divano	pohovka
	stolička	Hocker	stool	tool	tabouret	sgabello	stolička
	skříň	Schrank	wardrobe	riidekapp	armoire	armadio	skriňa
	polička	Regal	shelf	riiul	étagère	scaffale	polička
	stůl	Tisch	table	laud	table	tavolo	stôl
	noční_stolek	Nachttisch	bed_table	öökapp	table_de_chevet	comodino	nočný_stolík
	postel	Bett	bed	voodi	lit	letto	posteľ
	lednice	Kühlschrank	fridge	külmkapp	réfrigérateur	frigorifero	chladnička
	myčka	Spülmaschine	dishwasher	nõudepesumasin	lavevaisselle	lavastoviglie	umývačka
	odpadkový_koš	Abfalleimer	trash_can	prügikast	poubelle	pattumiera	odpadkový_kôš
	koště	Besen	broom	luud	balai	scopa	metla
	dřevo	Holz	wood	puit	bois	legno	drevo
	sklo	Glas	glass	klaas	verre	vetro	sklo
	komedie	Komödie	comedy	komöödia	comédie	commedia	komédia
	pero	Feder	feather	kübarasulg	plume	piuma	pero
	nouns	nouns	nouns	nouns	nouns	nouns	nouns
39	Gemstones	Gemstones	Gemstones	Gemstones	Gemstones	Gemstones	Gemstones
	diamant	Diamant	diamond	teemant	diamant	diamante	diamant
	perla	Perle	pearl	pärl	perle	perla	perla
	rubín	Rubin	ruby	rubiin	rubis	rubino	rubín
	smaragd	Smaragd	emerald	smaragd	émeraude	smeraldo	smaragd
	safír	Saphir	sapphire	safiir	saphir	zaffiro	zafír
	topaz	Topas	topaz	topaas	topaze	topazio	zafír
	jantar	Bernstein	amber	merevaik	ambre	ambra	jantár

pohanka	Buchweizen	buckwheat	tatar	sarrasin	grano_saraceno	pohánka
	Roggen	rye	rukis	seigle	segale	žito
oves	Hafer	oat	kaer		avena	ovos
proso		millet	hirss		miglio	proso
					orzo	jačmeň
1						pšenica
•						
						ryža
						kukurica
						Grain
nouns	nouns	nouns	nouns	nouns	nouns	nouns
						1
zákon	Gesetz	law	seadus	loi	legge	právo
hlava	Kopf	•		tête	testa	hlava
zlatník	-		kullassepp	orfèvre	2	klenotník
klenotnictví		-				klenotníctvo
prstýnek	Ring	ring	sõrmus	bague	anello	prsteň
náhrdelník	Halskette	necklace	kaelakee	collier	collana	náhrdelník
sklo	Glas	glass	klaas	verre	vetro	sklo
stříbro	Silber	silver	hõbe	argent	argento	striebro
	sklo náhrdelník prstýnek klenotnictví zlatník hlava zákon <b>nouns</b> <b>Grain</b> kukuřice rýže pšenice ječmen proso	stříbro   Silber     sklo   Glas     náhrdelník   Halskette     prstýnek   Ring     klenotnictví   Juweliergeschäft     zlatník   Goldschmied     hlava   Kopf     zákon   Gesetz     nouns   nouns     Grain   Grain     kukuřice   Mais     rýže   Reis     pšenice   Weizen     ječmen   Gerste     proso   Hirse     oves   Hafer     žito   Roggen	stříbroSilbersilverskloGlasglassnáhrdelníkHalskettenecklaceprstýnekRingringklenotnictvíJuweliergeschäftjewelry_storezlatníkGoldschmiedgoldsmithhlavaKopfheadzákonGesetzlawnounsnounsnounsGrainGrainGrainkukuřiceMaiscornrýžeReisricepšeniceWeizenwheatječmenGerstebarleyprosoHirsemilletovesHaferoatžitoRoggenrye	stříbroSilbersilverhõbeskloGlasglassklaasnáhrdelníkHalskettenecklacekaelakeeprstýnekRingringsõrmusklenotnictvíJuweliergeschäftjewelry_storeehezlatníkGoldschmiedgoldsmithkullassepphlavaKopfheadpeazákonGesetzlawseadusnounsnounsnounsnounsfýžeReisriceriisrýžeReisriceriispšeniceWeizenwheatnisuječmenGerstebarleyoderprosoHirsemillethirssovesHaferoatkaeržitoRoggenryerukis	stříbroSilbersilverhöbeargentskloGlasglassklaasverrenáhrdelníkHalskettenecklacekaelakeecollierprstýnekRingringsõrmusbagueklenotnictvíJuweliergeschäftjewelry_storeehejoailleriezlatníkGoldschmiedgoldsmithkullassepporfèvrehlavaKopfheadpeatêtezákonGesetzlawseadusloinounsnounsnounsnounsnounsGrainGrainGrainGrainGrainkukuřiceMaiscornmaismaisrýžeReisriceriisrizpšeniceWeizenwheatnisubléječmenGerstebarleyoderorgeprosoHirsemillethirssmilletovesHaferoatkaeravoinežitoRoggenryerukisseigle	stříbroSilbersilverhöbeargentargentoskloGlasglassklaasverrevetronáhrdelníkHalskettenecklacekaelakeecolliercollanaprstýnekRingringsörmusbagueanelloklenotnictvíJuweliergeschäftjewelry_storeehejoailleriegioielleriazlatníkGoldschmiedgoldsmithkullassepporfèvreoreficehlavaKopfheadpeatêtetestazákonGesetzlawseadusloileggenounsnounsnounsnounsnounsnounsGrainGrainGrainGrainGrainGraintkukuřiceMaiscornmaismaisrizpšeniceWeizenwheatnisubléfrumentoječmenGerstebarleyoderorgeorzoprosoHirsemillethirssmilletmiglioovesHaferoatkaeravoineavenažitoRoggenryerukisseiglesegale

	šedovlasý	grauhaarig	greyhaired	hallipäine	grisonnant	brizzolato	sivovlasý
	kudrnatý	kraus	frizzy	käharpäine	crêpelé	crespo	zvlnený
	kadeřavý	lockig	curly	lokkis	frisés	riccio	kučeravý
	rovný	rothaarig	straight	sirge	lisse	liscio	rovný
	rozcuchaný	ungekämmt	uncombed	kammimata	décoiffé	spettinato	ryšavý
	plešatý	glatzköpfig	bald	kiilas	chauve	pelato	plešatý
	oholený	rasiert	shaved	raseeritud	tondu	rasato	oholený
	vrásčitý	runzelig	wrinkled	kortsus	ridé	rugoso	zvráskavený
	vychrtlý	ausgemergelt	skinny	kõhn	décharné	scarno	vychudnutý
	růžový	rosig	rosy	roosiline	rose	roseo	ružový
	pihovatý	sommersprossig	freckled	tedretähniline	rubicond	lentigginoso	pehavý
	vystouplý	vorstehend	protruding	väljaulatuv	protubérant	sporgente	vypuklý
	ohebný	langbeinig	slender	sihvakas	élancé	slanciato	ohybný
	těžký	schwierig	difficult	keeruline	difficile	difficile	ťažký
	snadný	einfach	easy	lihtne	facile	facile	jednoduchý
	nouns	nouns	nouns	nouns	nouns	nouns	nouns
42	Herbs	Herbs	Herbs	Herbs	Herbs	Herbs	Herbs
	levandule	Lavendel	lavender	lavendel	lavande	lavanda	levandul'a
	tymián	Thymian	thyme	tüümian	thym	timo	tymián
	tymián rozmarýn	Thymian Rosmarin	thyme rosemary	tüümian rosmariin	thym romarin	timo rosmarino	tymián rozmarín
	rozmarýn máta	· ·			-		•
	rozmarýn máta majoránka	RosmarinMinzeMajoran	rosemary	rosmariin münt majoraan	romarin	rosmarino menta maggiorana	rozmarín mäta majorán
	rozmarýn máta	RosmarinMinzeMajoranSalbei	rosemary mint marjoram sage	rosmariin münt majoraan salvei	romarin menthe marjolaine sauge	rosmarino menta maggiorana salvia	rozmarín mäta majorán šalvia
	rozmarýn máta majoránka	RosmarinMinzeMajoran	rosemary mint marjoram	rosmariin münt majoraan	romarin menthe marjolaine	rosmarino menta maggiorana	rozmarín mäta majorán
	rozmarýn máta majoránka šalvěj	RosmarinMinzeMajoranSalbei	rosemary mint marjoram sage	rosmariin münt majoraan salvei	romarin menthe marjolaine sauge	rosmarino menta maggiorana salvia	rozmarín mäta majorán šalvia
	rozmarýn máta majoránka šalvěj bazalka kopr	RosmarinMinzeMajoranSalbeiBasilikumDill	rosemary mint marjoram sage basil dill	rosmariin münt majoraan salvei basiilik till	romarin menthe marjolaine sauge basilic aneth	rosmarino menta maggiorana salvia basilico aneto	rozmarín mäta majorán šalvia bazalka kôpor
	rozmarýn máta majoránka šalvěj bazalka kopr dub	Rosmarin     Minze     Majoran     Salbei     Basilikum     Dill     Eiche	rosemary mint marjoram sage basil dill oak	rosmariin münt majoraan salvei basiilik till tamm	romarin menthe marjolaine sauge basilic aneth chêne	rosmarino menta maggiorana salvia basilico aneto quercia	rozmarín mäta majorán šalvia bazalka kôpor dub
	rozmarýn máta majoránka šalvěj bazalka kopr dub jasan	Rosmarin     Minze     Majoran     Salbei     Basilikum     Dill     Eiche     Esche	rosemary mint marjoram sage basil dill oak oak ash	rosmariin münt majoraan salvei basiilik till till tamm saar	romarin menthe marjolaine sauge basilic aneth chêne frêne	rosmarino menta maggiorana salvia basilico aneto quercia frassino	rozmarín mäta majorán šalvia bazalka kôpor dub jaseň
	rozmarýn máta majoránka šalvěj bazalka kopr dub jasan jahodník	Rosmarin     Minze     Majoran     Salbei     Basilikum     Dill     Eiche     Esche     Erdbeere	rosemary mint marjoram sage basil dill oak ash strawberry	rosmariin münt majoraan salvei basiilik till till tamm saar maasikas	romarin menthe marjolaine sauge basilic aneth chêne frêne fraise	rosmarino menta maggiorana salvia basilico aneto quercia frassino fragola	rozmarín mäta majorán šalvia bazalka kôpor dub jaseň jahoda
	rozmarýn máta majoránka šalvěj bazalka kopr dub jasan jahodník malina	Rosmarin     Minze     Majoran     Salbei     Basilikum     Dill     Eiche     Esche     Erdbeere     Himbeere	rosemary mint marjoram sage basil dill oak ash strawberry raspberry	rosmariin münt majoraan salvei basiilik till tamm saar maasikas vaarikas	romarin menthe marjolaine sauge basilic aneth chêne frêne fraise framboise	rosmarino menta maggiorana salvia basilico aneto quercia frassino fragola lampone	rozmarín mäta majorán šalvia bazalka kôpor dub jaseň jahoda malina
	rozmarýn máta majoránka šalvěj bazalka kopr dub jasan jahodník	Rosmarin     Minze     Majoran     Salbei     Basilikum     Dill     Eiche     Esche     Erdbeere	rosemary mint marjoram sage basil dill oak ash strawberry	rosmariin münt majoraan salvei basiilik till till tamm saar maasikas	romarin menthe marjolaine sauge basilic aneth chêne frêne fraise	rosmarino menta maggiorana salvia basilico aneto quercia frassino fragola	rozmarín mäta majorán šalvia bazalka kôpor dub jaseň jahoda

	kostka	Würfel	dice	täring	dé	dado	kocka
	slunce	Sonne	sun	päike	soleil	sole	slnko
	adjectives						
43	Human_Features_ Negativity						
	sobecký	egoistisch	selfish	isekas	égoïste	egoista	sebecký
	drzý	unhöflich	rude	ebaviisakas	grossier	sgarbato	drzý
	lakomý	geizig	stingy	kitsi	avare	avaro	lakomý
	nevrlý	mürrisch	surly	tusane	maussade	scontroso	nevrlý
	neslušný	unfreundlich	unfriendly	ebasõbralik	impoli	maleducato	nepriateľský
	nepoctivý	unehrlich	dishonest	ebaaus	malhonnête	disonesto	neslušný
	závistivý	neidisch	envious	kade	envieux	invidioso	závistlivý
	nesympatický	unsympathisch	unpleasant	ebameeldiv	antipathique	antipatico	nesympatický
	spolehlivý	zuverlässig	reliable	usaldusväärne	fiable	affidabile	spoľahlivý
	optimistický	optimistisch	optimistic	optimistlik	optimiste	ottimista	optimistický
	mladý	jung	young	noor	jeune	giovane	mladý
	zdravý	gesund	healthy	terve	sain	sano	zdravý
	škodlivý	schädlich	harmful	kahjulik	nuisible	dannoso	škodlivý
	druhořadý	lausig	shoddy	armetu	minable	scadente	druhoradý
	polární	polar	polar	polaarne	polaire	polare	polárny
	enormní	enorm	enormous	tohutu	énorme	enorme	enormný
	adjectives						
44	Human_Features_ Positivity						
	krásný	schön	beautiful	ilus	beau	bello	krásny
	laskavý	liebenswürdig	kind	lahke	prévenant	gentile	láskavý
	milý	nett	nice	tore	gentil	carino	milý
	ochotný	hilfsbereit	helpful	abivalmis	aimable	collaborativo	ochotný
	obětavý	selbstlos	unselfish	isetu	désintéressé	altruista	obetavý
	pohledný	hübsch	handsome	nägus	joli	avvenente	pekný
	usměvavý	lächelnd	smiling	naeratav	souriant	sorridente	usmievavý

	syntetický	synthetisch	synthetic	sünteetiline	synthétique	sintetico	syntetický
	severní	nördlich	northern	põhjapoolne	septentrional	settentrionale	severný
	tlustý	fett	fat	paks	gros	grasso	tučný
	štíhlý	schlank	slim	sale	mince	snello	štíhly
	nemocný starý	alt	old	haige vana	vieux	vecchio	starý
	laskavý	nett krank	sick		gentil malade	gentile malato	chorý
	pokrytecký	heuchlerisch	twofaced kind	kahepalgeline lahke	hypocrite	ipocrita gantila	pokrytecký láskavý
	úzkostný	ängstlich	alixious	rahutu	anxieux		úzkostný
	zoufalý úzkostný	verzweifelt	desperate anxious	meeleheitel	désespéré	disperato ansioso	zúfalý
	veselý	heiter	cheerful	rõõmsameelne	joyeux	allegro	veselý
	ustaraný	besorgt	worried	mures	préoccupé	preoccupato	znepokojený
	nervózní	nervös	nervous	närviline	nerveux	nervoso	nervózny
	naštvaný	wütend	angry	vihane	fâché	arrabbiato	naštvaný
	smutný	traurig	sad	kurb	triste	triste	smutný
	šťastný	fröhlich	happy	õnnelik	heureux	felice	šťastný
5	Human_Moods	Human_Moods	Human_Moods	Human_Moods	Human_Moods	Human_Moods	Human_Moods
	adjectives	adjectives	adjectives	adjectives	adjectives	adjectives	adjectives
	barevný	bunt	colorful	värvikas	coloré	colorato	farebný
	počítačový	computergestützt	computerized	digitaalne	informatisé	computerizzato	počítačový
	měkký	weich	soft	pehme	moelleux	morbido	mäkký
	pohodlný	bequem	comfortable	mugav	confortable	comodo	pohodlný
	líný	faul	lazy	laisk	paresseux	pigro	lenivý
	neochotný	unwillig	reluctant	vastumeelne	réticent	riluttante	neochotný
	štíhlý	mager	slim	sale	maigre	magro	štíhly
	vysoký	groß	tall	pikk	grand	alto	vysoký
	silný		strong	tugev	fort	forte	silný

	hubený	mager	thin	õhuke	maigre	magro	chudý
	vysoký	groß	tall	pikk	grand	alto	vysoký
	robustní	untersetzt	sturdy	turske	robuste	robusto	robustný
	buclatý	rundlich	plump	pontsakas	dodu	grassottello	bacuľatý
	statný	stämmig	stocky	jässakas	trapu	tozzo	územčistý
	podsaditý	gedrungen	burly	paks	costaud	tarchiato	statný
	malý	klein	short	lühike	petit	basso	malý
	oplácaný	mollig	chubby	priske	rondelet	paffuto	bacuľatý
	zakřivený	gekrümmt	curved	kõver	courbé	ricurvo	zakrivený
	nepravidelný	unregelmäßig	irregular	ebakorrapärane	irrégulier	irregolare	nepravidelný
	usměvavý	lächelnd	smiling	naeratav	souriant	sorridente	usmievavý
	klidný	ruhig	calm	rahulik	calme	calmo	pokojný
	tajemný	geheimnisvoll	mysterious	salapärane	mystérieux	misterioso	záhadný
	okouzlující	faszinierend	charming	võluv	charmant	affascinante	očarujúci
	náboženský	religiös	religious	religioosne	religieux	religioso	náboženský
	umělý	künstlich	artificial	kunstlik	artificiel	artificiale	umelý
	nouns	nouns	nouns	nouns	nouns	nouns	nouns
47	Illnesses	Illnesses	Illnesses	Illnesses	Illnesses	Illnesses	Illnesses
	rýma	Erkältung	cold	nohu	rhume	raffreddore	nádcha
	chřipka	Grippe	flu	gripp	grippe	influenza	chrípka
	alergie	Allergie	allergy	allergia	allergie	allergia	alergia
	zánět_průdušek	Bronchitis	bronchitis	bronhiit	bronchite	bronchite	zápal_priedušiek
	cukrovka	Diabetes	diabetes	diabeet	diabète	diabete	cukrovka
	horečka	Fieber	fever	palavik	fièvre	febbre	horúčka
	mor	Pest	plague	katk	peste	peste	mor
	bolest_zubů	Zahnschmerzen	toothache	hambavalu	mal_de_dent	mal_di_denti	bolesť_zuba
	zlomenina	Fraktur	fracture	luumurd	fracture	frattura	zlomenina
	modřina	Bluterguss	bruise	sinikas	contusion	livido	modrina
	smrt	Tod	death	surm	mort	morte	smrť
	nemocnice	Krankenhaus	hospital	haigla	hôpital	ospedale	nemocnica
	teploměr	Thermometer	1	0	thermomètre	1	

	mast	Salbe	ointment	salv	pommade	pomata	masť
	poznávací_značka	Nummernschild	plate	numbrimärk	plaque	targa	poznávacia_značka
	svět	Welt	world	maailm	monde	mondo	svet
	nouns	nouns	nouns	nouns	nouns	nouns	nouns
48	Informatics	Informatics	Informatics	Informatics	Informatics	Informatics	Informatics
	heslo	Passwort	password	parool	mot_de_passe	password	heslo
	stahování	Herunterladen	download	allalaadimine	téléchargement	download	sťahovanie
	základní_deska	Hauptplatine	motherboard	emaplaat	carte_mère	scheda_madre	základná_doska
	software	Software	software	tarkvara	logiciel	software	softvér
	hacker	Hacker	hacker	häkker	hacker	hacker	hacker
	antivirový_program	Antivirus	antivirus	viirusetõrje	antivirus	antivirus	antivírus
	aplikace	Арр	app	aplikatsioon	application	app	aplikácia
	hardware	Hardware	hardware	riistvara	hardware	hardware	hardvér
	kniha	Buch	book	raamat	livre	libro	kniha
	obrázek	Bild	picture	pilt	image	immagine	obrázok
	bas	Bassgitarre	bass_guitar	basskitarr	basse	basso_elettrico	bas
	mixér	Mixer	mixer	mikser	mixeur	frullatore	mixér
	tabule	Tafel	blackboard	tahvel	tableau	lavagna	tabuľa
	noviny	Zeitung	newspaper	ajaleht	journal	giornale	noviny
	brankář	Torwart	goalkeeper	väravavaht	gardien_de_but	portiere	brankár
	předmět	Subjekt	subject	teema	sujet	soggetto	predmet
	nouns	nouns	nouns	nouns	nouns	nouns	nouns
49	Internal_Body_	Internal_Body_	Internal_Body_	Internal_Body_	Internal_Body_	Internal_Body_	Internal_Body_
	Parts	Parts	Parts	Parts	Parts	Parts	Parts
	mozek	Gehirn	brain	aju	cerveau	cervello	mozog
	srdce	Herz	heart	süda	cœur	cuore	srdce
	močový měchýř	Blase	bladder	kusepõis	vessie	vescica	močový mechúr
	střevo	Darm	intestine	sool	intestin	intestino	črevo
	plíce	Lunge	lung	kops	poumon	polmone	pľúca
	ledvina	Niere	kidney	neer	rein	rene	oblička
	játra	Leber	liver	maks	foie	fegato	pečeň

	žaludek	Magen	stomach	magu	estomac	stomaco	žalúdok
	duše	Seele	soul	hing	âme	anima	duša
	mysl	Geist	mind	mõistus	esprit	mente	rozum
	noha	Fuß	foot	jalg	pied	piede	noha
	prst	Finger	finger	sõrm	doigt	dito	prst
	chuť_k_jídlu	Appetit	appetite	söögiisu	appétit	appetito	chuť_do_jedla
	zánět	Entzündung	inflammation	põletik	inflammation	infiammazione	zápal
	dům	Haus	house	maja	maison	casa	dom
	dřevo	Holz	wood	puit	bois	legno	drevo
	nouns						
50	Kitchenware						
	otvírák_na_konzervy	Dosenöffner	tin_opener	konserviavaja	ouvreboîte	apriscatole	otvárač na konzervy
	prkénko	Schneidebrett	cutting_board	lõikelaud	planche_à_découper	tagliere	doska_na_krájanie
	cedník	Sieb	strainer	sõel	passoire	colino	vývrtka
	šlehač	Schneebesen	whisk	vispel	fouet	frusta_elettrica	šľahač
	struhadlo	Reibe	grater	riiv	râpe	grattugia	strúhadlo
	váleček	Nudelholz	rolling_pin	taignarull	rouleau_à_pâtisserie	mattarello	valček
	naběračka	Schöpflöffel	ladle	kulp	louche	mestolo	naberačka
	louskáček	Nussknacker	nutcracker	uhmer	cassenoisette	schiaccianoci	luskáčik
	myčka	Spülmaschine	dishwasher	nõudepesumasin	lavevaisselle	lavastoviglie	umývačka
	lednice	Kühlschrank	fridge	külmik	réfrigérateur	frigorifero	chladnička
	těsto	Teig	dough	tainas	pâte	impasto	cesto
	šlehačka	Schlagsahne	whipped_cream	vahukoor	crème_fouettée	panna montata	šľahačka
	restaurace	Restaurant	restaurant	restoran	restaurant	ristorante	reštaurácia
	dezert	Nachtisch	dessert	magustoit	dessert	dessert	dezert
	časopis	Zeitschrift	magazine	ajakiri	magazine	rivista	časopis
	list	Blatt	leaf	leht	feuille	foglia	list
	nouns						
51	Landscape_ Features						

	jezero	See	lake	järv	lac	lago	jazero
	hora	Berg	mountain	mägi	montagne	montagna	hora
	kopec	Hügel	hill	küngas	colline	collina	kopec
	vodopád	Wasserfall	waterfall	juga	chute_d'_eau	cascata	vodopád
	údolí	Tal	valley	org	vallée	valle	údolie
	řeka	Fluss	river	jõgi	rivière	fiume	rieka
	ledovec	Gletscher	glacier	liustik	glacier	ghiacciaio	ľadovec
	pláň	Ebene	plain	tasandik	plaine	pianura	planina
	město	Stadt	city	linn	ville	città	mesto
	cesta	Straße	road	tee	route	strada	cesta
	strom	Baum	tree	puu	arbre	albero	strom
	kaluž	Pfütze	puddle	loik	flaque	pozzanghera	kaluž
	turistika	Wandern	hiking	matkamine	randonnée_pédestre	escursionismo	turistika
	dovolená	Urlaub	holiday	puhkus	vacances	vacanza	dovolenka
	podpatek	Absatz	heel	konts	talon	tacco	podpätok
	seznam	Liste	list	nimekiri	liste	lista	zoznam
	nouns	nouns	nouns	nouns	nouns	nouns	nouns
52	Languages	Languages	Languages	Languages	Languages	Languages	Languages
	němčina	Deutsch	German	saksa	allemand	tedesco	nemčina
	polština	Polnisch	Polish	poola	polonais	polacco	poľština
	angličtina	Englisch	English	inglise	anglais	inglese	angličtina
	italština	Italienisch	Italian	itaalia	italien	italiano	taliančina
	ruština	Russisch	Russian	vene	russe	russo	ruština
	čeština	Tschechisch	Czech	tšehhi	tchèque	ceco	čeština
	francouzština	Französisch	French	prantsuse	français	francese	francúzština
	holandština	Niederländisch	Dutch	hollandi	néerlandais	olandese	holandčina
	Java	Java	Java	Java	Java	Java	Java
	Python	Python	Python	Python	Python	Python	Python
	azbuka	Kyrilliza	Cyrillic_alphabet	vene_tähestik	alphabet_cyrillique	alfabeto_cirillico	azbuka
	latinka	Lateinschrift	Latin_alphabet	ladina_tähestik	alphabet_latin	alfabeto_latino	latinka
	dvojjazyčnost	Zweisprachigkeit	bilingualism	kakskeelsus	bilinguisme	bilinguismo	dvojjazyčnosť

	učení	Lernen	learning	õppimine	apprentissage	apprendimento	učenie
	slimák	Nacktschnecke	snail	nälkjas	limace	lumaca	slimák
	řasa	Wimper	eyelash	ripse	cil	ciglio	riasa
	nouns						
53	Linguistics						
	samohláska	Vokal	vowel	vokaal	voyelle	vocale	samohlásky
	slabika	Silbe	syllable	silp	syllabe	sillaba	slabika
	jazyk	Sprache	language	keel	langue	linguaggio	jazyk
	předpona	Vorsilbe	prefix	eesliide	préfixe	prefisso	prefix
	slovní_zásoba	Wortschatz	lexicon	sõnastik	lexique	lessico	slovná_zásoba
	věta	Satz	sentence	lause	phrase	frase	veta
	přízvuk	Betonung	stress	rõhk	accent	accento	prízvuk
	slovo	Wort	word	sõna	mot	parola	slovo
	kniha	Buch	book	raamat	livre	libro	kniha
	učitel	Lehrer	teacher	õpetaja	enseignant	insegnante	učiteľ
	škola	Schule	school	kool	école	scuola	škola
	dělení	Dividieren	division	jagamine	division	divisione	delenie
	prvočíslo	Primzahl	prime_number	algarv	nombre_premier	numero_primo	prvočíslo
	vlastnost	Eigenschaft	property	omadus	propriété	proprietà	vlastnosť
	královna	Königin	queen	kuninganna	reine	regina	kráľovná
	hostel	Herberge	hostel	hostel	auberge	ostello	hostel
	nouns						
54	Liquid_Containers						
	láhev	Flasche	bottle	pudel	bouteille	bottiglia	fľaša
	plechovka	Dose	tin	konserv	canette	lattina	plechovka
	sklenice	Glas	glass	klaas	verre	bicchiere	pohár
	zkumavka	Fläschchen	vial	pits	fiole	fiala	skúmavka
	čutora	Flachmann	flask	veepudel	gourde	borraccia	džbán
	kelímek	Weinglas	wine_glass	pokaal	calice	calice	kalich
	džbán	Krug	jug	kann	cruche	brocca	džbán
	hrnek	Becher	mug	kruus	tasse	tazza	hrnček

	balík	Paket	packet	pakend	colis	pacchetto	balík
	zásuvka	Schublade	drawer	sahtel	tiroir	cassetto	zásuvka
	pivo	Bier	beer	õlu	bière	birra	pivo
	víno	Wein	wine	vein	vin	vino	víno
	barman	Barkeeper	barman	baarmen	barman	barista	barman
	nápoj	Getränk	drink	jook	boisson	drink	nápoj
	hanba	Schande	shame	häbi	honte	vergogna	hanba
	hrdina	Held	hero	kangelane	héros	eroe	hrdina
	nouns	nouns	nouns	nouns	nouns	nouns	nouns
55	Materials	Materials	Materials	Materials	Materials	Materials	Materials
	kov	Metall	metal	metall	métal	metallo	kov
	sklo	Glas	glass	klaas	verre	vetro	sklo
	dřevo	Holz	wood	puit	bois	legno	drevo
	látka	Gewebe	cloth	riie	tissu	stoffa	látka
	papír	Papier	paper	paber	papier	carta	papier
	kůže	Leder	leather	nahk	cuir	cuoio	koža
	hliník	Aluminium	aluminum	alumiinium	aluminium	alluminio	hliník
	zlato	Gold	gold	kuld	or	oro	zlato
	voda	Wasser	water	vesi	eau	acqua	voda
	zemina	Boden	soil	muld	terre	terra	zemina
	strom	Baum	tree	puu	arbre	albero	strom
	zářez	Einschnitt	incision	sisselõige	entaille	incisione	zárez
	oděv	Kleider	clothes	riideese	vêtement	vestito	odev
	blok	Notizbuch	notebook	märkmik	cahier	taccuino	blok
	kufr	Koffer	suitcase	kohver	valise	valigia	kufor
	kruh	Kreis	circle	ring	cercle	cerchio	kruh
	nouns	nouns	nouns	nouns	nouns	nouns	nouns
56	Maths	Maths	Maths	Maths	Maths	Maths	Maths
	číslo	Zahl	number	number	nombre	numero	číslo
	násobení	Multiplikation	multiplication	korrutamine	multiplication	moltiplicazione	násobenie

zlomek	Bruch	fraction	murdosa	fraction	frazione	zlomok
desetinné_číslo	Dezimalzahl	decimal	arvutus	décimale	cifra	výpočet
nekonečno	Unendlichkeit	infinity	lõpmatus	infini	infinito	nekonečno
odmocnina	Quadratwurzel	square_root	ruutjuur	racine_carrée	radice	odmocnina
odečítání	Subtraktion	subtraction	lahutamine	soustraction	sottrazione	odčítanie
rovnice	Gleichung	equation	võrrand	équation	equazione	rovnica
obdélník	Rechteck	rectangle	ristkülik	rectangle	rettangolo	obdĺžnik
trojúhelník	Dreieck	triangle	kolmnurk	triangle	triangolo	trojuholník
kalkulačka	Taschenrechner	calculator	kalkulaator	calculatrice	calcolatrice	kalkulačka
informatika	Informatik	computer_science	informaatika	informatique	informatica	informatika
poznámka	Notiz	note	märkus	note	nota	poznámka
zkouška	Prüfung	exam	eksam	examen	esame	skúška
louka	Wiese	meadow	niit	pelouse	prato	lúka
bratranec	Cousin	cousin	nõbu	cousin	cugino	bratranec
nouns	nouns	nouns	nouns	nouns	nouns	nouns
7 Means_of_Transport	Means_of_Transport	Means_of_Transport	Means_of_Transport	Means_of_Transport	Means_of_Transport	Means_of_Transport
1						
motorka	Motorrad	motorbike	mootorratas	moto	motocicletta	motorka
motorka loď			mootorratas laev	moto navire	motocicletta nave	motorka loď
	Motorrad	motorbike ship car				
lod' auto	Motorrad Boot	ship	laev	navire voiture	nave	loď
loď	Motorrad Boot Auto	ship car	laev auto	navire	nave macchina	lod' auto
loď auto tramvaj	Motorrad Boot Auto Straßenbahn	ship car tram	laev auto tramm	navire voiture tramway	nave macchina tram	loď auto električka
lod' auto tramvaj autobus	Motorrad Boot Auto Straßenbahn Bus	ship car tram bus	laev auto tramm buss	navire voiture tramway bus	nave macchina tram bus	loď auto električka autobus
lod' auto tramvaj autobus vlak	Motorrad Boot Auto Straßenbahn Bus Zug	ship car tram bus train	laev auto tramm buss rong	navire voiture tramway bus train	nave macchina tram bus treno	loď auto električka autobus vlak
loď   auto   tramvaj   autobus   vlak   letadlo   vrtulník	Motorrad Boot Auto Straßenbahn Bus Zug Flugzeug	ship car tram bus train plane helicopter	laev auto tramm buss rong lennuk	navire voiture tramway bus train avion hélicoptère	nave macchina tram bus treno aeroplano elicottero	loď auto električka autobus vlak lietadlo vrtuľník
loď     auto     tramvaj     autobus     vlak     letadlo     vrtulník     rotoped	Motorrad Boot Auto Straßenbahn Bus Zug Flugzeug Hubschrauber	ship car tram bus train plane	laev auto tramm buss rong lennuk helikopter	navire voiture tramway bus train avion hélicoptère vélo_d'_appartement	nave macchina tram bus treno aeroplano elicottero cyclette	loď auto električka autobus vlak lietadlo vrtuľník rotoped
loď auto tramvaj autobus vlak letadlo vrtulník	Motorrad Boot Auto Straßenbahn Bus Zug Flugzeug Hubschrauber Heimtrainer Laufband	ship car tram bus train plane helicopter exercise_bike	laev auto tramm buss rong lennuk helikopter trenažöör	navire voiture tramway bus train avion hélicoptère	nave macchina tram bus treno aeroplano elicottero	loď auto električka autobus vlak lietadlo vrtuľník
loď     auto     tramvaj     autobus     vlak     letadlo     vrtulník     rotoped     běžecký_pás	Motorrad Boot Auto Straßenbahn Bus Zug Flugzeug Hubschrauber Heimtrainer	ship car tram bus train plane helicopter exercise_bike treadmill	laev auto tramm buss rong lennuk helikopter trenažöör jooksulint	navire voiture tramway bus train avion hélicoptère vélo_d'_appartement tapis_de_course	nave macchina tram bus treno aeroplano elicottero cyclette tapis_roulant	loď auto električka autobus vlak lietadlo vrtuľník rotoped bežecký_pás
loď     auto     tramvaj     autobus     vlak     letadlo     vrtulník     rotoped     běžecký_pás     chodník	Motorrad Boot Auto Straßenbahn Bus Zug Flugzeug Hubschrauber Heimtrainer Laufband Gehweg	ship car tram bus train plane helicopter exercise_bike treadmill pavement	laev auto tramm buss rong lennuk helikopter trenažöör jooksulint kõnnitee	navire voiture tramway bus train avion hélicoptère vélo_d'_appartement tapis_de_course trottoir	nave macchina tram bus treno aeroplano elicottero cyclette tapis_roulant marciapiede	loď auto električka autobus vlak lietadlo vrtuľník rotoped bežecký_pás chodník
loď     auto     tramvaj     autobus     vlak     letadlo     vrtulník     rotoped     běžecký_pás     chodník     cesta	Motorrad Boot Auto Straßenbahn Bus Zug Flugzeug Hubschrauber Heimtrainer Laufband Gehweg Straße	ship car tram bus train plane helicopter exercise_bike treadmill pavement road	laev auto tramm buss rong lennuk helikopter trenažöör jooksulint kõnnitee tee	navire voiture tramway bus train avion hélicoptère vélo_d'_appartement tapis_de_course trottoir route	nave macchina tram bus treno aeroplano elicottero cyclette tapis_roulant marciapiede strada	loď auto električka autobus vlak lietadlo vrtuľník rotoped bežecký_pás chodník cesta

	bota	Schuh	shoe	king	chaussure	scarpa	topánka
	nouns	nouns	nouns	nouns	nouns	nouns	nouns
58	Medicine	Medicine	Medicine	Medicine	Medicine	Medicine	Medicine
	lékař	Arzt	physician	raviarst	médecin	dottore	doktor
	pilulka	Pille	pill	tablett	pilule	pillola	tabletka
	injekční_stříkačka	Spritze	syringe	süstal	seringue	siringa	injekčná_striekačk
	chirurg	Chirurg	surgeon	kirurg	chirurgien	chirurgo	chirurg
	nemocnice	Krankenhaus	hospital	haigla	hôpital	ospedale	nemocnica
	steh	Naht	suture	õmblus	suture	sutura	steh
	rentgen	Röntgenuntersuchung	xray	röntgen	radiographie	radiografia	röntgen
	obvaz	Verband	bandage	side	bandage	benda	obväz
	řezník	Fleischer	butcher	lihunik	boucher	macellaio	mäsiar
	vězení	Gefängnis	prison	vangla	prison	prigione	väzenie
	paprsek	Strahl	sunbeam	päikesekiir	rayon_de_soleil	raggio	lúč
	pila	Säge	saw	saag	scie	sega	píla
	osobní_trenér	Personal_Trainer	personal_trainer	personaaltreener	coach_sportif	personal_trainer	osobný_tréner
	trénink	Ausbildung	training	koolitus	entraînement	allenamento	školenie
	pláž	Strand	beach	rand	plage	spiaggia	pláž
	email	EMail	email	ekiri	courriel	email	email
	nouns	nouns	nouns	nouns	nouns	nouns	nouns
59	Metals	Metals	Metals	Metals	Metals	Metals	Metals
	olovo	Blei	lead	plii	plomb	piombo	olovo
	železo	Eisen	iron	raud	fer	ferro	železo
	zlato	Gold	gold	kuld	or	oro	zlato
	zinek	Zink	zinc	tsink	zinc	zinco	zinok
	platina	Platin	platinum	plaatina	platine	platino	platina
	mĕď	Kupfer	copper	vask	cuivre	rame	meď'
	hliník	Aluminium	aluminium	alumiinium	aluminium	alluminio	hliník
	nikl	Nickel	nickel	nikkel	nickel	nichel	nikel
	kyslík	Sauerstoff	oxygen	hapnik	oxygène	ossigeno	kyslík

	vodík	Wasserstoff	hydrogen	vesinik	hydrogène	idrogeno	vodík
	hřebík	Nagel	nail	nael	clou	chiodo	klinec
	klíč	Schlüssel	key	võti	clé	chiave	kľúč
	tavení	Schmelze	melting	sulatamine	fusion	fusione	tavenie
	koroze	Korrosion	corrosion	korrosioon	corrosion	corrosione	korózia
	petržel	Petersilie	parsley	petersell	persil	prezzemolo	petržlen
	roura	Rohr	tube	toru	tube	tubo	potrubie
	nouns	nouns	nouns	nouns	nouns	nouns	nouns
60	Music	Music	Music	Music	Music	Music	Music
	nota	Note	note	noot	note	nota	nota
	písnička	Lied	song	laul	chanson	canzone	pieseň
	kytara	Gitarre	guitar	kitarr	guitare	chitarra	gitara
	rock	Rockmusik	rock	rock	rock	rock	rock
	flétna	Flöte	flute	flööt	flûte	flauto	flauta
	zvuk	Klang	sound	heli	piano	suono	zvuk
	mikrofon	Mikrofon	microphone	mikrofon	micro	microfono	mikrofón
	zpěvák	Sänger	singer	laulja	chanteur	cantante	spevák
	písmeno	Buchstabe	letter	kiri	lettre	lettera	písmeno
	barva	Farbe	colour	värv	couleur	colore	farba
	kresba	Zeichnung	drawing	joonistus	dessin	disegno	kresba
	socha	Skulptur	sculpture	skulptuur	sculpture	scultura	socha
	spisovatel	Schriftsteller	writer	kirjanik	écrivain	scrittore	spisovateľ
	malíř	Maler	painter	maalikunstnik	peintre	pittore	maliar
	piknik	Picknick	picnic	piknik	piquenique	picnic	piknik
	kapsa	Tasche	pocket	tasku	poche	tasca	vrecko
	nouns	nouns	nouns	nouns	nouns	nouns	nouns
61	Music_Genres	Music_Genres	Music_Genres	Music_Genres	Music_Genres	Music_Genres	Music_Genres
	hip_hop	Hip_Hop	hip_hop	hiphop	hip_hop	hip_hop	hip_hop
	jazz	Jazz	jazz	džäss	jazz	jazz	jazz
	reggae	Reggae	reggae	reggae	reggae	reggae	reggae
	country	Country	country	kantri	country	country	country

	rock	Rockmusik	rock	rokk	rock	rock	rock
	soul						
	folk						
	heavy_metal						
	balet	Ballett	ballet	ballett	ballet	balletto	balet
	opera	Oper	opera	ooper	opéra	opera	opera
	trubka	Trompete	trumpet	trompet	trompette	tromba	trúbka
	xylofon	Xylophon	xylophone	ksülofon	xylophone	xilofono	xylofón
	kapela	Band	band	bänd	groupe	band	kapela
	písničkář	Liedermacher	songwriter	laulukirjutaja	chansonnier	cantautore	skladateľ
	míra	Мав	measure	mõõt	mesure	misura	miera
	synovec	Neffe	nephew	vennapoeg	neveu	nipote	synovec
	nouns						
62	Musical_Instruments						
	harfa	Harfe	harp	harf	harpe	arpa	harfa
	housle	Geige	violin	viiul	violon	violino	husle
	kytara	Gitarre	guitar	kitarr	guitare	chitarra	gitara
	klavír	Klavier	piano	klaver	piano	pianoforte	klavír
	flétna	Flöte	flute	flööt	flute	flauto	flauta
	saxofon	Saxophon	saxophone	saksofon	saxophone	sassofono	saxofón
	bubny	Schlagzeug	drums	trummid	batterie	batteria	bubny
	kontrabas	Kontrabass	double_bass	kontrabass	contrebasse	contrabbasso	kontrabas
	sluchátka	Kopfhörer	headphones	kõrvaklapid	écouteurs	cuffie	slúchadlá
	reproduktor	Lautsprecher	speaker	kõlar	hautparleur	altoparlante	reproduktor
	rádio	Radio	radio	raadio	radio	radio	rádio
	zvonek	Klingel	doorbell	kell	sonnette	campanello	zvonček
	hudba	Musik	music	muusika	musique	musica	hudba
	noty	Partitur	sheet_music	partituur	partition	spartito	noty
	-						
	živý_plot	Hecke	hedgerow	hekk	haie	siepe	živý_plot
	-			hekk õhtu	haie soir	siepe sera	živý_plot večer

	nouns						
63	Nonalcoholic_Drinks						
	limonáda	Limonade	lemonade	limonaad	limonade	limonata	limonáda
	sodovka	Smoothie	smoothie	smuuti	smoothie	aranciata	odvar
	džus	Saft	juice	mahl	jus	succo	džús
	mléčný_koktejl	Apfelsaft	milkshake	jäätisekokteil	frappé	frappè	nealko
	ledový_čaj	Eistee	iced_tea	jäätee	thé_glacé	tè_freddo	bylinkový_čaj
	mléko	Milch	milk	piim	lait	latte	mlieko
	káva	Kaffee	coffee	kohv	café	caffè	káva
	čaj	Tee	tea	tee	thé	tisana	čaj
	pivo	Bier	beer	õlu	bière	birra	pivo
	víno	Wein	wine	vein	vin	vino	víno
	led	Eis	ice	jää	glace	ghiaccio	ľad
	citron	Zitrone	lemon	sidrun	citron	limone	citrón
	sklenice	Glas	glass	klaas	verre	bicchiere	sklo
	brčko	Strohhalm	straw	kõrs	paille	cannuccia	slamka
	žárlivost	Eifersucht	jealousy	armukadedus	jalousie	gelosia	žiarlivosť
	popis	Beschreibung	description	kirjeldus	description	descrizione	popis
	nouns						
64	Nuts						
	kešu	Cashew	cashew	india_pähkel	cajou	anacardi	kešu
	arašíd	Erdnuss	peanut	maapähkel	cacahuète	arachidi	arašid
	mandle	Mandel	almond	mandel	amande	mandorla	mandl'a
	lískový_ořech	Haselnuss	hazelnut	sarapuupähkel	noisette	nocciola	lieskový_oriešok
	vlašský_ořech	Walnuss	walnut	kreeka_pähkel	noix	noci	orech
	piniový_oříšek	Pinienkern	pine_nut	seedripähkel	pigne	pinoli	píniový_oriešok
	pistácie	Pistazie	pistachio	pistaatsiapähkel	pistache	pistacchio	pistácia
	jedlý_kaštan	Kastanie	chestnut	kastan	châtaigne	castagna	gaštan
	čočka	Linse	lentil	lääts	lentille	lenticchia	šošovica
	hrozinka	Rosine	raisin	rosin	raisin_sec	uvetta	hrozienko
	banán	Banane	banana	banaan	banane	banana	banán

	meloun	Wassermelone	watermelon	arbuus	pastèque	anguria	melón
	arašídové_máslo	Erdnussbutter	peanut_butter	maapähklivõi	beurre_de_cacahuète	burro_di_arachidi	arašidové_maslo
	medovník	Nougat	nougat	martsipan	nougat	torrone	medovník
	zavazadlo	Gepäck	luggage	pagas	bagage	bagaglio	batožina
	oběť	Opfer	victim	ohver	victime	vittima	obeť
	nouns	nouns	nouns	nouns	nouns	nouns	nouns
65	Office_Supplies	Office_Supplies	Office_Supplies	Office_Supplies	Office_Supplies	Office_Supplies	Office_Supplies
	nůžky	Schere	scissors	käärid	ciseaux	forbici	nožnice
	pravítko	Lineal	ruler	joonlaud	règle	righello	pravítko
	pero	Kugelschreiber	ball_pen	pastakas	stylo	penna_a_sfera	guľôčkové pero
	tužka	Stift	pencil	pliiats	crayon	matita	ceruzka
	guma	Radiergummi	eraser	kustukumm	gomme	gomma	guma
	ořezávátko	Spitzer	pencil_sharpener	pliiatsiteritaja	taillecrayon	temperino	strúhadlo
	fix	Marker	marker	vildikas	feutre	pennarello	fixka
	penál	Etui	pencil_case	pinal	trousse	astuccio	peračník
	psací_stůl	Schreibtisch	desk	kirjutuslaud	bureau	scrivania	stôl
	židle	Stuhl	chair	tool	chaise	sedia	stolička
	kávovar	Kaffeemaschine	coffee_machine	kohvimasin	machine_à_café	macchina_del_caffè	kávovar
	hrnek	Tasse	mug	kruus	mug	tazza	hrnček
	štětec	Pinsel	paint_brush	pintsel	pinceau	pennello	štetec
	plátno	Leinwand	canvas	lõuend	toile	tela	plátno
	knoflík	Knopf	button	nööp	bouton	bottone	gombík
	stuha	Schleife	ribbon	pael	ruban	fiocco	stuha
	nouns	nouns	nouns	nouns	nouns	nouns	nouns
66	Parts_of_Head	Parts_of_Head	Parts_of_Head	Parts_of_Head	Parts_of_Head	Parts_of_Head	Parts_of_Head
	čelo	Stirn	forehead	laup	front	fronte	čelo
	brada	Kinn	chin	lõug	menton	mento	brada
	nos	Nase	nose	nina	nez	naso	nos
	tvář	Wange	cheek	põsk	joue	guancia	líce
	ucho	Ohr	ear	kõrv	oreille	orecchio	ucho
	oko	Auge	eye	silm	œil	occhio	oko

<b>58</b>	Parts_of_Skeleton	Parts_of_Skeleton	Parts_of_Skeleton	Parts_of_Skeleton	Parts_of_Skeleton	Parts_of_Skeleton	Parts_of_Skeleton
	nouns	nouns	nouns	nouns	nouns	nouns	nouns
	náboženství	Religion	religion	religioon	religion	religione	náboženstvo
	tajemství	Geheimnis	secret	salajane	secret	segreto	tajomstvo
	garáž	Garage	garage	garaaž	garage	garage	garáž
	kuchyně	Küche	kitchen	köök	cuisine	cucina	kuchyňa
	vila	Stadthaus	townhouse	ridaelamu	villa	villa	vila
	byt	Wohnung	apartment	korter	appartement	appartamento	apartmán
	dlaždice	Fliese	tile	plaat	carreau	piastrella	dlaždica
	linoleum	Linoleum	linoleum	linoleum	linoléum	linoleum	linoleum
	okno	Fenster	window	aken		finestra	okno
	-		pillar	sammas	pilastre fenêtre	pilastro finastro	pilier
	schodiště pilíř	Treppe Säule	stairs	trepp	escaliers	scale	schodisko
	podlaha	Fußboden	floor	põrand	plancher	pavimento	podlaha
	fasáda	Vorderfront	facade	fassaad	façade	parete	fasáda
	dveře	Tür	door	uks	porte	porta	dvere
	střecha	Dach	roof	katus	toit	tetto	strecha
	stěna	Wand	wall	sein	mur	muro	stena
7	Parts_of_House	Parts_of_House	Parts_of_House	Parts_of_House	Parts_of_House	Parts_of_House	Parts_of_House
_	nouns	nouns	nouns	nouns	nouns	nouns	nouns
	rodokmen	Stammbaum	pedigree	sugupuu	pedigree	pedigree	rodokmeň
	tenisák	Tennisball	tennis_ball	tennisepall	balle_de_tennis	pallina_da_tennis	tenis
	oční_stíny	Lidschatten	eye_shadow	lauvärvid	fard_à_paupières	ombretto	líčenie
	rtěnka	Lippenstift	lipstick	huulepulk	rouge_à_lèvres	rossetto	rúž
	náušnice	Ohrring	earring	kõrvarõngas	boucle_d'_oreille	orecchino	náušnica
	klobouk	Hut	hat	müts	chapeau	cappello	čiapka
	prst	Finger	finger	sõrm	doigt	dito	prst
	zadek	Rücken	butt	tagumik	derrière	sedere	zadok
	ústa	Mund	mouth	suu	bouche	bocca	ústa

	hrudní_kost	Brustbein	breastbone	kolju	sternum	sterno	lebka
	čelist	Oberkiefer	jawbone	lõualuu	mâchoire	mascella	sánka
	stehenní_kost	Oberschenkelknochen	femur	reieluu	fémur	femore	rebro
	kostrč	Steißbein	соссух	selgroog	соссух	coccige	chrbtica
	lopatka	Schulterblatt	shoulder_blade	abaluu	omoplate	scapola	lopatka
	pánevní_kost	Schlüsselbein	collarbone	rangluu	clavicule	clavicola	kostrč
	klíční_kost	Wadenbein	fibula	sääreluu	péroné	perone	ihlica
	holenní_kost	Schienbein	tibia	ribi	tibia	tibia	píšťala
	kloub	Gelenk	joint	liiges	articulation	articolazione	kĺb
	chrupavka	Knorpel	cartilage	kõhr	cartilage	cartilagine	chrupka
	krev	Blut	blood	veri	sang	sangue	krv
	sval	Muskel	muscle	lihas	muscle	muscolo	sval
	zlomenina	Fraktur	fracture	murd	fracture	frattura	zlomenina
	vykloubení	Verrenkung	dislocation	nihestus	luxation	lussazione	vykĺbenie
	plíseň	Schimmel	mould	hallitus	moisissure	muffa	pleseň
	koš_na_odpadky	Papierkorb	bin	prügikorv	bac_de_recyclage	cestino	kôš_na_odpadky
	nouns	nouns	nouns	nouns	nouns	nouns	nouns
69	Parts_of_Speech	Parts_of_Speech	Parts_of_Speech	Parts_of_Speech	Parts_of_Speech	Parts_of_Speech	Parts_of_Speech
	podstatné_jméno	Substantiv	noun	nimisõna	nom	nome	podstatné_meno
	přídavné_jméno	Adjektiv	adjective	omadussõna	adjectif	aggettivo	prídavné_meno
	zájmeno	Pronomen	pronoun	asesõna	pronom	pronome	zámeno
	číslovka	Zahlwort	numeral	arvsõna	numéral	numerale	číslovka
	sloveso	Verb	verb	tegusõna	verbe	verbo	sloveso
	spojka	Konjunktion	conjunction	sidesõna	conjonction	congiunzione	spojka
	příslovce	Adverb	adverb	määrsõna	adverbe	avverbio	príslovka
	citoslovce	Interjektion	interjection	hüüdsõna	interjection	interiezione	citoslovce
	jednotné_číslo	Singular	singular	ainsus	singulier	singolare	jednotné číslo
	čas	Zeitform	tense	aeg	temps	tempo	čas
	časování	Konjugation	conjugation	pööramine	conjugaison	coniugazione	časovanie
	skloňování	Deklination	inflection	käänamine	déclinaison	declinazione	skloňovanie

	souhláska	Konsonant	consonant	konsonant	consonne	consonante	spoluhláska
	nerv	Nerv	nerve	närv	nerf	nervo	nerv
	párek	Bratwurst	hot_dog	viiner	saucisse	salsiccia	párok
	nouns	nouns	nouns		nouns	nouns	nouns
70	Politics	Politics	Politics	Politics	Politics	Politics	Politics
	volby	Wahl	election	valimised	élection	elezione	voľby
	vláda	Regierung	government	valitsus	gouvernement	governo	vláda
	strana	Partei	party	partei	parti	partito	strana
	hlasování	Abstimmung	poll	küsitlus	vote	votazione	hlasovanie
	parlament	Parlament	parliament	parlament	parlement	parlamento	parlament
	hlasovací_lístek	Stimmzettel	ballot	hääletussedel	bulletin_de_vote	scheda_elettorale	hlasovací_lístok
	republika	Republik	republic	vabariik	république	repubblica	republika
	premiér	Ministerpräsident	prime_minister	peaminister	premier_ministre	primo_ministro	premiér
	vězení	Gefängnis	prison	vangla	prison	prigione	väzenie
	kreditní_karta	Kreditkarte	credit_card	krediitkaart	carte_de_crédit	carta_di_credito	kreditná_karta
	vedoucí	Chef	boss	boss	directeur	direttore	vedúci
	soutěž	Wettbewerb	contest	konkurss	compétition	competizione	súťaž
	tým	Mannschaft	team	meeskond	équipe	squadra	tím
	voják	Soldat	soldier	sõdur	soldat	soldato	vojak
	lehátko	Liegestuhl	deckchair	lamamistool	transat	sdraio	polohovateľné_kreslo
	čára	Linie	line	joon	ligne	linea	čiara
	nouns	nouns	nouns	nouns	nouns	nouns	nouns
71	Professions	Professions	Professions	Professions	Professions	Professions	Professions
	hasič	Feuerwehrmann	firefighter	tuletõrjuja	pompier	pompiere	hasič
	doktor	Arzt	doctor	arst	médecin	medico	doktor
	kovář	Schmied	smith	sepp	forgeron	fabbro	kováč
	prodavač	Verkäufer	seller	müüja	vendeur	venditore	predavač
	kuchař	Koch	cook	kokk	cuisinier	cuoco	kuchár
	švadlena	Schneider	dressmaker	õmbleja	couturier	sarto	krajčírka
	právník	Rechtsanwalt	lawyer	advokaat	avocat	avvocato	právnik
	policista	Polizist	police_officer	politseinik	policier	poliziotto	policajt

	dospělý	Erwachsene	adult	täiskasvanu	adulte	adulto	dospelý
	puberťák	Teenager	teenager	teismeline	adolescent	adolescente	adolescent
	dobrovolník	Freiwillige	volunteer	volontäär	volontaire	volontario	dobrovoľník
	tchyně	Schwiegermutter	mother_in_law	ämm	bellemère	suocera	svokra
	práce	Arbeit	job	töö	travail	lavoro	práca
	kancelář	Büro	office	büroo	bureau	ufficio	kancelária
	deník	Tagebuch	diary	päevik	journal	diario	denník
	místo	Platz	place	koht	place	luogo	miesto
	nouns						
2	Reptiles						
	krokodýl	Krokodil	crocodile	krokodill	crocodile	coccodrillo	krokodíl
	had	Schlange	snake	madu	serpent	serpente	had
	ještěrka	Eidechse	lizard	sisalik	lézard	lucertola	jašterica
	želva	Schildkröte	turtle	kilpkonn	tortue	tartaruga	korytnačka
	gekon	Gecko	gecko	geko	gecko	geco	gekón
	chameleon	Chamäleon	chameleon	kameeleon	caméléon	camaleonte	chameleón
	leguán	Leguan	iguana	iguaan	iguane	iguana	leguán
	aligátor	Alligator	alligator	alligaator	alligator	alligatore	aligátor
	žába	Frosch	frog	konn	grenouille	rana	žaba
	hroch	Nilpferd	hippo	jõehobu	hippopotame	ippopotamo	hroch
	červ	Wurm	worm	uss	ver	verme	červ
	pavouk	Spinne	spider	ämblik	araignée	ragno	pavúk
	syčení	Zischen	hiss	kahin	sifflement	sibilo	syčanie
	krunýř	Häutung	molt	nahk	mue	muta	pancier
	škrábanec	Kratzer	scratch	kriimustus	égratignure	graffio	škrabanec
	rukáv	Ärmel	sleeve	varrukas	manche	manica	rukáv
	nouns						
'3	Road_Means_of_ Transport						
	auto	Auto	car	auto	voiture	automobile	auto

	autobus	Bus	bus	buss	bus	autobus	autobus
	taxi	Taxi	taxi	takso	taxi	taxi	taxi
	jízdní_kolo	Fahrrad	bike	jalgratas	vélo	bicicletta	bicykel
	motocykl	Motorrad	motorbike	mootorratas	moto	moto	motocykel
	trolejbus	Obus	trolleybus	trollibuss	trolleybus	filobus	trolejbus
	dodávka	Lieferwagen	van	kaubik	camionnette	furgone	van
	skútr	Motorroller	scooter	roller	scooter	scooter	skúter
	brusle	Schlittschuh	skates	rulluisud	patins	pattini	korčule
	letadlo	Flugzeug	airplane	lennuk	avion	aeroplano	lietadlo
	cesta	Straße	road	maantee	route	strada	cesta
	kruhový_objezd	Kreisverkehr	roundabout	ringristmik	rondpoint	rotonda	kruhový_objazd
	provoz	Verkehr	traffic	liiklus	trafic	traffico	premávka
	dopravní_nehoda	Verkehrsunfall	car_crash	autoõnnetus	accident	incidente_stradale	dopravná_nehoda
	ragby	Rugby	rugby	ragbi	rugby	rugby	rugby
	topinkovač	Toaster	toaster	röster	grillepain	tostapane	hriankovač
	nouns						
74	Rooms_in_the_ House						
	kuchyně	Küche	kitchen	köök	cuisine	cucina	kuchyňa
	toaleta	Toilette	toilet	tualett	toilettes	gabinetto	toaleta
	obývací_pokoj	Wohnzimmer	living_room	elutuba	salon	salotto	obývačka
	koupelna	Badezimmer	bathroom	vannituba	salle_de_bain	bagno	kúpeľňa
	jídelna	Esszimmer	dining_room	söögituba	salle_à_manger	sala_da_pranzo	jedáleň
	dětský_pokoj	Kinderzimmer	utility_room	lastetuba	buanderie	ripostiglio	detská_izba
	ložnice	Schlafzimmer	bedroom	magamistuba	chambre_à_coucher	camera_da_letto	spálňa
	sklep	Keller	cellar	kelder	cave	cantina	pivnica
	vagón	Wagen	wagon	vagun	wagon	vagone	vagón
	pilotní_kabina	Cockpit	cockpit	kokpit	cockpit	cabina_di_pilotaggio	pilotná_kabína
	koncertní_sál	Konzertsaal	concert_hall	kontserdisaal	salle_de_concert	sala_concerti	koncertná_sála
	Koneertin_5u						
	recepce	Rezeption	reception	retseptsioon	réception	reception	recepcia

	sporák	Kocher	cooker	pliit	cuisinière	fornello	sporák
	poznámka	Anmerkung	comment	märkus	remarque	commento	poznámka
	student	Student	student	tudeng	étudiant	studente	študent
	nouns	nouns	nouns	nouns	nouns	nouns	nouns
75	Savanna_Animals	Savanna_Animals	Savanna_Animals	Savanna_Animals	Savanna_Animals	Savanna_Animals	Savanna_Animals
	lev	Löwe	lion	lõvi	lion	leone	lev
	slon	Elefant	elephant	elevant	éléphant	elefante	slon
	gazela	Gazelle	gazelle	gasell	gazelle	gazzella	gazela
	leopard	Leopard	leopard	leopard	léopard	leopardo	leopard
	hroch	Nilpferd	hippo	jõehobu	hippopotame	ippopotamo	hroch
	zebra	Zebra	zebra	sebra	zèbre	zebra	zebra
	hyena	Hyäne	hyena	hüään	hyène	iena	hyena
	žirafa	Giraffe	giraffe	kaelkirjak	girafe	giraffa	žirafa
	medvěd	Bär	bear	karu	ours	orso	medveď
	zajíc	Hase	hare	jänes	lièvre	lepre	zajac
	křeček	Hamster	hamster	hamster	hamster	criceto	škrečok
	kočka	Katze	cat	kass	chat	gatto	mačka
	pytlák	Wilderer	poacher	salakütt	braconnier	bracconiere	pytliak
	safari	Safari	safari	safari	safari	safari	safari
	rovnováha	Gleichgewicht	balance	tasakaal	équilibre	equilibrio	rovnováha
	anténa	Antenne	antenna	antenn	antenne	antenna	anténa
	nouns	nouns	nouns	nouns	nouns	nouns	nouns
76	School_Subjects	School_Subjects	School_Subjects	School_Subjects	School_Subjects	School_Subjects	School_Subjects
	geografie	Geografie	geography	geograafia	géographie	geografia	geografia
	fyzika	Physik	physics	füüsika	physique	fisica	fyzika
	chemie	Chemie	chemistry	keemia	chimie	chimica	chémia
	biologie	Biologie	biology	bioloogia	biologie	biologia	biológia
	tělesná_výchova	Sportunterricht	physical_education	kehaline_kasvatus	éducation_physique	educazione_fisica	telesná_výchova
	cizí_jazyk	Fremdsprache	foreign_language	võõrkeel	langue_étrangère	lingua_straniera	cudzí_jazyk
	výtvarná_výchova	Kunstvermittlung	art_education	kunstiõpetus	éducation_artistique	educazione_artistica	výtvarná_výchova
	dějepis	Geschichte	history	ajalugu	histoire	storia	dejepis

	skaut	Pfadfinder	scouting	skaut	scout	scout	skaut
	keramika	Keramik	ceramics	keraamika	céramique	ceramica	keramika
	přestávka	Pause	break	vaheaeg	pause	intervallo	prestávka
	oběd	Mittagessen	lunch	lõuna	déjeuner	pranzo	obed
	doučování	Nachhilfe	tutoring	õpetamine	cours_particuliers	ripetizioni	doučovanie
	učitel	Lehrer	teacher	õpetaja	enseignant	insegnante	učiteľ
	vařečka	Kochlöffel	wooden_spoon	puulusikas	cuillère_en_bois	cucchiaio_di_legno	varecha
	most	Brücke	bridge	sild	pont	ponte	most
	nouns	nouns	nouns	nouns	nouns	nouns	nouns
17	Shapes	Shapes	Shapes	Shapes	Shapes	Shapes	Shapes
	krychle	Würfel	cube	kuup	cube	cubo	kocka
	trojúhelník	Dreieck	triangle	kolmnurk	triangle	triangolo	trojuholník
	koule	Kugel	sphere	kera	sphère	sfera	gul'a
	obdélník	Rechteck	rectangle	ristkülik	rectangle	rettangolo	obdĺžnik
	kosočtverec	Rhombus	rhombus	romb	losange	rombo	kosoštvorec
	jehlan	Pyramide	pyramid	püramiid	pyramide	piramide	ihlan
	kužel	Kegel	cone	koonus	cône	cono	kužeľ
	válec	Zylinder	cylinder	silinder	cylindre	cilindro	valec
	obvod	Umfang	perimeter	ümbermõõt	périmètre	perimetro	obvod
	úhel	Winkel	angle	nurk	angle	angolo	uhol
	geometrie	Geometrie	geometry	geomeetria	géométrie	geometria	geometria
	matematika	Mathematik	maths	matemaatika	mathématiques	matematica	matematika
	objem	Volumen	volume	maht	volume	volume	objem
	kružítko	Zirkel	compass	sirkel	compas	compasso	kružidlo
	mušle	Muschel	shell	merekarp	coquille	conchiglia	mušľa
	vlajka	Flagge	flag	lipp	drapeau	bandiera	vlajka
	nouns	nouns	nouns	nouns	nouns	nouns	nouns
78	Shoes	Shoes	Shoes	Shoes	Shoes	Shoes	Shoes
	lodičky	Pumps	high_heels	ketsid	talons_hauts	scarpe_col_tacco	lodičky
	holínky	Gummistiefel	rain_boots	kummikud	bottes_de_pluie	stivali_da_pioggia	gumáky

sandály	Sandalen	sandals	sandaalid	sandales	sandali	sandále
boty	Stiefel	boots	saapad	bottes	stivali	čižmy
tenisky	Turnschuhe	sneakers	tossud	sneakers	sneakers	tenisky
sportovní_boty	Sportschuhe	trainers	tennised	baskets	scarpe_da_tennis	drevák
žabky	Flipflops	flipflops	plätud	tongs	infradito	žabky
bačkory	Hausschuhe	slippers	sussid	chaussons	ciabatte	papuče
punčochy	Strumpfhose	stockings	sukk	bas	collant	pančuchy
ponožka	Socke	sock	sokk	chaussette	calza	ponožka
kalhoty	Hose	trousers	püksid	pantalon	pantaloni	nohavice
džíny	Jeans	jeans	teksad	jeans	jeans	džínsy
procházka	Treck	walk	jalutuskäik	randonnée	camminata	prechádzka
běh	Lauf	run	jooks	course	corsa	beh
klec	Käfig	cage	puur	cage	gabbia	klietka
krása	Schönheit	beauty	ilu	beauté	bellezza	krása
nouns	nouns	nouns	nouns	nouns	nouns	nouns
Shops	Shops	Shops	Shops	Shops	Shops	Shops
lékárna	Apotheke	pharmacy	apteek	pharmacie	farmacia	lekáreň
knihkupectví	Buchhandlung	book_shop	raamatupood	librairie	libreria	kníhkupectvo
trafika	Kiosk	tobacco_shop	tubakapood	kiosque	tabaccheria	trafika
butik	Bekleidungsgeschäft	clothes_store	riidepood	magasin_de_vêtements	negozio_di_vestiti	butik
supermarket	Supermarkt	supermarket	supermarket	supermarché	supermercato	supermarket
řeznictví	Fleischerei	butcher_shop	lihapood	boucherie	macelleria	mäsiarstvo
cukrárna	Lebensmittelgeschäft	grocery	toidupood	épicerie	ortofrutta	cukráreň
pekárna	Bäckerei	bakery	pagariäri	boulangerie	panificio	pekáreň
kino	Kino	cinema	kino	cinéma	cinema	kino
stanice	Bahnhof	station	jaam	gare	stazione	stanica
cigareta	Zigarette	cigarette	sigaret	cigarette	sigaretta	cigareta
nákup	Einkaufen	shopping	šoppamine	achat	spesa	nakupovanie
	0.10	receipt	kviitung	reçu	scontrino	účtenka
účtenka	Quittung	receipt	Kvintung	1030		
účtenka sleva	Rabatt	discount	allahindlus	rabais	sconto	zľava

	tajemník	Sekretär	clerk	müüja	secrétaire	segretario	tajomník
	nouns	nouns	nouns	nouns	nouns	nouns	nouns
80	Sources_of_Energy	Sources_of_Energy	Sources_of_Energy	Sources_of_Energy	Sources_of_Energy	Sources_of_Energy	Sources_of_Energy
	ropa	Öl	oil	õli	pétrole	petrolio	ropa
	jaderná_energie	Kernkraft	nuclear_power	tuumaenergia	énergie_nucléaire	energia_nucleare	jadrová_energia
	uhlí	Kohle	coal	kivisüsi	charbon	carbone	uhlie
	plyn	Gas	gas	gaas	gaz	gas_naturale	plyn
	metan	Methan	methane	metaan	méthane	metano	metán
	větrná_energie	Windenergie	wind_power	tuuleenergia	énergie_éolienne	energia_eolica	veterná_energia
	sluneční_energie	Sonnenenergie	solar_energy	päikeseenergia	énergie_solaire	energia_solare	solárna_energia
	vodní_energie	Wasserkraft	hydropower	hüdroenergia	hydroélectricité	energia_idroelettrica	vodná_energia
	teplo	Wärme	heat	soojus	chaleur	calore	teplo
	elektřina	Elektrizität	electricity	elekter	électricité	elettricità	elektrina
	čerpací_stanice	Tankstelle	petrol_station	bensiinijaam	stationservice	benzinaio	čerpacie_stanice
	elektrárna	Kraftwerk	power_plant	elektrijaam	centrale_électrique	centrale_elettrica	elektráreň
	těžba	Gewinnung	extraction	kaevandamine	extraction	estrazione	baníctvo
	atomová_bomba	Kernspaltung	nuclear_fission	tuumalõhustumine	fission_nucléaire	fissione_nucleare	atómová_bomba
	polibek	Kuss	kiss	suudlus	bise	bacio	bozk
	batoh	Rucksack	backpack	seljakott	sac_à_dos	zaino	batoh
	nouns	nouns	nouns	nouns	nouns	nouns	nouns
31	Spices	Spices	Spices	Spices	Spices	Spices	Spices
	hřebíček	Nelken	cloves	nelk	clous_de_girofle	chiodi_di_garofano	klinček
	pepř	Pfeffer	pepper	pipar	poivre	рере	čierne_korenie
	skořice	Zimt	cinnamon	kaneel	cannelle	cannella	škorica
	šafrán	Safran	saffron	safran	safran	zafferano	šafran
	zázvor	Ingwer	ginger	ingver	gingembre	zenzero	zázvor
	paprika	Paprika	paprika	paprika	paprika	paprika	paprika
	kurkuma	Kurkuma	turmeric	kurkum	curcuma	curcuma	kurkuma
	muškátový_oříšek	Muskatnuss	nutmeg	muskaatpähkel	muscade	noce_moscata	muškátový_oriešok
	sůl	Salz	salt	sool	sel	sale	soľ

	cukr	Zucker	sugar	suhkur	sucre	zucchero	cukor
	med	Honig	honey	mesi	miel	miele	med
	mandle	Mandeln	almond	mandel	amande	mandorla	mandl'a
	chod	Mahlzeit	course	käik	mets	portata	chod
	recept	Rezept	recipe	retsept	recette	ricetta	recept
	řetěz	Kette	chain	kett	chaîne	catena	reťaz
	abeceda	Alphabet	alphabet	tähestik	alphabet	alfabeto	abeceda
	nouns	nouns	nouns	nouns	nouns	nouns	nouns
82	Spirits	Spirits	Spirits	Spirits	Spirits	Spirits	Spirits
04	gin	Gin	gin	džinn	gin	gin	gin
	whisky	Whiskey	whisky	viski	whisky	whisky	whisky
	vodka	Wodka	vodka	viin	vodka	vodka	vodka
	rum	Rum	rum	rumm	rhum	rum	rum
	víno	Wein	wine	vein	vin	vino	víno
	pivo	Bier	beer	õlu	bière	birra	pivo
	tequila	Tequila	tequila	tekiila	tequila	tequila	tequila
	koňak	Cognac	-	konjak	-	-	koňak
	Kollak	Cognac	cognac	KOIIJAK	cognac	cognac	KOIIAK
	voda	Wasser	water	vesi	eau	acqua	voda
	džus	Saft	juice	mahl	jus	succo	džús
	čaj	Tee	tea	tee	thé	tè	čaj
	káva	Kaffee	coffee	kohv	café	caffè	káva
	brambor	Kartoffel	potato	kartul	pomme_de_terre	patata	zemiak
	ječmen	Gerste	barley	oder	orge	orzo	jačmeň
	kněžna	Prinzessin	princess	printsess	princesse	principessa	kňažná
	park	Park	park	park	parc	parco	park
	nouns	nouns	nouns	nouns	nouns	nouns	nouns
83	Sport	Sport	Sport	Sport	Sport	Sport	Sport
	raketa	Schläger	racket	reket	raquette	racchetta	raketa
	házená	Handball	handball	käsipall	handball	pallamano	hádzaná
	sportovec	Sportler	athlete	sportlane	athlète	atleta	športovec
	rozhodčí	Schiedsrichter	referee	kohtunik	arbitre	arbitro	rozhodca
tenis	Tennis	tennis	tennis	tennis	tennis	tenis	
-------------	---	--	---	--	--	---	
balon	Ball	ball	pall	ballon	pallone	lopta	
zápas	Spiel	match	matš	match	partita	zápas	
lyžování	Ski	ski	suusatamine	ski	sci	lyžovanie	
schovávaná	Verstecken	hide_and_seek	peitus	cachecache	nascondino	schovávačka	
deskovka	Brettspiel	board_game	lauamäng	jeu_de_société	gioco_da_tavolo	stolová_hra	
palice	Fischernetz	small_net	võrk	épuisette	retino	palice	
naběračka	Schöpfkelle	ladle	kulp	louche	mestolo	naberačka	
číšník	Kellner	waiter	kelner	garçon	cameriere	čašník	
zahradník	Gärtner	gardener	aednik	jardinier	giardiniere	záhradník	
rýč	Spaten	spade	labidas	bêche	vanga	rýľ	
hrášek	Erbse	pea	hernes	pois	pisello	hrášok	
nouns	nouns	nouns	nouns	nouns	nouns	nouns	
Sports	Sports	Sports	Sports	Sports	Sports	Sports	
atletika	Leichtathletik	athletics	kergejõustik	athlétisme	atletica_leggera	atletika	
gymnastika	Kunstturnen	artistic_gymnastics	riistvõimlemine	gymnastique_artistique	ginnastica_artistica	gymnastika	
plavání	Schwimmen	swimming	ujumine	natation	nuoto	plávanie	
	Skilanglauf	crosscountry_skiing	murdmaasuusatamine			beh_na_lyžiach	
		cycling	jalgrattasõit	cyclisme	ciclismo	cyklistika	
bruslení	Schlittschuhlaufen	, e	uisutamine	patinage	pattinaggio	korčuľovanie	
volejbal	Volleyball	volleyball	võrkpall	volley	pallavolo	volejbal	
windsurfing	Windsurfen	windsurf	purjelauasõit	planche_à_voile	windsurf	windsurfing	
jogging	Joggen	jogging	jooksmine	jogging	jogging	jogging	
turistika	Wandern	hiking	matkamine	randonnée_pédestre	escursionismo	turistika	
uniforma	Uniform	uniform	vormiriietus	uniforme	divisa	uniforma	
míč	Ball	ball	pall	ballon	palla	lopta	
sportovec	Sportler	athlete	sportlane	athlète	atleta	športovec	
tým	Mannschaft	team	meeskond	équipe	squadra	tím	
šíp	Pfeil	arrow	nool	flèche	freccia	šíp	
	Dieb	thief		voleur	ladro	zlodej	
	balon zápas lyžování schovávaná deskovka palice naběračka číšník zahradník rýč hrášek <b>nouns</b> <b>Sports</b> atletika gymnastika plavání běh_na_lyžích cyklistika bruslení volejbal windsurfing jogging turistika uniforma míč sportovec	balonBallzápasSpiellyžováníSkischovávanáVersteckendeskovkaBrettspielpaliceFischernetznaběračkaSchöpfkellečíšníkKellnerzahradníkGärtnerrýčSpatenhrášekErbsenounsnounssportsSportsatletikaLeichtathletikgymnastikaKunstturnenběh_na_lyžíchSkilanglaufcyklistikaRadfahrenbrusleníSchlittschuhlaufenvolejbalVolleyballwindsurfingJoggenturistikaWandernuniformaUniformmíčBallsportovecSportler	balonBallballzápasSpielmatchlyžováníSkiskilyžováníSkiskischovávanáVersteckenhide_and_seekdeskovkaBrettspielboard_gamepaliceFischernetzsmall_netnaběračkaSchöpfkelleladlečíšníkKellnerwaiterzahradníkGärtnergardenerrýčSpatenspadehrášekErbsepeanounsnounsnounssoportsSportsSportsatletikaLeichtathletikathleticsgymnastikaKunstturnenswimmingběh_na_lyžíchSkilanglaufcrosscountry_skiingvolejbalVolleyballvolleyballvolejbalVolleyballvolleyballwindsurfingJoggenjoggingturistikaBallballsportovecSportlerathlete	balonBallballpallzápasSpielmatchmatšlyžováníSkiskisuusatamineschovávanáVersteckenhide_and_seekpeitusdeskovkaBrettspielboard_gamelauamängpaliceFischernetzsmall_netvörknaběračkaSchöpfkelleladlekulpčíšníkKellnerwaiterkelnerzahradníkGärtnergardeneraednikrýčSpatenspadelabidashrášekErbsepeahernesnounsnounsnounsnounsSportsSportsSportsSportsgymnastikaKunstturnenartistic_gymnasticsristvöinlemineplaváníSchlittschuhlaufenskatingujuminevolejbalVolleyballvolleyballvörkpallvolejbalVolleyballvolleyballvörkpallwindsurfingJoggenjoggingjooksmineuniformaUniforminiformvormirietusmičBallballpallsportovecSportlerathletesportlane	balonBallballpallballonzápasSpielmatchmatšmatchlyžováníSkiskisuusatamineskilyžováníSkiskisuusatamineskischovávanáVersteckenhide_and_seekpeituscachecachedeskovkaBrettspielboard_gamelauamängjeu_de_sociétépaliceFischernetzsmall_netvõrképuisettenaběračkaSchöpfkelleladlekulplouchečišníkKellnerwaiterkelnergarçonzahradníkGärtnergardeneraednikjardinierrýčSpatenspadelabidasbêchehrášekErbsepeahernespoisnounsnounsnounsnounsnounsSportsSportsSportsSportssportsstatetikaLeichtathletikathleticskergejõustikathlétismegymnastikaKunstrurnenartistic_gymnasticsriistvõimleminegymnastique_artistiqueplaváníSchwimmenswimmingujuminenatationběh_na_lyžichSkilanglaufcrosscountry_skiigmurdmaasustatamineski_de_fondcyklistkaRadfahrencyclingjalgrattasõitcyclismejoggingJoggenjoggingjooksminejoggingjoggingturistikaWandernhikingmatkaminerandonnée_pédestreuniformaUniformuniformvoririie	balon     Ball     ball     pall     ballon     pallone       zápas     Spiel     match     matš     match     parita       lyžování     Ski     ski     suustamine     ski     sci       schovávaná     Verstecken     hide_and_seek     peitus     cachecache     nascondino       glace     Fischernetz     small_net     võrk     épuisette     retino       naběračka     Schöpfkelle     lalde     kulp     louche     mestolo       číšnik     Kellner     waiter     kelner     garçon     camericre       zahradník     Gärtner     gardener     aednik     jardinier     giardiniere       rýč     Spaten     spade     labidas     běche     vanga       nouns     nouns     nouns     nouns     nouns     nouns       sports     Sports     Sports     Sports     Sports     Sports       symastika     Kunsturnen     artistic_gymnastics     ristvoimlemine     ski_de_fond       gymnastika     Kunsturnen     artistic_gymnastics     ristvoimlemine     sci_di_fondo       cyklistika     Radfahren     cycling     jalgratasöit     cyclisme     ciclismo       plavání     Schiubrahuhalen     skating	

	nouns						
85	Sweets						
	bonbón	Bonbon	praline	kompvek	bonbon	caramella	cukrík
	cukrová_vata	Zuckerwatte	cotton_candy	suhkruvatt	barbe_à_papa	zucchero_filato	cukrová_vata
	sušenka	Biskuit	biscuit	küpsis	biscuit	biscotto	sušienka
	bábovka	Käsekuchen	cheesecake	juustukook	dragée	confetto	koláč
	kobliha	Gebäck	pastry	kondiitritoode	pâtisseries	pasticcino	oblátka
	dort	Kuchen	cake	kook	tarte	torta	torta
	pudink	Pudding	custard	piparkook	gâteau	budino	puding
	čokoláda	Schokolade	chocolate	šokolaad	chocolat	cioccolato	čokoláda
	chléb	Brot	bread	leib	pain	pane	chlieb
	pizza	Pizza	pizza	pitsa	pizza	pizza	pizza
	švestka	Pflaume	plum	ploom	prune	prugna	slivka
	cukr	Zucker	sugar	suhkur	sucre	zucchero	cukor
	cukrář	Konditor	pastry_chef	kondiiter	pâtissier	pasticciere	cukrárka
	zubař	Zahnarzt	dentist	hambaarst	dentiste	dentista	zubár
	koncert	Konzert	concert	kontsert	concert	concerto	koncert
	kost	Knochen	bone	luu	OS	OSSO	kosť
	adjectives						
86	Temperature_ Features						
	teplý	heiß	hot	kuum	chaud	caldo	teplý
	studený	kalt	cold	külm	froid	freddo	studený
	vlažný	lau	warm	soe	tiède	tiepido	vlažný
	vřelý	kochend	boiling	keev	bouillant	bollente	vrelý
	chladný	kühl	cool	jahe	frais	fresco	chladný
	ledový	gefroren	freezing	jäine	gelé	ghiacciato	ľadový
	mrazivý	eiskalt	gelid	jääkülm	glacial	gelido	mrazivý
	žhavý	glühend	scorching	kõrvetav	brûlant	rovente	žeravý
	kapalný	flüssig	liquid	vedel	liquide	liquido	kvapalný
	plynný	gasförmig	gaseous	gaasiline	gazeux	gassoso	plynný

	suchý	trocken	dry	kuiv	sec	asciutto	suchý
	mokrý	nass	wet	märg	mouillé	bagnato	mokrý
	bledý	blass	pale	kahvatu	pâle	pallido	bledý
	neprůhledný	undurchsichtig	opaque	läbipaistmatu	opaque	opaco	nepriehľadný
	zničený	zerstört	destroyed	hävitatud	détruit	distrutto	zničený
	západní	westlich	western	läänepoolne	occidental	occidentale	západný
	nouns	nouns	nouns	nouns	nouns	nouns	nouns
87	Textile_Fibres	Textile_Fibres	Textile_Fibres	Textile_Fibres	Textile_Fibres	Textile_Fibres	Textile_Fibres
	vlna	Wolle	wool	vill	laine	lana	vlna
	hedvábí	Seide	silk	siid	soie	seta	hodváb
	bavlna	Baumwolle	cotton	puuvill	coton	cotone	bavlna
	nylon	Nylon	nylon	nailon	nylon	canapa	nylon
	viskóza	Viskose	viscose	viskoos	viscose	viscosa	viskóza
	len	Leinen	linen	lina	lin	lino	l'an
	kašmír	Kaschmir	cashmere	kašmiir	cachemire	cashmere	kašmír
	fleece	Fleece	fleece	fliis	molleton	pile	satén
	vlákno	Faden	thread	niit	fil	filo	vlákno
	víakno výšivka	Stickerei	embroidery	tikkimine	broderie	ricamo	víakno výšivka
	kalhoty	Hose	trousers	püksid	pantalon	pantaloni	nohavice
	svetřík	Pullover		kampsun	pull	felpa	
	šicí_stroj	Nähmaschine	sweater sewing_machine	õmblusmasin	machine_à_coudre	macchina_da_cucire	sveter šijací stroj
	šití	Nähen	needlework	õmblemine	couture	cucito	šitie
	zub	Zahn	tooth	hammas	dent	dente	zub
	přátelství	Freundschaft	friendship	sõprus	amitié	amicizia	priateľstvo
	praceistvi	Treundsenart	mendship	soprus		anneizia	priaterstvo
	adjectives	adjectives	adjectives	adjectives	adjectives	adjectives	adjectives
88	Touch_Features	Touch_Features	Touch_Features	Touch_Features	Touch_Features	Touch_Features	Touch_Features
	lepkavý	klebrig	sticky	kleepuv	collant	appiccicoso	lepkavý
	hebký	weich	soft	pehme	mou	soffice	hebký
	drsný	grob	coarse	kare	rugueux	ruvido	hrubý
	pichlavý	scharf	sharp	terav	pointu	appuntito	pichľavý
	tuhý	steif	stiff	jäik	rigide	rigido	tuhý

	slizovitý	schleimig	slimy	limane	gluant	viscido	slizký
	hladký	glatt	smooth	sile	lisse	liscio	hladký
	zvlněný	wellig	wavy	laineline	ondulé	ondulato	zvlnený
	masivní	massiv	massive	massiivne	massif	massiccio	masívny
	choulostivý	delikat	mild	kerge	délicat	delicato	chúlostivý
	silný	stark	strong	tugev	fort	forte	silný
	slabý	schwach	weak	nõrk	faible	debole	slabý
	tmavý	düster	somber	sünge	sombre	cupo	tmavý
	průzračný	klar	limpid	läbipaistev	limpide	limpido	priezračný
	jižní	südlich	southern	lõunapoolne	méridional	meridionale	južný
	daleký	fern	distant	kauge	lointain	distante	ďaleký
	noung	noung		noung	noung	noung	nouns
89	nouns Trees	nouns Trees	nouns Trees	nouns Trees	nouns Trees	nouns Trees	nouns Trees
39	dub	Eiche	oak		chêne		dub
				tamm		quercia	
	bříza	Birke	birch	kask	bouleau	betulla	breza
	buk	Buche	beech	pöök 	hêtre	faggio	buk
	lípa	Linde	linden	pärn	tilleul	tiglio	lipa
	javor	Ahorn	maple	vaher	érable	acero	javor
	smrk	Fichte	spruce	kuusk	épicéa	abete_rosso	smrek
	borovice	Kiefer	pine	mänd	pin	pino	borovica
	jedle	Tanne	fir	nulg	sapin	abete	jedľa
	kůra	Rinde	bark	koor	écorce	corteccia	kôra
	kmen	Stamm	trunk	tüvi	tronc	tronco	kmeň
	plevel	Unkraut	weed	umbrohi	mauvaise_herbe	erbaccia	burina
	květina	Blüte	flower	lill	fleur	fiore	kvet
	rodokmen	Stammbaum	family_tree	sugupuu	arbre_généalogique	albero_genealogico	rodokmeň
	fotosyntéza	Photosynthese	photosynthesis	fotosüntees	photosynthèse	fotosintesi	fotosyntéza
	tělo	Körper	body	keha	corps	corpo	telo
	krematorium	Krematorium	crematorium	krematoorium	crématorium	crematorio	krematórium
	nouns	nouns	nouns	nouns	nouns	nouns	nouns

90	Units_of_Time	Units_of_Time	Units_of_Time	Units_of_Time	Units_of_Time	Units_of_Time	Units_of_Time
	den	Tag	day	päev	jour	giorno	deň
	měsíc	Monat	month	kuu	mois	mese	mesiac
	rok	Jahr	year	aasta	an	anno	rok
	týden	Woche	week	nädal	semaine	settimana	týždeň
	vteřina	Sekunde	second	sekund	seconde	biennio	sekunda
	minuta	Minute	minute	minut	minute	minuto	minúta
	tisíciletí	Jahrtausend	millennium	millennium	millénaire	millennio	tisícročie
	století	Jahrhundert	century	sajand	siècle	secolo	storočie
	litr	Liter	litre	liiter	litre	litro	liter
	metr	Meter	metre	meeter	mètre	metro	meter
	ráno	Vormittag	morning	hommik	matin	mattina	ráno
	odpoledne	Nachmittag	afternoon	pärastlõuna	aprèsmidi	pomeriggio	popoludnie
	rytmus	Rhythmus	rhythm	rütm	rythme	ritmo	rytmus
	rychlost	Geschwindigkeit	speed	kiirus	vitesse	velocità	rýchlosť
	vdova	Witwe	widow	lesk	veuve	vedova	vdova
	lano	Seil	rope	köis	corde	corda	lano
	nouns	nouns	nouns	nouns	nouns	nouns	nouns
91	Vegetables	Vegetables	Vegetables	Vegetables	Vegetables	Vegetables	Vegetables
	dýně	Kürbis	pumpkin	kõrvits	citrouille	zucca	tekvica
	mrkev	Karotte	carrot	porgand	carotte	carota	mrkva
	celer	Staudensellerie	celery	seller	céleri	sedano	zeler
	zelí	Kohl	cabbage	kapsas	chou	cavolo	kapusta
	salát	Salat	salad	salat	salade	insalata	šalát
	řepa	Rübe	beetroot	peet	betterave	barbabietola	repa
	květák	Blumenkohl	cauliflower	lillkapsas	choufleur	cavolfiore	karfiol
	cibule	Zwiebel	onion	sibul	oignon	cipolla	cibul'a
	jablko	Apfel	apple	õun	pomme	mela	jablko
	mango	Mango	mango	mango	mangue	mango	mango
	mandle	Mandel	almond	mandel	amande	mandorla	mandl'a
	čokoláda	Schokolade	chocolate	šokolaad	chocolat	cioccolato	čokoláda

	vejce	Ei	egg	muna	œuf	uovo	vajce
	mléko	Milch	milk	piim	lait	latte	mlieko
	plast	Plastik	plastic	plast	bouteille	plastica	plast
	věž	Turm	tower	torn	tour	torre	veža
	verbs						
92	Verbs_Animal_						
	Sounds						
	štěkat	bellen	bark	haukuma	aboyer	abbaiare	štekať
	mňoukat	miauen	meow	näuguma	miauler	miagolare	mňaukať
	mečet	blöken	bleat	mökitama	bêler	belare	mečať
	bučet	muhen	moo	ammuma	mugir	muggire	bučať
	kuňkat	quaken	croak	krooksuma	coasser	gracidare	kvákať
	pípat	piepen	cheep	piiksuma	piauler	pigolare	pípať
	ržát	wiehern	neigh	hirnuma	hennir	nitrire	erdžať
	vrnět	schnurren	purr	nurruma	ronronner	fare_le_fusa	priasť
	pošeptat	kichern	chuckle	sosistama	ricaner	ridacchiare	šepkať
	škytat	schluchzen	hiccup	nuuksuma	sangloter	singhiozzare	štikútať
	trkat	zermalmen	gore	puksima	encorner	incornare	bodať
	dupat	stampfen	stomp	trampima	piétiner	calpestare	dupať
	bušit	schlagen	bang	taguma	cogner	colpire	búšiť
	kopnout	treten	kick	lööma	frapper	calciare	kopnúť
	vydělat	verdienen	earn	teenima	gagner	guadagnare	zarobiť
	zmrazit	einfrieren	freeze	külmutama	congeler	congelare	zmraziť
	verbs						
93	Verbs Cognition	Verbs Cognition	Verbs_Cognition	Verbs Cognition	Verbs_Cognition	Verbs_Cognition	Verbs_Cognition
	vědět	kennen	know	teadma	savoir	sapere	vedieť
	věřit	glauben	believe	uskuma	croire	credere	veriť
	myslet	denken	think	mõtlema	penser	pensare	myslieť
	rozumět	verstehen	understand	mõistma	comprendre	capire	porozumieť
	pamatovat	erinnern	remember	mäletama	rappeler	ricordare	pamätať
	zapomenout	vergessen	forget	unustama	oublier	dimenticare	zabudnúť

	přemýšlet	meditieren	meditate	mõtlema	réfléchir	riflettere	premýšľať
	interpretovat	interpretieren	interpret	tõlgendama	interpréter	interpretare	interpretovať
	milovat	lieben	love	armastama	aimer	amare	milovať
	nenávidět	hassen	hate	vihkama	haïr	odiare	nenávidieť
	poslouchat	zuhören	listen	kuulama	écouter	ascoltare	počúvať
	slyšet	hören	hear	kuulma	entendre	udire	počuť
	hádat_se	streiten	quarrel	tülitsema	disputer	litigare	hádať_sa
	potvrdit	bestätigen	confirm	kinnitama	confirmer	confermare	potvrdiť
	umýt	waschen	wash	pesema	laver	lavare	umývať
	odstranit	entfernen	remove	eemaldama	éliminer	rimuovere	odstrániť
	1	1	1	1	1		
0.4	verbs	verbs	verbs	verbs	verbs	verbs	verbs
94	Verbs_ Communication 1	Verbs_	Verbs_ Communication 1				
	říct	sagen	say	ütlema	dire	Communication_1 dire	povedať
	tvrdit	behaupten	claim	nõudma	affirmer	affermare	tvrdiť
	ptát_se	fragen	ask	küsima	demander	chiedere	pýtať sa
	odpovědět	antworten	reply	vastama	répondre	rispondere	odpovedať
	volat	zurufen	call	helistama	héler	chiamare	volať
	oznámit	benachrichtigen	announce	teatama	annoncer	annunciare	oznámiť
	opakovat	wiederholen	repeat	kordama	répéter	ripetere	opakovať
	zmínit	erwähnen	mention	mainima	mentionner	menzionare	zmieniť
			mention	mamma		menzionare	
	poslouchat	zuhören	listen	kuulama	écouter	ascoltare	počúvať
	slyšet	hören	hear	kuulma	entendre	udire	počuť
	kašlat	husten	cough	köhima	tousser	tossire	kašľať
	kýchnout	niesen	sneeze	aevastama	éternuer	starnutire	kýchnuť
	spát	schlafen	sleep	magama	dormir	dormire	spať
	jíst	essen	eat	sööma	manger	mangiare	jesť
	zmenšit	verringern	reduce	vähendama	baisser	abbassare	znížiť
	sabotovat	sabotieren	sabotage	saboteerima	manipuler	manomettere	sabotovať
	verbs	verbs	verbs	verbs	verbs	verbs	verbs

95	Verbs_						
	Communication_2						
	vyhlásit	erklären	state	konstateerima	déclarer	dichiarare	vyhlásiť
	svěřit_se	anvertrauen	confide	tunnistama	confier	confidare	zveriť_sa
	vyprávět	erzählen	tell	jutustama	raconter	raccontare	rozprávať
	prozradit	verraten	reveal	paljastama	révéler	rivelare	prezradiť
	napovědět	vorschlagen	suggest	soovitama	suggérer	suggerire	navrhnúť
	doznat_se	bekennen	confess	pihtima	avouer	confessare	priznať_sa
	uznat	zugeben	admit	möönma	admettre	ammettere	uznať
	podotknout	kommentieren	remark	märkima	commenter	commentare	poznamenať
	urazit	beleidigen	insult	solvama	insulter	insultare	uraziť
	proklít	verfluchen	curse	needma	maudire	maledire	prekliať
	ranit	ärgern	offend	solvuma	offenser	offendere	raniť
	umlčet	schweigen	hush	vaikima	faire_taire	zittire	umlčať
	analyzovat	prüfen	examine	uurima	analyser	analizzare	analyzovať
	zahrnout	einbeziehen	include	sisaldama	inclure	includere	zahrnúť
	vyfouknout	ablassen	deflate	tühjaks_minema	dégonfler	sgonfiare	fúkať
	vytesat	ausschnitzen	carve	nikerdama	sculpter	scolpire	vytesať
	verbs						
96	Verbs_Cooking_1						
	smažit	braten	fry	praadima	frire	friggere	smažiť
	vařit	kochen	cook	keetma	cuire	cuocere	variť
	nakrájet	aufschneiden	slice	viilutama	couper	affettare	krájať
	pražit	dünsten	roast	röstima	rôtir	arrostire	opražiť
	upéct	backen	bake	küpsetama	enfourner	infornare	piecť
	strouhat	reiben	grate	riivima	râper	grattugiare	strúhať
	hníst	kneten	knead	sõtkuma	pétrir	impastare	vyprážať
	osmahnout	anrösten	pan_fry	pruunistama	sauter	saltare	prevarit
	prostřít	Tisch_decken	set_the_table	lauda_katma	mettre_la_table	apparecchiare	prestrieť
	servírovat	servieren	serve	teenindama	servir	servire	slúžiť
	krmit	füttern	feed	söötma	nourrir	nutrire	kŕmiť

	hladovět	hungern	starve	nälgima	affamer	affamare	hladovať
	explodovat	explodieren	explode	plahvatama	exploser	esplodere	explodovať
	hnít	faulen	rot	mädanema	pourrir	marcire	hniť
	kopírovat	kopieren	сору	kopeerima	copier	copiare	kopírovať
	rozvíjet	entwickeln	develop	arendama	développer	sviluppare	rozvíjať
	verbs						
97	Verbs_Cooking_2						
	rozdrobit	zerquetschen	mince	hakkima	émietter	sminuzzare	rozdrobiť
	marinovat	marinieren	marinate	marineerima	mariner	marinare	marinovať
	grilovat	grillen	grill	grillima	griller	grigliare	grilovať
	míchat	rühren	stir	segama	mélanger	mescolare	miešať
	loupat	schälen	peel	koorima	éplucher	pelare	lúpať
	drtit	hacken	chop	tükeldama	hacher	tritare	drviť
	toustovat	rösten	toast	röstima	rissoler	tostare	opekať
	podusit	köcheln	simmer	hautama	étuver	stufare	dusiť
	sežrat	auffressen	devour	Impistore	dévorer	divorare	zožrať
				kugistama			
	polknout	schlucken	swallow	neelama	avaler	deglutire	zhltnúť
	hnusit_se	ekeln	disgust	jälestama	dégoûter	disgustare	hnusiť_sa
	porazit	schlachten	butcher	veristama	abattre	macellare	zarezať
	vyrvat	zerreißen	tear	rebima	déchirer	strappare	vyrvať
	bodnout	einstechen	stab	torkama	poignarder	pugnalare	bodnúť
	emigrovat	migrieren	migrate	migreerima	émigrer	emigrare	emigrovať
	protestovat	protestieren	protest	protesteerima	protester	protestare	protestovať
	verbs						
98	Verbs_Crime						
	napálit	reinlegen	cheat	petma	feinter	truffare	podviesť
	oklamat	täuschen	deceive	tüssama	tromper	ingannare	oklamať
	falšovat	fälschen	fake	võltsima	falsifier	falsificare	falšovať
	krást	stehlen	steal	varastama	voler	rubare	ukradnúť
	unést	entführen	kidnap	röövima	kidnapper	rapire	uniesť
	hrozit	drohen	threaten	ähvardama	menacer	minacciare	hroziť

	lhát	lügen	lie	valetama	mentir	mentire	klamať
	potrestat	bestrafen	punish	karistama	punir	punire	potrestať
	odsoudit	verurteilen	convict	süüdi_mõistma	condamner	condannare	odsúdiť
	vynadat	schimpfen	scold	noomima	engueuler	sgridare	vynadať
	diskutovat	diskutieren	argue	vaidlema	discuter	discutere	diskutovať
	odpustit	verzeihen	forgive	andestama	pardonner	perdonare	odpustiť
	osvobodit	befreien	free	vabastama	libérer	liberare	oslobodiť
	koulet	rollen	roll	rullima	enrouler	rotolare	kotúľať
	otáčet_se	rotieren	rotate	pöörlema	tourner	ruotare	otáčať_sa
	verbs						
99	Verbs_Destroying						
	zbořit	demolieren	demolish	lammutama	démolir	demolire	zbúrať
	zničit	zerstören	destroy	hävitama	détruire	distruggere	zničiť
	rozbít	brechen	break	murdma	briser	rompere	rozbiť
	pokazit	vernichten	wreck	purustama	gâcher	rovinare	pokaziť
	zdevastovat	verwüsten	devastate	laastama	dévaster	devastare	zdevastovať
	roztřískat	beschädigen	ruin	vigastama	abîmer	sfasciare	rozbiť
	vyhubit	ausrotten	exterminate	maha_tapma	exterminer	sterminare	vyhubiť
	srazit	abreißen	tear_down	maha_kiskuma	abattre	abbattere	zraziť
	zrušit	abschaffen	abolish	kaotama	abolir	abolire	zrušiť
	vybuchnout	explodieren	burst	paugatama	éclater	scoppiare	explodovať
	opravit	reparieren	repair	parandama	réparer	riparare	opraviť
	rekonstruovat	rekonstruieren	reconstruct	rekonstrueerima	reconstruire	ricostruire	rekonštruovať
	hrozit	bedrohen	threaten	ähvardama	menacer	minacciare	hroziť
	terorizovat	terrorisieren	terrorise	terroriseerima	terroriser	terrorizzare	terorizovať
	uvolnit	lockern	loosen	lõdvendama	desserrer	allentare	uvoľniť
	zacházet	behandeln	treat	ravima	traiter	trattare	zaobchádzať
	verbs						
100	Verbs_Dog						

	štěkat	bellen	bark	haukuma	aboyer	abbaiare	štekať
	potrhat	zerfleischen	maul	lõrisema	mutiler	azzannare	potrhať
	čmuchat	schnuppern	sniff	nuusutama	flairer	fiutare	ňuchať
	vrtět	wedeln	wag	liputama	remuer	scodinzolare	vrtieť
	vrčet	knurren	growl	urisema	grogner	ringhiare	vrčať
	výt	heulen	howl	ulguma	hurler	ululare	vyť
	kňučet	winseln	yelp	klähvima	japper	guaire	kňučať
	pokousat	beißen	bite	hammustama	mordre	mordere	pohrýzť
	zařvat	brüllen	roar	möirgama	rugir	ruggire	zarevať
	pištět	quieken	squeak	kiljuma	couiner	squittire	pišťať
	umřít	sterben	die	surema	mourir	morire	umrieť
	zranit	verwunden	wound	haavama	blesser	ferire	zraniť
	mluvit	sprechen	speak	rääkima	parler	parlare	hovoriť
	křičet	schreien	shout	karjuma	crier	gridare	kričať
	držet	halten	keep	hoidma	tenir	tenere	držať
	ozývat_se	widerhallen	echo	kajama	résonner	rimbombare	ozývať_sa
	verbs	verbs	verbs	verbs	verbs	verbs	verbs
101	Verbs_Driving	Verbs_Driving	Verbs_Driving	Verbs_Driving	Verbs_Driving	Verbs_Driving	Verbs_Driving
101	zrychlit	beschleunigen	hasten	kiirendama	accélérer	accelerare	zrýchliť
	brzdit	bremsen	brake	pidurdama	freiner	frenare	brzdiť
	parkovat	parken	park	parkima	garer	parcheggiare	parkovať
	řídit	fahren	drive	sõitma	conduire	guidare	šoférovať
	zahnout	abbiegen	turn	rooli keerama	tourner	svoltare	zahnúť
	zpomalit	abbremsen	slow_down	aeglustama	ralentir	rallentare	spomaliť
	strhnout	reißen	steer	juhtima	entraîner	sterzare	stiahnuť
		ausschwenken	swerve	libisema	déraper	sbandare	uháňať
	dostat_smyk	aussenwenken	Swerve	noiseina	1		
						navigara	vynlávať
	plout	segeln	sail	purjetama	naviguer	navigare	vyplávať
	plout vykolejit	segeln entgleisen	sail derail	purjetama maabuma	naviguer dérailler	deragliare	vykoľajiť
	plout	segeln	sail	purjetama	naviguer		

	vyhrát	gewinnen	win	võitma	gagner	vincere	vyhrať
	pozdravit	grüßen	greet	tervitama	saluer	salutare	pozdraviť
	hloubit	graben	dig	kaevama	creuser	scavare	kopať
	verbs	verbs	verbs	verbs	verbs	verbs	verbs
102	Verbs_Eating	Verbs_Eating	Verbs_Eating	Verbs_Eating	Verbs_Eating	Verbs_Eating	Verbs_Eating
	jíst	essen	eat	sööma	manger	mangiare	jesť
	žvýkat	kauen	chew	närima	mâcher	masticare	žuť
	polknout	schlucken	swallow	neelama	avaler	deglutire	prehltnúť
	požít	schlingen	ingest	mäluma	ingérer	ingoiare	požiť
	pokousat	beißen	bite	hammustama	mordre	mordere	pohrýzť
	kousnout_si	knabbern	nibble	näksima	mordiller	addentare	zhltnúť
	strávit	verdauen	digest	seedima	digérer	digerire	stráviť
	cpát_se	schlemmen	gorge	õgima	se_gaver	abbuffarsi	napchávať
	zvracet	erbrechen	vomit	oksendama	vomir	vomitare	vracať
	postit_se	fasten	fast	paastuma	jeûner	digiunare	postiť_sa
	obědvat	zu_Mittag_essen	have_lunch	lõunatama	déjeuner	pranzare	obedovať
	večeřet	zu_Abend_essen	dine	einestama	dîner	cenare	večerať
	přibírat	zunehmen	gain_weight	kaalus_juurde_võtma	grossir	ingrassare	pribrať
	znechutit	erkranken	sicken	tülgastama	écœurer	nauseare	znechutiť
	ukázat	zeigen	show	näitama	montrer	mostrare	ukázať
	bušit	schlagen	beat	lööma	frapper	battere	búšiť
	verbs	verbs	verbs	verbs	verbs	verbs	verbs
103	Verbs_Economics	Verbs_Economics	Verbs_Economics	Verbs_Economics	Verbs_Economics	Verbs_Economics	Verbs_Economics
	prodávat	verkaufen	sell	müüma	vendre	vendere	predať
	koupit	kaufen	buy	ostma	acheter	comprare	kúpiť
	splatit	bezahlen	pay	maksma	payer	pagare	platiť
	naúčtovat	berechnen	charge	tasuma	débiter	addebitare	naúčtovať
	zadlužit_se	verschuldet_sein	get_into_debt	võlgu_jääma	s'_endetter	indebitare	zadĺžiť_sa
	zdanit	besteuern	tax	maksustama	taxer	tassare	zdaniť
	vložit	investieren	invest	investeerima	investir	investire	vložiť
	financovat	finanzieren	finance	finantseerima	financer	finanziare	financovať

	vyloupit	berauben	rob	röövima	rapiner	rapinare	vykradnúť
	podvést	schummeln	cozen	tüssama	duper	imbrogliare	podvádzať
	zprostředkovat	vermitteln	mediate	vahendama	arbitrer	mediare	sprostredkovať
	podepsat	unterschreiben	sign	allkirjastama	signer	firmare	podpísať
	měřit	messen	measure	mõõtma	mesurer	misurare	merať
	vážit	wiegen	weigh	kaaluma	peser	pesare	vážiť
	čekat	warten	wait	ootama	attendre	aspettare	čakať
	koktat	stottern	stutter	kogelema	bégayer	balbettare	koktať
	verbs	verbs	verbs	verbs	verbs	verbs	verbs
104	Verbs_Farming	Verbs_Farming	Verbs_Farming	Verbs_Farming	Verbs_Farming	Verbs_Farming	Verbs_Farming
	obdělávat	anpflanzen	cultivate	harima	cultiver	coltivare	obrábať
	orat	pflügen	plough	kündma	labourer	arare	orať
	sít	säen	sow	külvama	semer	seminare	siať
	zavlažovat	bewässern	water	kastma	irriguer	irrigare	zavlažovať
	sázet	pflanzen	plant	istutama	planter	piantare	sadiť
	hnojit	düngen	fertilise	väetama	fertiliser	concimare	hnojiť
	okopat	hacken	hoe	kõplama	biner	zappare	okopávať
	sklidit	ernten	harvest	koristama	récolter	raccogliere	pozberať
	chovat	züchten	breed	aretama	élever	allevare	chovať
	dojit	melken	milk	lüpsma	traire	mungere	dojiť
	umlít	mahlen	grind	ihuma	moudre	macinare	mlieť
	posypat	panieren	flour	iahvatama	fariner	infarinare	posypať
	mést	fegen	sweep	pühkima	balayer	spazzare	zametať
	čistit	reinigen	clean	puhastama	nettoyer	pulire	čistiť
	zatřepat	schütteln	shake	raputama	secouer	agitare	pretrepať
	hostit	bewirten	host	majutama	régaler	ospitare	hostiť
	verbs	verbs	verbs	verbs	verbs	verbs	verbs
105	Verbs_Hair	Verbs_Hair	Verbs_Hair	Verbs_Hair	Verbs_Hair	Verbs_Hair	Verbs_Hair
	rozcuchat	zerzausen	ruffle	sasima	décoiffer	spettinare	rozhádzať
	ostříhat	trimmen	trim	lõikama	couper	tagliare	ostrihať

	splést	flechten	braid	punuma	tresser	intrecciare	upliesť
	barvit	färben	dye	värvima	teindre	tingere	farbiť
	uhladit	glätten	straighten	sirgendama	lisser	lisciare	uhladiť
	zastřihnout	rasieren	shave	raseerima	se_raser	rasare	oholiť
	rozčesat	kämmen	comb	kammima	coiffer	pettinare	česať
	nakadeřit	wellen	curl	lokitama	friser	arricciare	zvlniť
	vyžehlit	bügeln	iron	triikima	repasser	stirare	vyžehliť
	natřít	bestreichen	rub_on	määrima	enduire	spalmare	natrieť
	malovat	malen	paint	maalima	peindre	dipingere	maľovať
	pokosit	beschneiden	prune	pügama	élaguer	potare	pokosiť
	začervenat se	erröten	blush	punastama	rougir	arrossire	červenať sa
	blednout	erbleichen	pale	kahvatuma	pâlir	impallidire	blednúť
	informovat	informieren	inform	teavitama	informer	informare	informovať
	létat	fliegen	fly	lendama	voler	volare	lietať
	verbs	verbs	verbs	verbs	verbs	verbs	verbs
106	Verbs_Human_	Verbs_Human_	Verbs_Human_	Verbs_Human_	Verbs_Human_	Verbs_Human_	Verbs_Human_
	Sounds	Sounds	Sounds	Sounds	Sounds	Sounds	Sounds
	pískat	pfeifen	whistle	vilistama	siffloter	fischiettare	pískať
	mumlat	murmeln	grumble	pomisema	marmonner	borbottare	mumlať
	škytat	schluchzen	hiccup	1.1	1 .	• 1•	×1.1 / 1 / 1
		beindenzen	meeup	luksuma	hoqueter	singhiozzare	štikútať
	šeptat	flüstern	whisper	sosistama	susurrer	singhiozzare sussurrare	šepkať
	šeptat zpívat		_		-	e	
		flüstern	whisper	sosistama	susurrer	sussurrare	šepkať
	zpívat	flüstern singen	whisper sing	sosistama laulma	susurrer chanter	sussurrare cantare	šepkať spievať
	zpívat krknout	flüstern singen aufstoßen	whisper sing burp	sosistama laulma röhitsema	susurrer chanter roter	sussurrare cantare ruttare	šepkať spievať grgať
	zpívat krknout zívnout	flüstern singen aufstoßen gähnen	whisper sing burp yawn chuckle	sosistama laulma röhitsema haigutama	susurrer chanter roter bâiller	sussurrare cantare ruttare sbadigliare ridacchiare	šepkať spievať grgať zívať
	zpívat krknout zívnout chichotat_se hýkat	flüstern singen aufstoßen gähnen kichern	whisper sing burp yawn	sosistama laulma röhitsema haigutama itsitama kisama	susurrer chanter roter bâiller ricaner	sussurrare cantare ruttare sbadigliare	šepkať         spievať         grgať         zívať         chichotať_sa
	zpívat krknout zívnout chichotat_se	flüstern singen aufstoßen gähnen kichern blöken	whisper sing burp yawn chuckle bray	sosistama laulma röhitsema haigutama itsitama	susurrer chanter roter bâiller ricaner braire	sussurrare cantare ruttare sbadigliare ridacchiare ragliare sibilare	šepkať         spievať         grgať         zívať         chichotať_sa         híkať
	zpívat krknout zívnout chichotat_se hýkat zasyčet	flüstern singen aufstoßen gähnen kichern blöken zischen	whisper         sing         burp         yawn         chuckle         bray         hiss	sosistama laulma röhitsema haigutama itsitama kisama sisisema ärkama	susurrer chanter roter bâiller ricaner braire coasser	sussurrare cantare ruttare sbadigliare ridacchiare ragliare	šepkať         spievať         grgať         zívať         chichotať_sa         híkať         syčať
	zpívat krknout zívnout chichotat_se hýkat zasyčet probudit	flüstern singen aufstoßen gähnen kichern blöken zischen wecken	whispersingburpyawnchucklebrayhisswake	sosistama laulma röhitsema haigutama itsitama kisama sisisema	susurrer chanter roter bâiller ricaner braire coasser réveiller	sussurrare cantare ruttare sbadigliare ridacchiare ragliare sibilare svegliare	šepkať         spievať         grgať         zívať         chichotať_sa         híkať         syčať         zobudiť_sa

	sklidit	abräumen	clear_the_table	puhastama	débarrasser	sparecchiare	spratať
	pracovat	arbeiten	work	töötama	travailler	lavorare	pracovať
	verbs	verbs	verbs	verbs	verbs	verbs	verbs
107	Verbs_Killing	Verbs_Killing	Verbs_Killing	Verbs_Killing	Verbs_Killing	Verbs_Killing	Verbs_Killing
	usmrtit	töten	kill	surmama	trucider	uccidere	usmrtiť
	vraždit	morden	assassinate	mõrvama	assassiner	assassinare	vraždiť
	zmasakrovat	massakrieren	batter	surnuks_peksma	massacrer	massacrare	pozabíjať
	zabít	umbringen	murder	tapma	tuer	ammazzare	zabiť
	dusit	ersticken	suffocate	surnuks_lämmatama	étouffer	soffocare	udusiť
	utopit	ertränken	drown	uputama	noyer	affogare	utopiť
	škrtit	erwürgen	strangle	surnuks_kägistama	étrangler	strangolare	škrtiť
	otrávit	vergiften	poison	mürgitama	empoisonner	avvelenare	otráviť
	spáchat_sebevraždu	Selbstmord_begehen	commit_suicide	enesetappu_tegema	se_suicider	suicidarsi	spáchať_samovraždu
	umřít	sterben	die	surema	mourir	morire	zomrieť
	uvěznit	einsperren	imprison	vangistama	emprisonner	incarcerare	väzenie
	obvinit	anklagen	accuse	süüdistama	accuser	accusare	obviňovať
	zprostit	freisprechen	acquit	õigeks_mõistma	acquitter	assolvere	zbaviť
	pohřbít	begraben	bury	matma	enterrer	seppellire	pochovať
	připojit	anschließen	connect	ühendama	connecter	collegare	pripojiť
	navštívit	besuchen	visit	külastama	visiter	visitare	navštíviť
	verbs	verbs	verbs	verbs	verbs	verbs	verbs
108	Verbs_Measures	Verbs_Measures	Verbs_Measures	Verbs_Measures	Verbs_Measures	Verbs_Measures	Verbs_Measures
	prodloužit	verlängern	lengthen	pikendama	allonger	allungare	predĺžiť
	zkrátit	kürzen	shorten	lühendama	raccourcir	accorciare	skrátiť
	zúžit	verengen	tighten	kitsendama	restreindre	restringere	zúžiť
	snížit	verkleinern	reduce	vähendama	réduire	ridurre	znížiť
	rozložit	ausbreiten	extend	laiendama	agrandir	estendere	rozložiť
	zmenšit	schrumpfen	shrink	kokku_tõmbuma	rétrécir	rimpicciolire	zmenšiť
	zvětšit	vergrößern	enlarge	suurendama	élargir	ingrandire	zväčšiť
	zvýšit	erhöhen	increase	kõrgendama	hausser	accrescere	zvýšiť

	měřit	messen	measure	mõõtma	mesurer	misurare	merať
	vážit	wiegen	weigh	kaaluma	peser	pesare	vážiť
	uvolnit	entspannen	relax	lõdvestuma	délasser	rilassare	uvoľniť
	změkčit	erweichen	soften	pehmendama	mollir	ammorbidire	zmäkčiť
	zatížit	verstopfen	encumber	koormama	encombrer	ingombrare	zaťažiť
	obsadit	besetzen	occupy	hõivama	occuper	occupare	obsadiť
	pomýlit	irreführen	mislead	eksitama	enjôler	illudere	pomýliť
	opovážit_se	wagen	dare	julgema	oser	osare	opovážiť_sa
00	verbs	verbs	verbs	verbs	verbs	verbs	verbs
.09	Verbs_Motion	Verbs_Motion	Verbs_Motion	Verbs_Motion	Verbs_Motion	Verbs_Motion	Verbs_Motion
	jít	gehen	go	minema	aller	andare	ísť
	vrátit_se	zurückkehren	return	tagasi_minema	retourner	tornare	vrátiť_sa
	vstoupit	eintreten	enter	sisenema	entrer	entrare	vstúpiť
	vyjít	ausgehen	exit	väljuma	sortir	uscire	vyjsť
	dorazit	ankommen	arrive	jõudma	arriver	arrivare	prísť
	odejít	weggehen	leave	lahkuma	partir	partire	odísť
	sestoupit	heruntergehen	go_down	alla_minema	descendre	scendere	zostúpiť
	vystoupat	hinaufgehen	go_up	üles_tulema	monter	salire	vystúpiť
	hodit	werfen	throw	viskama	jeter	lanciare	hodiť
	vzít	nehmen	catch	püüdma	prendre	prendere	vziať
	spát	schlafen	sleep	magama	dormir	dormire	spať
	poslouchat	zuhören	listen	kuulama	écouter	ascoltare	počúvať
	milovat	lieben	love	armastama	aimer	amare	milovať
	znát	kennen	know	teadma	connaître	conoscere	vedieť
	oddělit	trennen	separate	eraldama	séparer	separare	oddeliť
	osolit	salzen	salt	soolama	saler	salare	osoliť
	verbs			verbs		verbs	verbs
110		verbs	verbs		verbs		
10	Verbs_Mouth	Verbs_Mouth	Verbs_Mouth	Verbs_Mouth	Verbs_Mouth	Verbs_Mouth	Verbs_Mouth
	políbit	küssen	kiss	suudlema	embrasser	baciare	pobozkať
	lízat	lecken	lick	limpsima	lécher	leccare	olízať
	pokousat	beißen	bite	hammustama	mordre	mordere	hrýzť

	pofoukat	pusten	blow	puhuma	souffler	soffiare	pofúkať
	cucat	aussaugen	suck	imema	sucer	succhiare	sať
	usmát_se	lächeln	smile	naeratama	sourire	sorridere	usmievať_sa
	zívat	gähnen	yawn	haigutama	bâiller	sbadigliare	zívať
	plivnout	spucken	spit	sülitama	cracher	sputare	pľuvať
	čichat	schnüffeln	sniff	nuusutama	flairer	annusare	pričuchnúť
	vydechnout	ausatmen	exhale	välja_hingama	exhaler	espirare	vydýchnuť
	mluvit	sprechen	speak	rääkima	parler	parlare	hovoriť
	vyslovit	aussprechen	pronounce	hääldama	prononcer	pronunciare	vyslovovať
	pozorovat	beobachten	observe	jälgima	observer	osservare	pozorovať
	tleskat	klatschen	clap	plaksutama	applaudir	applaudire	tlieskať
	cestovat	reisen	travel	reisima	voyager	viaggiare	cestovať
	odblokovat	entriegeln	unlock	avama	débloquer	sbloccare	odblokovať
	verbs	verbs	verbs	verbs	verbs	verbs	verbs
11	Verbs_Music	Verbs_Music	Verbs_Music	Verbs_Music	Verbs_Music	Verbs_Music	Verbs_Music
	hrát	spielen	play	musitseerima	jouer	suonare	hrať
	ladit	intonieren	tune	häälestama	entonner	intonare	naladiť
	komponovat	komponieren	compose	komponeerima	composer	comporre	komponovať
	zpívat	singen	sing	laulma	chanter	cantare	spievať
	nahrát	aufnehmen	record	lindistama	enregistrer	registrare	nahrať
	improvizovat	improvisieren	improvise	improviseerima	improviser	improvvisare	improvizovať
	poslouchat	zuhören	listen	kuulama	écouter	ascoltare	počúvať
	aranžovat	arrangieren	arrange	orkestreerima	arranger	arrangiare	usporiadať
	nakreslit	zeichnen	draw	joonistama	dessiner	disegnare	nakresliť
	účinkovat	rezitieren	act	deklameerima	réciter	recitare	účinkovať
	hulákat	schreien	scream	karjuma	brailler	urlare	jačať
	pofoukat	pusten	blow	puhuma	souffler	soffiare	pofúkať
	ocenit	schätzen	appreciate	väärtustama	apprécier	apprezzare	oceniť
	následovat	folgen	follow	järgnema	suivre	seguire	sledovať
	vyhnout_se	vermeiden	avoid	vältima	éviter	evitare	vyhnúť sa

	verbs						
112	Verbs_Perception						
	vidět	sehen	see	nägema	voir	vedere	vidieť
	spatřit	erblicken	glimpse	pilku_heitma	entrevoir	scorgere	uvidieť
	slyšet	hören	hear	kuulma	entendre	udire	počuť
	cítit	spüren	feel	tundma	sentir	sentire	cítiť
	poslouchat	zuhören	listen	kuulama	écouter	ascoltare	počúvať
	dívat_se	anschauen	watch	vaatama	regarder	guardare	pozerať
	vnímat	wahrnehmen	sense	tajuma	percevoir	percepire	vnímať
	všimnout_si	bemerken	notice	märkama	remarquer	notare	všimnúť_si
	oslepit	blenden	blind	pimestama	aveugler	accecare	oslepiť
	ztlumit	schweigen	silence	vaigistama	faire_taire	silenziare	stlmiť
	rozumět	verstehen	understand	mõistma	comprendre	capire	rozumieť
	ignorovat	ignorieren	ignore	ignoreerima	ignorer	ignorare	ignorovať
	schválit	billigen	approve	heaks_kiitma	approuver	approvare	schváliť
	zapřít	leugnen	deny	salgama	nier	negare	odoprieť
	spadnout	fallen	fall	langema	tomber	cadere	spadnúť
	garantovat	garantieren	guarantee	garanteerima	garantir	garantire	garantovať
	verbs						
13	Verbs_Plants						
	kvést	blühen	bloom	õitsema	fleurir	fiorire	kvitnúť
	růst	wachsen	grow	kasvama	pousser	crescere	pestovať
	klíčit	keimen	germinate	idanema	germer	germinare	klíčiť
	pučet	sprießen	sprout	võrsuma	faner	germogliare	pučať
	množit se	sich_vermehren	breed	aretama	se_reproduire	riprodursi	množiť sa
	kořenit	sich_entfalten	spring	tärkama	bourgeonner	nascere	zakoreniť
	rašit	blühen	blossom	puhkema	s'_épanouir	sbocciare	pestovať
	vadnout	vertrocknen	photosynthesize	fotosünteesima	flétrir	appassire	rozkvitnúť
	spát	schlafen	sleep	magama	dormir	dormire	spať
	smát_se	lachen	laugh	naerma	rire	ridere	smiať sa

	muset	müssen	have_to	pidama	devoir	dovere	musieť
	vědět	können	know	teadma	savoir	sapere	vedieť
	chtít	wollen	want	tahtma	vouloir	volere	chcieť
	rozumět	verstehen	understand	aru_saama	comprendre	capire	rozumieť
	oddálit	entfernen	move_away	lükkama	éloigner	allontanare	oddialiť
	najít	finden	find	leidma	trouver	trovare	nájsť
	verbs						
114	Verbs_Psych						
	zarmoutit	betrüben	sadden	kurvastama	attrister	rattristare	zarmútiť
	dělat_starosti	bekümmern	worry	muret_tekitama	préoccuper	preoccupare	robiť_starosti
	trápit	beunruhigen	anguish	vaevama	tracasser	angosciare	trápiť
	nudit	langweilen	bore	tüütama	ennuyer	annoiare	nudiť
	deprimovat	deprimieren	depress	masendama	déprimer	deprimere	deprimovať
	postrašit	erschrecken	frighten	hirmutama	effrayer	spaventare	postrašiť
	rozveselit	aufheitern	cheer_up	tuju_tõstma	égayer	rallegrare	rozveseliť
	povzbudit	ermutigen	encourage	julgustama	encourager	incoraggiare	povzbudiť
	preferovat	bevorzugen	prefer	eelistama	préférer	preferire	uprednostňovať
	toužit	wünschen	desire	ihaldama	désirer	desiderare	túžiť
	hloubat	meditieren	meditate	mediteerima	méditer	meditare	hĺbať
	kritizovat	kritisieren	criticize	kritiseerima	critiquer	criticare	kritizovať
	zlepšit	verbessern	improve	parandama	améliorer	migliorare	zlepšiť
	zhoršit	verschlimmern	worsen	halvendama	empirer	peggiorare	zhoršiť
	fackovat	ohrfeigen	slap	lööma	gifler	schiaffeggiare	fackať
	dopravit	transportieren	transport	transportima	transporter	trasportare	prepraviť
	verbs						
115	Verbs_Religion						
	požehnat	segnen	bless	õnnistama	bénir	benedire	požehnať
	vysvětit	weihen	consecrate	pühitsema	consacrer	consacrare	vysvätiť
	pokřtít	taufen	baptize	ristima	baptiser	battezzare	pokrstiť
	posvětit	heiligen	sanctify	kanoniseerima	sanctifier	santificare	posvätiť
	konvertovat	konvertieren	convert	usku_vahetama	convertir	convertire	konvertovať

	krást	stehlen	steal	varastama	voler	rubare	ukradnúť
	rozvést_se	sich_scheiden_lassen	divorce	lahutuma	divorcer	divorziare	rozviesť_sa
	usmrtit	töten	kill			uccidere	zabiť
	volit	wählen		tapma hääletama	tuer		voliť
		umschreiben	vote transcribe		voter	votare	
	přepsat			üles_kirjutama	transcrire	trascrivere	prepísať
	oslnit	blenden	dazzle	pimestama	éblouir	abbagliare	oslniť
	dokončit	beenden	complete	täiendama	terminer	terminare	dokončiť
	verbs	verbs	verbs	verbs	verbs	verbs	verbs
16	Verbs_School	Verbs_School	Verbs_School	Verbs_School	Verbs_School	Verbs_School	Verbs_School
	studovat	studieren	study	uurima	étudier	studiare	študovať
	naučit_se	lernen	learn	õppima	apprendre	imparare	naučiť_sa
	napsat	schreiben	write	kirjutama	écrire	scrivere	napísať
	číst	lesen	read	lugema	lire	leggere	čítať
	počítat	zählen	count	loendama	compter	contare	počítať
	zapamatovat_si	auswendig_lernen	memorize	meelde_jätma	mémoriser	memorizzare	zapamätať_si
	osvojit_si	erwerben	acquire	omandama	acquérir	apprendere	osvojiť_si
	zopakovat	wiederholen	repeat	kordama	répéter	ripetere	opakovať
	vyučovat	beibringen	teach	õpetama	enseigner	insegnare	učiť
	vyvolat	aufrufen	question	üle_kuulama		interrogare	vyvolať
	poslouchat		obey	kuuletuma	interroger obéir	obbedire	počúvať
	dodržovat	gehorchen	•				dodržovať
		respektieren	respect	austama	respecter	rispettare	
	pokročit	vorangehen	progress	edenema	progresser	progredire	pokročiť
	chodit	besuchen	attend	käima	fréquenter	frequentare	zúčastniť_sa
	vysušit	austrocknen	drain	kuivatama	assécher	prosciugare	vysušiť
	zachovat	bewahren	maintain	säilitama	maintenir	mantenere	zachovať

117	Verbs_Smell	Verbs_Smell	Verbs_Smell	Verbs_Smell	Verbs_Smell	Verbs_Smell	Verbs_Smell
	čichat	schnüffeln	sniff	nuusutama	flairer	annusare	pričuchnúť
	páchnout	riechen	smell	haistma	sentir	odorare	páchnuť
	vonět	parfümieren	perfume	lehkama	parfumer	profumare	voňať
	zapáchat	stinken	stink	haisema	puer	puzzare	smrdieť
	vydat	abgeben	effuse	piserdama	émettre	effondere	vydať
	vydechnout	ausatmen	exhale	välja_hingama	exhaler	espirare	vydýchnuť
	vydávat	ausströmen	emanate	lõhnama	émaner	emanare	vydávať
	vdechovat	einatmen	inhale	sisse_hingama	inhaler	inspirare	vdychovať
	okusit	kosten	taste	maitsma	goûter	assaggiare	okúsiť
	dotknout_se	berühren	touch	puudutama	toucher	toccare	dotknúť_sa
	poslouchat	zuhören	listen	kuulama	écouter	ascoltare	počúvať
	vidět	sehen	see	nägema	voir	vedere	vidieť
	krvácet	bluten	bleed	veritsema	saigner	sanguinare	krvácať
	omdlít	ohnmächtig_werden	faint	minestama	s'_évanouir	svenire	omdlieť
	nosit	tragen	wear	kandma	habiller	vestire	nosiť
	zachránit	speichern	save	päästma	sauver	salvare	zachrániť
	verbs	verbs	verbs	verbs	verbs	verbs	verbs
18	Verbs_Sport	Verbs_Sport	Verbs_Sport	Verbs_Sport	Verbs_Sport	Verbs_Sport	Verbs_Sport
	bruslit	eislaufen	skate	uisutama	patiner	pattinare	korčuľovať
	běhat	laufen	run	jooksma	courir	correre	behať
	lyžovat	skifahren	ski	suusatama	skier	sciare	lyžovať
	skákat	springen	jump	hüppama	sauter	saltare	skákať
	plavat	schwimmen	swim	ujuma	nager	nuotare	plávať
	potápět_se	tauchen	dive	sukelduma	plonger	tuffare	potápať_sa
	lézt	klettern	climb	ronima	gravir	arrampicare	liezť
	tancovat	tanzen	dance	tantsima	danser	ballare	tancovať
	ležet	liegen	lie	lamama	coucher	giacere	ležať
	opalovat_se	sonnen	sunbathe	päevitama	prendre_le_soleil	prendere_il_sole	opaľovať_sa
	vařit	kochen	cook	keetma	cuisinier	cucinare	variť
	uklízet	aufräumen	tidy_up	koristama	nettoyer	riordinare	upratovať

	potit_se	schwitzen	sweat	higistama	suer	sudare	potiť_sa
	zhubnout	abnehmen	lose_weight	kaalust_alla_võtma	maigrir	dimagrire	schudnúť
	představovat	darstellen	represent	kujutama	représenter	rappresentare	predstavovať
	otevřít	öffnen	open	avama	ouvrir	aprire	otvoriť
	verbs	verbs	verbs	verbs	verbs	verbs	verbs
119	Verbs_Telephone	Verbs_Telephone	Verbs_Telephone	Verbs_Telephone	Verbs_Telephone	Verbs_Telephone	Verbs_Telephone
	zavolat	anrufen	call	helistama	appeler	chiamare	zavolať
	odpovědět	antworten	answer	vastama	répondre	rispondere	odpovedať
	zavěsit	auflegen	hang_up	ära_panema	raccrocher	riattaccare	zavesiť
	textovat	simsen	text	sõnumit_saatma	envoyer_un_messag	messaggiare	poslať_správu
	řinčet	anklingeln	ring	helisema	sonner	squillare	rinčať
	konverzovat	plaudern	chat	vestlema	chatter	conversare	konverzovať
	telefonovat	chatten	phone	vastu_võtma	téléphoner	telefonare	telefonovať
	zavolat_zpátky	zurückrufen	call_back	tagasi_helistama	rappeler	richiamare	zložiť
	zpívat	singen	sing	laulma	chanter	cantare	spievať
	vřískat	schreien	yell	karjuma	crier	strillare	kričať
	programovat	programmieren	program	programmeerima	programmer	programmare	programovať
	dělat_rozhovor	interviewen	interview	intervjueerima	interviewer	intervistare	robiť_rozhovor
	filmovat	filmen	film	filmima	filmer	filmare	filmovať
	odebírat	abonnieren	subscribe	tellima	abonner	abbonare	odoberať
	přestat	aufhören	stop	peatuma	cesser	smettere	zastaviť
	zatlačit	drücken	push	lükkama	pousser	spingere	zatlačiť
	verbs	verbs	verbs	verbs	verbs	verbs	verbs
120	Verbs_Touch	Verbs_Touch	Verbs_Touch	Verbs_Touch	Verbs_Touch	Verbs_Touch	Verbs_Touch
	hladit	streicheln	caress	paitama	caresser	accarezzare	pohladkať
	poškrábat	kratzen	scratch	kriimustama	griffer	graffiare	poškriabať
	ohmatat	abtasten	feel	kompima	tâter	tastare	cítiť
	dotknout_se	berühren	touch	puudutama	toucher	toccare	dotknúť_sa
	zavadit	streifen	stroke	riivama	frôler	sfiorare	zavadiť
	sevřít	fassen	grip	haarama	serrer	stringere	zovrieť

22	War	War	War	War	War	War	War
	nouns	nouns	nouns	nouns	nouns	nouns	nouns
	zvednout	heben	lift	tõstma	soulever	sollevare	zdvihnúť
	položit	legen	lay	lamama	étendre	distendere	položiť
	uklidnit	beruhigen	calm_down	rahunema	calmer	calmare	upokojiť_sa
	výt	heulen	howl	ulguma	hurler	ululare	vyť
	zničit	zerstören	destroy	hävitama	détruire	distruggere	zničiť
	plakat	weinen	cry	nutma	pleurer	piangere	plakať
	vřít	kochen	boil	keema	bouillir	bollire	vrieť
	odpařit	verdampfen	evaporate	aurustuma	s'_évaporer	evaporare	vypariť
	mrznout	gefrieren	freeze	külmetama	geler	ghiacciare	mrznúť
	mrholit	nieseln	drizzle	tibutama	bruiner	piovigginare	mrholiť
	zatáhnout_se	bewölken	cloud_over	pilve_minema	ennuager	annuvolare	zatiahnuť_sa
	foukat	donnern	thunder	müristama	tonner	tuonare	fúkať
	lít	gießen	pour	kallama	flotter	diluviare	liať
	padat_kroupy	hageln	hail	rahet_sadama	grêler	grandinare	viať
	sněžit	schneien	snow	lund_sadama	neiger	nevicare	snežiť
	pršet	regnen	rain	sadama	pleuvoir	piovere	pršať
L	Verbs_Weather	Verbs_Weather	Verbs_Weather	Verbs_Weather	Verbs_Weather	Verbs_Weather	Verbs_Weather
	verbs	verbs	verbs	verbs	verbs	verbs	verbs
	IIIZet	verschwinden	disappear	каципа	disparaître	scomparire	Zimiznut
	mizet	verschwinden	seem	kaduma	1		zmiznúť
	opálit_se připadat	bräunen scheinen	tan	päevituma tunduma	bronzer paraître	abbronzare sembrare	opáliť_sa pripadať
		jucken	itch	sügelema	démanger	prudere	
	dýchat svědit	atmen	breathe	hingama	respirer	respirare	dýchať svrbieť
	nasytit	sättigen	satiate	rahuldama	rassasier	saziare	nasýtiť
	čichat	schnüffeln	sniff	nuusutama	flairer	annusare	pričuchnúť
	pofoukat	pusten	blow	puhuma	souffler	soffiare	fúkať
			11.	1			C' 1- 42
	popadnout	ergreifen	grasp	kahmama	agripper	afferrare	uchopiť
	stisknout	drücken	press	vajutama	appuyer	premere	stlačiť

	voják	Soldat	soldier	sõdur	soldat	soldato	vojak
	zákop	Schützengraben	trench	kaevik	tranchée	trincea	zákop
	invaze	Invasion	invasion	invasioon	invasion	invasione	invázia
	bitva	Schlacht	battle	lahing	bataille	battaglia	bitka
	kulka	Kugel	bullet	kuul	balle	pallottola	guľka
	mírová_smlouva	Friedensvertrag	peace_treaty	rahuleping	traité_de_paix	trattato_di_pace	mierová_zmluva
	mina	Landmine	land_mine	maamiin	mine	mina_antiuomo	mína
	veterán	Veteran	veteran	veteran	vétéran	veterano	veterán
	sportovec	Sportler	athlete	sportlane	athlète	atleta	športovec
	povodeň	Überschwemmung	flood	üleujutus	inondation	inondazione	povodeň
	hádka	Streit	quarrel	tüli	querelle	litigio	hádka
	hůl	Stock	stick	kepp	bâton	bastone	palica
	soutěž	Wettkampf	competition	võistlus	tournoi	gara	súťaž
	rozhodčí	Schiedsrichter	referee	kohtunik	arbitre	arbitro	rozhodca
	filozofie	Philosophie	philosophy	filosoofia	philosophie	filosofia	filozofia
	peněženka	Brieftasche	wallet	rahakott	portefeuille	portafoglio	peňaženka
	nouns	nouns	nouns	nouns	nouns	nouns	nouns
123	Water_Means_of_	Water_Means_of_	Water_Means_of_	Water_Means_of_	Water_Means_of_	Water_Means_of_	Water_Means_of_
	Transport	Transport	Transport	Transport	Transport	Transport	Transport
	kánoe	Kanu	canoe	kanuu	canoë	canoa	kanoe
	člun	Boot	boat	paat	barque	barca	čln
	loď	Schiff	ship	laev	navire	nave	loď
	parník	Dampfschiff	steamship	aurulaev	bateau_à_vapeur	piroscafo	parník
	trajekt	Fähre	ferry	parvlaev	ferry	traghetto	trajekt
	jachta	Jacht	yacht	jaht	yacht	gondola	jachta
	výletní_loď	Kreuzfahrtschiff	cruise_ship	kruiisilaev	bateau_de_croisière	nave_da_crociera	výletná_loď
	motorový_člun	Motorboot	motorboat	mootorpaat	vedette	motoscafo	motorový_čln
	auto	Auto	car	auto	voiture	automobile	auto
	horkovzdušný balón	Heißluftballon	hot_air_balloon	kuumaõhupall	montgolfière	mongolfiera	teplovzdušný balón
	tuleň	Seehund	seal	hüljes	phoque	foca	tuleň
	bóje	Boje	buoy	poi	bouée	boa	plavák

	přístav	Hafen	harbour	sadam	port	porto	prístav
	plavba	Schifffahrt	sailing	purjetamine	navigation	navigazione	plavba
	medaile	Medaille	medal	medal	médaille	medaglia	medaila
	paruka	Perücke	wig	parukas	perruque	parrucca	parochňa
	nouns	nouns	nouns	nouns	nouns	nouns	nouns
124	Weapons	Weapons	Weapons	Weapons	Weapons	Weapons	Weapons
	bomba	Bombe	bomb	pomm	bombe	bomba	bomba
	meč	Schwert	sword	mõõk	épée	spada	meč
	atomová_bomba	Atombombe	nuclear_weapon	tuumarelv	arme_nucléaire	arma_nucleare	atómová_bomba
	kulomet	Maschinengewehr	machine_gun	kuulipilduja	mitrailleuse	mitragliatrice	gul'omet
	granát	Granate	grenade	granaat	grenade	granata	granát
	dělo	Kanone	cannon	suurtükk	canon	cannone	kanón
	dýka	Dolch	dagger	pistoda	poignard	pugnale	dýka
	puška	Gewehr	rifle	vintpüss	fusil	fucile	puška
	váleček	Nudelholz	rolling_pin	taignarull	rouleau_à_pâtisserie	mattarello	valec
	kladivo	Baseballschläger	baseball_bat	pesapallimüts	batte_de_baseball	mazza_da_baseball	kladivo
	výbuch	Explosion	explosion	plahvatus	explosion	esplosione	výbuch
	masakr	Gemetzel	massacre	massimõrv	massacre	strage	masaker
	válka	Krieg	war	sõda	guerre	guerra	vojna
	palba	Schuss	gunshot	lask	coup_de_feu	sparo	paľba
	kalendář	Kalender	calendar	kalender	calendrier	calendario	kalendár
	kříž	Kreuz	cross	rist	croix	croce	kríž
	adjectives	adjectives	adjectives	adjectives	adjectives	adjectives	adjectives
125	Weather_Conditions	Weather_Conditions	Weather_Conditions	Weather_Conditions	Weather_Conditions	Weather_Conditions	Weather_Conditions
	slunný	sonnig	sunny	päikeseline	ensoleillé	soleggiato	slnečný
	oblačný	bewölkt	cloudy	pilvine	nuageux	nuvoloso	zamračený
	deštivý	regnerisch	rainy	vihmane	pluvieux	piovoso	daždivý
	větrný	windig	windy	tuuline	venteux	ventilato	veterný
	mlhavý	neblig	foggy	udune	brumeux	nebbioso	hmlistý
	bouřkový	stürmisch	stormy	tormine	orageux	temporalesco	búrlivý
	sněžný	verschneit	snowy	lumine	neigeux	nevoso	snehový

	dusný	schwül	muggy	umbne	étouffant	afoso	sparný
		1 1 1	1.1		1		
	temný	dunkel	dark	tume	sombre	scuro	tmavý
	jasný	hell	bright	ere	clair	chiaro	jasný
	lesklý	funkelnd	shiny	läikiv	brillant	brillante	lesklý
	černý	schwarz	black	must	noir	nero	čierny
	šťastný	glücklich	happy	õnnelik	heureux	felice	šťastný
	rozzlobený	wütend	angry	vihane	fâché	arrabbiato	nahnevaný
	nerovný	ungleich	unequal	ebavõrdne	inégal	disuguale	nerovný
	východní	östlich	eastern	idapoolne	oriental	orientale	východný
	nouns	nouns	nouns	nouns	nouns	nouns	
126	Weather_Events	Weather_Events	Weather_Events	Weather_Events	Weather_Events	Weather_Events	Weather_Events
	liják	Gewitter	storm	torm	orage	temporale	lejak
	bouřka	Regenguss	thunderstorm	äikesetorm	averse	acquazzone	búrka
	mrak	Wolke	cloud	pilv	nuage	nuvola	mrak
	déšť	Regen	rainfall	vihmasadu	pluie	pioggia	dážď
	sníh	Schnee	snow	lumi	neige	neve	sneh
	mlha	Nebel	fog	udu	brouillard	nebbia	hmla
	krupobití	Hagel	hail	rahe	grêle	grandine	krupobitie
	vítr	Wind	wind	tuul	vent	vento	vietor
	led	Eis	ice	jää	glace	ghiaccio	ľad
	pára	Dampf	steam	aur	vapeur	vapore	para
	deštník	Regenschirm	umbrella	vihmavari	parapluie	ombrello	dáždnik
	příval	Strom	stream	oja	torrent	torrente	lejak
	změna klimatu	Klimawandel	climate_change	kliimamuutus	changement_climatique	cambiamento_climatico	klimatické_zmeny
	předpověď počasí	Wettervorhersage	weather_forecast	ilmaprognoos	prévision_météo	previsioni_meteo	predpoveď počasia
	rozhodnutí	Entscheidung	decision	otsus	décision	decisione	rozhodnutie
	zápalka	Streichholz	matchstick	tuletikk	allumette	fiammifero	zápalka
	nouns	nouns	nouns	nouns	nouns	nouns	nouns
127	Wild_Animals	Wild_Animals	Wild_Animals	Wild_Animals	Wild Animals	Wild_Animals	Wild_Animals
	divočák	Wildschwein	wild_boar	metssiga	sanglier	cinghiale	diviak

	ježek	Igel	hedgehog	siil	hérisson	riccio	ježko
	vlk	Wolf	wolf	hunt	loup	lupo	vlk
	jelen	Hirsch	deer	hirv	cerf	cervo	jeleň
	bobr	Murmeltier	marmot	kobras	marmotte	marmotta	bobor
	sokol	Falke	hawk	kotkas	faucon	falco	jastrab
	sova	Eule	howl	uluk	hurlement	gufo	sova
	tchoř	Iltis	polecat	tuhkur	putois	puzzola	tchor
	panda	Panda	panda	panda	panda	panda	panda
	tuleň	Seehund	seal	hüljes	phoque	foca	tuleň
	kožešina	Pelz	fur	karvkate	fourrure	pelliccia	kožušina
	roh	Hörner	horns	sarved	corne	corna	roh
	lov	Jagd	hunting	jahtimine	chasse	caccia	lov
	vyhynutí	Aussterben	extinction	väljasuremine	extinction	estinzione	vyhynutie
	poklad	Schatz	treasure	aare	trésor	tesoro	poklad
	žárovka	Glühbirne	light_bulb	pirn	ampoule	lampadina	žiarovka
	nouns	nouns	nouns	nouns	nouns	nouns	nouns
128	Zodiac_Signs	Zodiac_Signs	Zodiac_Signs	Zodiac_Signs	Zodiac_Signs	Zodiac_Signs	Zodiac_Signs
	štír	Skorpion	scorpio	Skorpion	scorpion	scorpione	škorpión
	lev	Löwe	leo	Lõvi	lion	leone	lev
	beran	Widder	aries	Jäär	bélier	ariete	baran
	blíženec	Zwillinge	gemini	Kaksikud	gémeaux	gemelli	blíženec
	býk	Stier	taurus	Sõnn	taureau	toro	býk
	vodnář	Wassermann	aquarius	Veevalaja	verseau	acquario	vodnár
	váhy	Waage	libra	Kaalud	balance	bilancia	váhy
		I		Neitsi	vierge	vergine	panna
	panna	Jungfrau	virgo	INCIUSI	(leige		Puilla
	panna	Jungirau	Virgo	INCIUSI			- Pullin
	puma	Puma	puma	puuma	puma	puma	puma
							1
	puma	Puma	puma	puuma	puma	puma	puma
	puma vlaštovka	Puma Schwalbe Ratte Tiger	puma swallow	puuma pääsuke	puma avaler	puma rondine	puma lastovička
	puma vlaštovka krysa	Puma Schwalbe Ratte	puma swallow rat	puuma pääsuke rott	puma avaler rat	puma rondine ratto	puma lastovička potkan

vlas	Haar	hair	juus	cheveu	capello	vlas
ručník	Handtuch	towel	rätik	serviette	asciugamano	uterák

# Appendix 2. Guidelines for HAMOD Dataset Creation and Translation

# Summary

1. Dataset Creation	
1.1. Selection of the Semantic Categories	
1.2. Inliers Selection	
1.3. Outliers Selection	
1.4. Formal Requirements for the encoding of the dataset	285
1.5. Format	285
2. Translation and Adaptation of the Dataset	

# **1.** Dataset Creation

This part of the document (part 1) has been adapted from Camacho-Collados & Navigli (2016) guidelines.<sup>1</sup>

HAMOD dataset (High Agreement Multi-lingual Outlier Detection dataset) is a dataset for exercising the outlier detection task on distributional models.

<sup>&</sup>lt;sup>1</sup> See <u>http://lcl.uniroma1.it/outlier-detection/</u> (last access: 24/06/2022).

It consists of several sets. Each set (equivalent to *cluster* in Camacho-Collados & Navigli, 2016) is made by a group of 8 words belonging to a *semantic category* or pertaining to a specific *topic*<sup>2</sup> and 8 words which do not fit to the first 8 at different degrees (i.e., they do not have relevant or enough properties to be considered parts of the semantic category or the topic). We call the words belonging to the semantic category/topic *inliers* and the words which do not fit *outliers*.

Examples of semantic categories (included in our dataset) are: "School Subjects", "Means of transport", "Clothes", "Parts of Skeleton", "Trees". Examples of topics are: "Music", "Informatics", "Linguistics", "Cooking".

#### 1.1. Selection of the Semantic Categories

Here are some requirements in the selection of the topics, the semantic categories, and the specific inliers and outliers:

*1. Named Entities and Proper Names.* Named Entities and proper names have to be avoided. "Solar System Planets", "South American Countries", "Presidents of Czech Republic" are not suitable candidates for a set, "Musical Instruments", "Shapes" and "Professions" are suitable candidates for a set.

2. General Knowledge. Categories and topics should belong to some general knowledge, thus avoid narrow and domain-specific categories or topics. "Farm Animals" (*cow*, *pig*, *goose*, *dog*, etc.) is a suitable category, but "Dog Breeds" (*basset hound*, *bohemian shepherd*, *poodle*, *bulldog*, etc.) may be too specific and may not belong to some shared knowledge.

*3. 12-year-old Vocabulary.* Words chosen as inliers and outliers should be easily understood by a 12-year-old person. If you cannot test this, try to use highly frequent and try not to use domain-specific vocabulary.

4. Semantics, nothing else. As we are evaluating thesauri (which encode semantic relations), the criteria for the identification of the sets must be semantic. "Interrogative Pronouns" or "Time Preposition" are not suitable candidates for a set, as the criterion would be syntactic or morphological.

## 1.2. Inliers Selection

We intend *semantic categories* as sets of words referring to entities, properties or events sharing some common features: for example, "Means of Transport" contains items which share the feature of being human-made artifacts, used by human beings to move around. The features are implicitly reflected in the label assigned to each set by the annotator, which is only needed to identify the set (but will not be used in the task nor signaled

<sup>&</sup>lt;sup>2</sup> See Paragraph 1.2. for a definition.

to the human evaluators). For example, "Means of Transport" is a set based on a semantic category. Here are its 8 inliers:

- 01 motorbike
  02 ship
  03 car
  04 tram
  05 bus
  06 train
- 07 plane
- 08 helicopter

Even if there are several types of means of transport (road, water, and flying vehicle), they all share the property of being human-made artifacts used by human beings to move around. A more specific set, derived from this, is "Road Means of Transport":

01 car
02 bus
03 taxi
04 bike
05 motorbike
06 trolleybus
07 van
08 scooter

In this set, *train*, *plane* and *boat* would be outliers because they are not means of transport used on the road (they are used on rails, air, and water, instead).

Sets based on *topics* are less strict: they can contain words that do not necessarily shared relevant features (e.g., abstract and concrete, human-made and natural objects, events, and properties can be mixed). For example, "Music" is a set based on a topic. Here are its 8 inliers:

01	note
02	song
03	guitar
04	rock
05	flute
06	sound
07	microphone
08	singer

In this case, items may belong to different semantic categories: for example, *guitar*, *flute* are musical instruments, *singer* and is an artistic profession, *rock* and *note* are abstract entities etc. All the inliers clearly pertain to the same topic.

### 1.3. Outliers Selection

Once you have identified the topic or the semantic category and its corresponding inliers, choosing the outliers may be trickier and clear guidelines are needed. Inliers are among themselves similar and related; outliers should have instead a lower degree of similarity and relatedness, at different hierarchical levels.

We follow Camacho-Collados & Navigli (2016) in dividing the 8 outliers in 4 sub-sets (thus, 2 words per each sub-set), defining each sub-set as follows:

- 1. sub-set 1: two words that are closely related to the inliers, thus sharing a high number of features with them, but not enough to be part of the inliers. For example, in "Road Means of Transport" (see above), *skates* and *airplane* are means of transport, but an airplane is a flying vehicle, not a road one, and the skates are more like a sport equipment, therefore they are distinguished from the inliers.
- 2. sub-set 2: two words that are less related by sharing less features. For example, *road* and *roundabout* are always human-made entities but cannot be said to be means of transport, but they are still related to the semantic category of driving.
- 3. sub-set 3: two words that are even less related, but still pertaining to the semantic category. They may be words referring to different kind of entities (from concrete to abstract, in the case of our examples) or to events or properties. For example, *traffic* and *car\_crash* refer to kinds of events that can involve road vehicles.
- 4. sub-set 4: two words that are not related at all to the inliers. They can be random words. For example, *rugby* and *toaster* do not share any feature with the inliers nor pertain to the semantic category.

To sum up, the outliers of the set "Road Means of Transport" are:

01 skates
02 airplane
03 road
04 roundabout
05 traffic
06 car\_crash
07 rugby
08 toaster

For what concerns topics, the selection of the first six outliers does not have to follow the same hierarchy as for semantic categories, as different entities, events, and properties may be included among the inliers. The last two outliers always need to be random words. See for example the outliers for the set "Music" (mentioned above):

> 01 letter 02 colour 284

- 03 drawing
- 04 sculpture
- 05 writer
- 06 painter
- 07 picnic
- 08 pocket

### 1.4. Formal Requirements for the encoding of the dataset

Here are some formal requirements in the selection and encoding of the inliers and the outliers in the sets:

1. Parts of Speech. Sets only contain words with the same part of speech, considering inliers and outliers altogether. There are only-noun, only-verb and only-adjectives sets in the current state of the dataset.

2. *Multiword Expressions*. In order to be processed by the evaluation script, multiword expressions need to be encoded with an underscore joining each word of the term. See, for example, *peanut\_butter*, *salle\_de\_bain* (Eng., 'bathroom'), *cambiamento\_climatico* (Eng., 'climate change').

*3. Lemmas.* Words have to be encoded as their lemma. For example, singular form is the lemma for a noun, singular masculine for an adjective, infinitive form for a verb in Italian. Plural forms are not accepted unless the word is a *pluralia tantum* - i.e., only have a plural form (e.g., *trousers* in English).

#### 1.5. Format

The format of the sets has to be a simple .txt file containing the inliers and the outliers as follows. Inliers can be in a random order, as there is no hierarchical relation among the 8 elements. Outliers need to follow the order outlined in the paragraph above. Inliers and outliers must be separated by an empty line. A different .txt file must be created for each language.

inlier 1 inlier 2 inlier 3 inlier 4 inlier 5 inlier 6 inlier 7 inlier 8 <Empty line> outlier from sub-set 1 outlier from sub-set 2 outlier from sub-set 2 outlier from sub-set 3 outlier from sub-set 3 outlier from sub-set 4 outlier from sub-set 4

See the set "Means of Transport" as an example:

motorbike ship car tram bus train plane helicopter exercise\_bike treadmill pavement road driver pilot needle shoe

## 2. Translation and Adaptation of the Dataset

In this part of the document, we summarize the steps needed in order to translate and adapt the HAMOD dataset to other languages. English must be kept as source language for any translation/adaptation of the sets. Starting from English, the 8 inliers + 8 outliers have to be accurately translated into their equivalent in the new language(s).

One may choose not to translate directly - but to adapt the word into something similar that still fits among the inliers or the outliers - in the following cases:

- 1. there is no exact correspondence in the translation;
- 2. the corresponding translation is infrequent (according to a reference corpus) in the new language(s);
- 3. the corresponding translation is too polysemous and/or ambiguous;
- 4. the word is too culture-specific (e.g., names of food, means of transports, animals) and therefore absent in the new language(s).
In these circumstances, a multiword expression can be used if it is attested in the corpus - i.e., it is not Out-Of-Vocabulary (OOV). For example, *pet* in English would be translated *animale\_domestico* in Italian.

In case this is not possible, the target word can be replaced with a completely different one - but still semantically related according to the guidelines for the selection of the inliers and the outliers. For example, as *custard* in English is culture-specific, it can be adapted to *tvaroh* in Czech (thus always referring to a dairy product).

Other desiderata to be followed:

*1. Usage.* The translated word has to be commonly used in the language (in particular for what concerns multiword expressions) – in case of uncertainty, a corpus in the target language can be checked.

2. *Parts of Speech*. The translated word has to be in the same part of speech of the other words in the set (there cannot be mixed sets with both adjectives, verbs, and nouns: only-verbs, only-nouns and only-adjectives sets have to be created).

*3. Lemmas.* The translated word has to be encoded as its lemma. Plural forms are not accepted unless the word is a *pluralia tantum* - i.e., only have a plural form (e.g., *trousers* in English).

4. Semantic Polysemy/Ambiguity. In case of semantic polysemy/ambiguity, the translated word has to be contextually consistent with the others in the set. For example, *oil* in the set "Cooking" must refer to the ingredient used to cook or fry (thus, *olio* in Italian, *huile* in French etc.). If it is in the set "Sources of Energy", then it refers to *petroleum* used to power engines (thus, *petrolio* in Italian, *pétrole* in French etc.).

5. Part of Speech Ambiguity. In case of part of speech ambiguity, the translated word has to be consistent with the others in the set. For example, *fast* in the set "Verbs Eating" cannot refer to its homonymous adjective meaning "quick, rapid", but instead it would mean "to eat no food" (and thus translated as *paastuma* in Estonian, *fasten* in German etc.).

## Appendix 3. Quantitative Results of the Evaluation of Distributional Models

In this Appendix we report the detailed results of the evaluation of the distributional models for each of the 128 sets in the dataset, divided per language. Each Table is structured as follows: in column 1 there is the set name; in column 2 the accuracy of the Sketch Engine Thesaurus; in column 3 the OPP of the Sketch Engine Thesaurus; in column 4 the accuracy of the Word Embeddings with attribute "word"; in column 5 the OPP of the Word Embeddings with attribute "word"; in column 6 the accuracy of the Word Embeddings with attribute "lemma"; in column 7 the OPP of the Word Embeddings with attribute "lemma".

	CS	CS	CS	CS	CS	CS
set name	SkeThe	SkeThe	WE_	WE_	WE_	WE_
	accuracy	OPP	word	word	lemma	lemma
			accuracy	OPP	accuracy	OPP
Art	0.375	0.722	0.750	0.889	0.000	0.806
Astronomical_Objects	0.625	0.958	0.375	0.806	0.250	0.819
Biomes	0.625	0.958	0.875	0.986	1.000	1.000
Birds	1.000	1.000	0.875	0.986	0.875	0.986
Bodies_of_Water	0.250	0.792	0.875	0.958	0.875	0.931
Book_Genres	0.500	0.917	0.250	0.819	0.000	0.639
Bugs	0.625	0.889	0.625	0.833	0.500	0.819
Building_Materials	1.000	1.000	1.000	1.000	1.000	1.000
Buildings	0.125	0.889	0.625	0.806	0.375	0.750
Car_Components	0.875	0.986	0.875	0.972	0.875	0.986
Chemical_Elements	0.750	0.972	1.000	1.000	0.625	0.917
Clothes	1.000	1.000	1.000	1.000	0.000	0.778
Colours	1.000	1.000	1.000	1.000	1.000	1.000
Computer_Components	0.375	0.778	0.625	0.903	0.625	0.708
Containers	0.750	0.889	0.625	0.875	0.375	0.875
Cooking	0.375	0.875	0.500	0.917	0.875	0.986
Dairy_Products	0.875	0.986	0.750	0.861	0.625	0.792
Dances	0.375	0.806	0.750	0.972	0.875	0.917
Dimensional_Features_1	0.500	0.903	1.000	1.000	1.000	1.000
Dimensional_Features_2	0.875	0.958	0.625	0.903	0.750	0.944
Dishes_and_Cutlery	0.875	0.986	0.750	0.972	0.500	0.944
Economics	0.375	0.806	1.000	1.000	1.000	1.000

Table 1. Sketch Engine Thesaurus and Word Embeddings evaluation; accuracy and OPP computed per set, for Czech

Electronics	0.000	0.889	1.000	1.000	1.000	1.000
External_Body_Parts	0.875	0.986	1.000	1.000	1.000	1.000
Extreme_Natural_Events	0.875	0.986	0.625	0.931	0.750	0.944
Family_Members	1.000	1.000	1.000	1.000	1.000	1.000
Fantasy_Characters	0.375	0.861	0.500	0.847	0.500	0.903
Farm Animals	0.750	0.944	0.750	0.917	0.625	0.958
Firearms	1.000	1.000	1.000	1.000	1.000	1.000
Fish	0.875	0.986	1.000	1.000	0.750	0.847
Flowers	0.000	0.861	0.000	0.847	0.000	0.861
Flying_Means_of_Transport	0.375	0.875	0.875	0.972	0.875	0.972
Food	0.875	0.972	0.500	0.931	0.625	0.958
Food_Features	0.500	0.931	0.625	0.958	0.750	0.972
Free_Time_Activities	0.000	0.611	0.500	0.917	0.000	0.806
Fruit	0.500	0.931	0.500	0.875	0.250	0.764
Fruit Trees	0.625	0.958	1.000	1.000	1.000	1.000
Furniture	0.125	0.861	1.000	1.000	1.000	1.000
Gemstones	1.000	1.000	0.625	0.958	1.000	1.000
Grain	1.000	1.000	1.000	1.000	1.000	1.000
Hair_Features	0.125	0.736	0.375	0.722	0.375	0.736
Herbs	1.000	1.000	0.875	0.986	0.875	0.972
Human_Features_Negativity	0.375	0.792	0.750	0.972	0.000	0.667
Human_Features_Positivity	0.875	0.944	0.625	0.958	0.500	0.833
Human_Moods	0.250	0.833	0.875	0.986	0.750	0.972
Human_Physical_Features	0.250	0.667	0.500	0.917	0.500	0.806
Illnesses	0.875	0.986	0.625	0.958	0.500	0.944
Informatics	0.000	0.542	0.625	0.958	0.875	0.875
Internal_Body_Parts	0.000	0.889	0.875	0.958	0.875	0.972
Kitchenware	0.625	0.944	0.875	0.972	0.375	0.806
Landscape_Features	0.625	0.875	1.000	1.000	1.000	1.000
Languages	1.000	1.000	1.000	1.000	1.000	1.000
Linguistics	0.250	0.681	0.875	0.986	1.000	1.000
Liquid_Containers	0.250	0.861	0.875	0.972	0.625	0.944
Materials	0.625	0.958	0.625	0.917	1.000	1.000
Maths	0.250	0.681	0.625	0.944	0.625	0.958
Means_of_Transport	0.750	0.972	0.750	0.944	0.750	0.931
Medicine	0.125	0.486	1.000	1.000	1.000	1.000
Metals	0.750	0.944	0.500	0.944	0.875	0.986
Music	0.250	0.750	1.000	1.000	1.000	1.000
Music_Genres	0.375	0.861	1.000	1.000	0.750	0.958
Musical_Instruments	1.000	1.000	0.875	0.986	0.000	0.764
Non-alcoholic_Drinks	0.250	0.736	0.500	0.931	0.500	0.944
Nuts	0.500	0.819	0.625	0.806	0.625	0.931
Office_Supplies	0.125	0.681	0.500	0.833	0.625	0.903
Parts_of_Head	0.750	0.958	0.750	0.917	0.375	0.931
Parts_of_House	0.625	0.958	0.250	0.875	0.375	0.931
Parts_of_Skeleton	0.750	0.958	0.375	0.917	0.375	0.847
Parts_of_Speech	0.250	0.750	0.250	0.806	0.375	0.806
Politics	0.250	0.833	1.000	1.000	1.000	1.000
Professions	0.250	0.875	1.000	1.000	0.750	0.958

Reptiles	0.625	0.861	0.625	0.875	0.750	0.889
Road_Means_of_Transport	0.500	0.792	0.375	0.931	0.375	0.889
Rooms_in_the_House	0.250	0.833	0.750	0.917	0.625	0.944
Savanna_Animals	0.750	0.917	0.500	0.792	0.625	0.889
School_Subjects	0.875	0.986	0.750	0.972	0.750	0.958
Shapes	0.750	0.972	0.875	0.972	0.750	0.972
Shoes	0.000	0.736	0.750	0.889	0.000	0.292
Shops	0.875	0.986	1.000	1.000	1.000	1.000
Sources_of_Energy	0.250	0.722	0.625	0.889	0.625	0.875
Spices	0.500	0.931	0.625	0.958	0.500	0.889
Spirits	0.625	0.861	1.000	1.000	1.000	1.000
Sport	1.000	1.000	1.000	1.000	1.000	1.000
Sports	0.875	0.958	0.750	0.958	0.625	0.931
Sweets	0.250	0.847	0.625	0.944	0.375	0.931
Temperature_Features	0.625	0.861	0.625	0.917	0.375	0.736
Textile_Fibres	0.500	0.917	0.375	0.875	0.250	0.833
Touch_Features	0.000	0.639	0.500	0.875	0.230	0.931
Trees	1.000	1.000	1.000	1.000	1.000	1.000
Units_of_Time	0.125	0.903	0.500	0.861	0.750	0.903
Vegetables	0.875	0.972	1.000	1.000	0.750	0.972
Verbs_Animal_Sounds	0.000	0.708	0.500	0.806	0.250	0.736
Verbs_Cognition	0.250	0.833	0.375	0.875	0.125	0.819
Verbs_Communication_1	0.500	0.847	0.625	0.917	0.750	0.917
Verbs_Communication_2	0.625	0.833	0.250	0.917	0.250	0.889
Verbs_Cooking_1	0.500	0.917	0.875	0.972	0.875	0.972
Verbs_Cooking_2	0.000	0.806	0.875	0.972	0.875	0.986
Verbs_Crime	0.125	0.611	0.375	0.819	0.250	0.694
Verbs_Destroy	0.000	0.514	0.500	0.917	0.500	0.847
Verbs_Dog	0.250	0.806	0.625	0.847	0.500	0.889
Verbs_Driving	0.125	0.736	0.500	0.778	0.375	0.750
Verbs_Eating	0.750	0.833	0.125	0.722	0.000	0.653
Verbs_Economics	0.375	0.736	0.625	0.958	0.250	0.875
Verbs_Farming	0.875	0.986	0.250	0.917	1.000	1.000
Verbs_Hair	0.000	0.778	0.250	0.875	0.375	0.917
Verbs_Human_Sounds	0.000	0.417	0.500	0.792	0.250	0.806
Verbs_Killing	0.125	0.639	0.375	0.750	0.375	0.819
Verbs_Measure	0.375	0.903	0.750	0.972	0.625	0.958
Verbs_Motion	0.250	0.764	0.875	0.972	0.750	0.917
Verbs_Mouth	0.250	0.667	0.625	0.847	0.750	0.847
Verbs_Music	0.375	0.722	0.875	0.986	0.875	0.972
Verbs_Perception	0.875	0.958	0.625	0.889	0.750	0.917
Verbs_Plants	0.125	0.750	0.875	0.986	0.875	0.986
Verbs_Psych	0.250	0.500	0.500	0.861	0.625	0.917
Verbs_Religion	0.000	0.889	0.625	0.931	0.625	0.903
Verbs_School	0.250	0.736	0.375	0.889	0.250	0.875
Verbs_Smell	0.125	0.597	0.125	0.819	0.125	0.833
Verbs_Sport	0.375	0.861	1.000	1.000	0.750	0.861
Verbs_Telephone	0.125	0.569	0.000	0.694	0.375	0.736
Verbs_Touch	0.250	0.847	0.375	0.833	0.250	0.889

Verbs_Weather	0.375	0.833	0.625	0.958	1.000	1.000
War	0.000	0.486	0.750	0.972	0.375	0.889
Water_Means_of_Transport	0.500	0.861	0.750	0.972	0.750	0.903
Weapons	0.000	0.889	0.625	0.903	0.625	0.861
Weather_Conditions	0.875	0.958	0.875	0.944	0.750	0.847
Weather_Events	0.625	0.944	1.000	1.000	1.000	1.000
Wild_Animals	0.250	0.889	0.875	0.944	0.875	0.944
Zodiac_Signs	0.500	0.931	0.250	0.875	0.000	0.875

	DE	DE	DE	DE	DE	DE
set name	SkeThe	SkeThe	WE_	WE_	WE_	WE_
	accuracy	OPP	word	word	lemma	lemma
			accuracy	OPP	accuracy	OPP
Art	0.250	0.722	1.000	1.000	0.750	0.972
Astronomical_Objects	0.000	0.889	0.250	0.722	0.250	0.722
Biomes	0.875	0.986	1.000	1.000	1.000	1.000
Birds	0.875	0.986	0.750	0.972	1.000	1.000
Bodies_of_Water	0.625	0.833	0.500	0.903	0.500	0.931
Book_Genres	0.375	0.931	0.375	0.819	0.375	0.764
Bugs	0.500	0.889	0.625	0.903	0.750	0.917
Building_Materials	1.000	1.000	0.875	0.972	0.875	0.972
Buildings	0.750	0.972	0.375	0.806	0.625	0.833
Car_Components	1.000	1.000	1.000	1.000	0.875	0.986
Chemical_Elements	0.750	0.972	0.750	0.944	0.750	0.972
Clothes	0.750	0.972	0.500	0.931	0.500	0.931
Colours	0.750	0.931	1.000	1.000	1.000	1.000
Computer_Components	0.125	0.778	0.750	0.917	0.500	0.806
Containers	0.375	0.833	0.500	0.847	0.750	0.889
Cooking	0.250	0.750	0.750	0.972	0.625	0.958
Dairy_Products	0.500	0.944	0.500	0.833	0.500	0.833
Dances	0.500	0.917	0.875	0.986	0.875	0.972
Dimensional_Features_1	0.625	0.833	0.375	0.917	1.000	1.000
Dimensional_Features_2	0.250	0.833	0.125	0.903	0.375	0.931
Dishes_and_Cutlery	0.500	0.944	0.500	0.875	0.625	0.889
Economics	0.375	0.819	0.750	0.972	1.000	1.000
Electronics	0.375	0.931	1.000	1.000	1.000	1.000
External_Body_Parts	0.750	0.972	1.000	1.000	1.000	1.000
Extreme_Natural_Events	0.625	0.931	0.625	0.958	0.500	0.944
Family_Members	1.000	1.000	1.000	1.000	1.000	1.000
Fantasy_Characters	0.750	0.944	0.500	0.847	0.625	0.889
Farm_Animals	1.000	1.000	1.000	1.000	1.000	1.000
Firearms	0.875	0.986	0.875	0.986	0.625	0.958
Fish	0.875	0.986	0.625	0.958	0.375	0.819
Flowers	0.750	0.972	0.750	0.972	0.750	0.944
Flying_Means_of_Transport	0.625	0.958	0.875	0.972	0.875	0.958
Food	0.000	0.889	0.000	0.806	0.000	0.819
Food_Features	0.625	0.875	0.625	0.875	0.750	0.917
Free_Time_Activities	0.000	0.444	0.750	0.917	0.500	0.889
Fruit	0.625	0.917	0.375	0.889	0.375	0.875
Fruit_Trees	0.750	0.972	0.875	0.986	1.000	1.000
Furniture	0.500	0.903	1.000	1.000	1.000	1.000
Gemstones	1.000	1.000	0.750	0.944	0.750	0.958
Grain	0.875	0.986	1.000	1.000	0.875	0.875
Hair_Features	1.000	1.000	0.625	0.875	0.500	0.875
Herbs	0.875	0.986	0.625	0.944	0.500	0.944
Human_Features_Negativity	0.500	0.875	1.000	1.000	1.000	1.000

Table 2. Sketch Engine Thesaurus and Word Embeddings evaluation; accuracy and OPP computed per set, for German

Human_Features_Positivity	0.250	0.764	0.125	0.833	0.125	0.861
Human_Moods	0.625	0.875	0.625	0.917	1.000	1.000
Human_Physical_Features	0.000	0.514	0.875	0.944	0.750	0.903
Illnesses	1.000	1.000	0.750	0.944	0.000	0.847
Informatics	0.250	0.681	1.000	1.000	1.000	1.000
Internal_Body_Parts	0.750	0.972	0.750	0.944	0.750	0.944
Kitchenware	0.375	0.625	0.375	0.833	0.375	0.833
Landscape_Features	0.500	0.875	0.250	0.917	0.375	0.931
Languages	0.875	0.875	1.000	1.000	0.000	0.000
Linguistics	0.375	0.708	0.875	0.986	1.000	1.000
Liquid_Containers	0.250	0.750	0.750	0.972	0.875	0.986
Materials	0.375	0.847	0.875	0.972	0.625	0.833
Maths	0.250	0.750	0.250	0.833	0.500	0.861
Means_of_Transport	0.500	0.889	0.500	0.903	0.500	0.875
Medicine	0.000	0.528	0.250	0.917	0.375	0.875
Metals	0.750	0.931	1.000	1.000	0.750	0.972
Music	0.125	0.694	0.750	0.972	1.000	1.000
Music_Genres	0.250	0.917	1.000	1.000	1.000	1.000
Musical_Instruments	1.000	1.000	1.000	1.000	1.000	1.000
Non-alcoholic_Drinks	0.375	0.806	0.875	0.958	0.750	0.958
Nuts	0.250	0.847	0.625	0.847	0.500	0.903
Office_Supplies	0.125	0.694	0.375	0.833	0.625	0.875
Parts_of_Head	0.750	0.903	0.250	0.861	1.000	1.000
Parts_of_House	0.125	0.833	0.375	0.931	0.250	0.847
Parts_of_Skeleton	0.500	0.944	0.750	0.958	0.750	0.972
Parts_of_Speech	0.375	0.833	0.500	0.889	0.625	0.806
Politics	0.125	0.778	1.000	1.000	1.000	1.000
Professions	0.250	0.819	0.625	0.958	0.375	0.903
Reptiles	0.625	0.847	0.625	0.778	0.625	0.833
Road_Means_of_Transport	0.375	0.819	0.875	0.986	1.000	1.000
Rooms_in_the_House	1.000	1.000	0.875	0.958	0.875	0.944
Savanna_Animals	0.750	0.917	0.625	0.889	0.875	0.958
School_Subjects	0.250	0.806	0.750	0.917	0.750	0.917
Shapes	0.000	0.847	0.750	0.903	0.750	0.917
Shoes	0.500	0.944	0.500	0.861	0.000	0.389
Shops	0.625	0.917	0.875	0.986	1.000	1.000
Sources_of_Energy	0.625	0.833	0.500	0.861	0.500	0.847
Spices	0.875	0.986	0.500	0.944	0.000	0.736
Spirits	0.625	0.931	1.000	1.000	1.000	1.000
Sport	1.000	1.000	1.000	1.000	0.875	0.972
Sports	0.625	0.833	0.750	0.958	0.750	0.972
Sweets	0.500	0.792	0.625	0.958	0.875	0.972
Temperature_Features	0.500	0.806	0.625	0.903	0.750	0.889
Textile_Fibres	1.000	1.000	0.625	0.931	0.625	0.931
Touch_Features	0.250	0.764	0.750	0.944	1.000	1.000
Trees	1.000	1.000	0.875	0.972	0.875	0.972
Units_of_Time	0.500	0.944	0.375	0.847	0.500	0.889
Vegetables	0.250	0.792	0.375	0.903	0.250	0.903
Verbs_Animal_Sounds	0.750	0.861	0.750	0.972	0.625	0.903

Verbs_Cognition	0.000	0.792	0.250	0.806	0.375	0.819
Verbs_Communication_1	0.375	0.819	0.250	0.778	0.250	0.792
Verbs_Communication_2	0.250	0.750	0.250	0.667	0.500	0.833
Verbs_Cooking_1	0.625	0.931	0.875	0.972	0.875	0.972
Verbs_Cooking_2	0.500	0.944	0.875	0.958	0.875	0.986
Verbs_Crime	0.125	0.528	0.375	0.764	0.500	0.833
Verbs_Destroy	0.125	0.736	0.750	0.875	0.500	0.903
Verbs_Dog	0.125	0.528	0.000	0.625	0.125	0.667
Verbs_Driving	0.000	0.722	0.625	0.875	0.750	0.917
Verbs_Eating	0.000	0.750	0.625	0.833	0.500	0.764
Verbs_Economics	0.375	0.833	0.750	0.972	0.750	0.972
Verbs_Farming	0.500	0.917	0.500	0.944	0.375	0.931
Verbs_Hair	0.375	0.806	0.500	0.917	0.500	0.875
Verbs_Human_Sounds	0.250	0.694	0.500	0.861	0.375	0.819
Verbs_Killing	0.250	0.736	0.625	0.917	0.500	0.931
Verbs_Measure	0.375	0.833	1.000	1.000	0.875	0.986
Verbs_Motion	0.125	0.444	0.875	0.958	0.750	0.931
Verbs_Mouth	0.125	0.764	0.750	0.931	0.625	0.875
Verbs_Music	0.375	0.639	0.625	0.931	0.625	0.917
Verbs_Perception	0.250	0.806	0.375	0.847	0.375	0.819
Verbs_Plants	0.000	0.694	1.000	1.000	1.000	1.000
Verbs_Psych	0.375	0.750	0.625	0.944	0.500	0.944
Verbs_Religion	0.250	0.861	0.500	0.944	0.750	0.972
Verbs_School	0.250	0.861	0.125	0.750	0.250	0.750
Verbs_Smell	0.125	0.556	0.250	0.917	0.250	0.917
Verbs_Sport	0.000	0.722	0.750	0.889	0.875	0.931
Verbs_Telephone	0.000	0.389	0.000	0.889	0.125	0.889
Verbs_Touch	0.375	0.819	0.125	0.694	0.250	0.694
Verbs_Weather	0.000	0.389	0.750	0.917	0.625	0.903
War	0.125	0.472	0.500	0.944	0.500	0.944
Water_Means_of_Transport	0.625	0.903	0.875	0.958	0.625	0.903
Weapons	0.875	0.986	0.875	0.958	0.625	0.917
Weather_Conditions	1.000	1.000	0.875	0.986	1.000	1.000
Weather_Events	0.375	0.875	1.000	1.000	1.000	1.000
Wild_Animals	0.375	0.917	0.625	0.903	0.625	0.833
Zodiac_Signs	0.875	0.986	0.875	0.986	0.000	0.861

	EN	EN	EN	EN	EN	EN
set name	SkeThe	SkeThe	WE_	WE_	WE_	WE_
	accuracy	OPP	word	word	lemma	lemma
			accuracy	OPP	accuracy	OPP
Art	0.125	0.625	0.750	0.889	0.750	0.889
Astronomical_Objects	0.250	0.861	0.375	0.750	0.250	0.722
Biomes	0.125	0.903	0.750	0.931	0.750	0.917
Birds	0.875	0.986	0.750	0.958	1.000	1.000
Bodies_of_Water	0.375	0.806	0.500	0.889	0.500	0.833
Book_Genres	0.375	0.736	0.625	0.847	0.500	0.764
Bugs	0.375	0.833	0.625	0.847	0.625	0.833
Building_Materials	0.875	0.986	0.875	0.917	1.000	1.000
Buildings	0.750	0.972	0.500	0.833	0.125	0.806
Car_Components	1.000	1.000	0.750	0.972	0.500	0.931
Chemical_Elements	0.625	0.958	0.625	0.889	0.500	0.833
Clothes	0.250	0.917	0.500	0.931	0.500	0.889
Colours	1.000	1.000	1.000	1.000	1.000	1.000
Computer_Components	0.250	0.750	0.875	0.958	0.500	0.931
Containers	0.750	0.903	0.625	0.819	0.250	0.764
Cooking	0.125	0.847	0.500	0.917	0.750	0.958
Dairy_Products	0.250	0.875	0.750	0.806	0.625	0.792
Dances	0.125	0.722	0.750	0.931	0.500	0.861
Dimensional_Features_1	0.625	0.944	0.875	0.917	0.875	0.917
Dimensional_Features_2	0.000	0.889	0.750	0.972	0.250	0.889
Dishes_and_Cutlery	0.500	0.889	0.250	0.861	0.375	0.903
Economics	0.500	0.889	0.375	0.931	0.750	0.972
Electronics	0.000	0.792	0.000	0.792	0.250	0.833
External_Body_Parts	0.750	0.958	0.875	0.986	0.875	0.986
Extreme_Natural_Events	0.000	0.833	0.750	0.917	0.625	0.903
Family_Members	1.000	1.000	1.000	1.000	1.000	1.000
Fantasy_Characters	0.625	0.931	0.750	0.931	0.625	0.917
Farm_Animals	0.625	0.875	0.500	0.931	0.875	0.986
Firearms	0.125	0.806	1.000	1.000	0.875	0.986
Fish	1.000	1.000	0.750	0.931	0.625	0.903
Flowers	1.000	1.000	0.375	0.819	0.625	0.931
Flying_Means_of_Transport	0.000	0.528	0.625	0.944	0.625	0.889
Food	0.375	0.931	0.500	0.819	0.500	0.931
Food_Features	0.375	0.903	0.500	0.944	0.500	0.833
Free_Time_Activities	0.000	0.500	0.375	0.875	0.500	0.875
Fruit	0.625	0.903	0.500	0.833	0.250	0.833
Fruit_Trees	1.000	1.000	1.000	1.000	1.000	1.000
Furniture	0.625	0.875	0.500	0.875	0.750	0.917
Gemstones	0.500	0.944	0.750	0.972	0.750	0.972
Grain	0.750	0.944	1.000	1.000	1.000	1.000
Hair_Features	0.125	0.653	0.250	0.625	0.000	0.472
Herbs	0.625	0.944	0.500	0.875	0.375	0.861
Human_Features_Negativity	0.625	0.958	1.000	1.000	1.000	1.000

Table 3. Sketch Engine Thesaurus and Word Embeddings evaluation; accuracy and OPP computed per set, for English

	0.105	0.670	0.075	0.006	0.500	0.075
Human_Features_Positivity	0.125	0.653	0.375	0.806	0.500	0.875
Human_Moods	0.750	0.944	0.875	0.972	0.750	0.958
Human_Physical_Features	0.375	0.833	0.500	0.944	0.625	0.847
Illnesses	0.875	0.986	0.750	0.958	0.625	0.917
Informatics	0.250	0.750	1.000	1.000	1.000	1.000
Internal_Body_Parts	1.000	1.000	0.875	0.917	0.875	0.958
Kitchenware	0.000	0.875	0.250	0.750	0.125	0.764
Landscape_Features	0.625	0.917	0.500	0.944	0.750	0.972
Languages	1.000	1.000	1.000	1.000	0.875	0.875
Linguistics	0.250	0.639	0.375	0.931	0.375	0.931
Liquid_Containers	0.750	0.833	0.750	0.958	0.750	0.944
Materials	0.375	0.917	0.750	0.958	0.750	0.958
Maths	0.250	0.681	0.625	0.917	0.500	0.875
Means_of_Transport	0.125	0.806	0.500	0.917	0.375	0.889
Medicine	0.375	0.722	1.000	1.000	1.000	1.000
Metals	0.500	0.944	0.125	0.889	0.125	0.903
Music	0.125	0.750	0.625	0.958	0.500	0.944
Music_Genres	0.125	0.722	0.250	0.764	0.250	0.764
Musical_Instruments	0.250	0.917	0.875	0.986	0.000	0.875
Non-alcoholic_Drinks	0.125	0.778	0.750	0.972	0.750	0.972
Nuts	0.250	0.847	0.250	0.708	0.375	0.750
Office_Supplies	0.000	0.514	0.125	0.792	0.125	0.806
Parts_of_Head	0.875	0.958	0.750	0.958	0.875	0.972
Parts_of_House	0.500	0.833	0.375	0.861	0.500	0.889
Parts_of_Skeleton	0.375	0.819	0.750	0.944	0.625	0.944
Parts_of_Speech	0.250	0.778	0.000	0.778	0.000	0.778
Politics	0.125	0.681	1.000	1.000	0.500	0.944
Professions	0.000	0.528	0.375	0.722	0.250	0.597
Reptiles	0.500	0.861	0.625	0.847	0.750	0.875
Road_Means_of_Transport	0.125	0.833	0.250	0.847	0.375	0.875
Rooms_in_the_House	0.500	0.917	0.750	0.931	0.875	0.958
Savanna_Animals	1.000	1.000	1.000	1.000	1.000	1.000
School_Subjects	0.375	0.861	0.750	0.944	0.750	0.958
Shapes	0.000	0.861	0.500	0.861	0.500	0.861
Shoes	0.000	0.569	0.500	0.778	0.000	0.194
Shops	0.125	0.681	0.875	0.972	0.875	0.958
Sources_of_Energy	0.125	0.639	0.500	0.778	0.500	0.778
Spices	0.875	0.958	0.750	0.958	0.000	0.778
Spirits	0.625	0.917	1.000	1.000	1.000	1.000
Sport	0.750	0.861	0.875	0.958	0.875	0.958
Sports	0.250	0.806	0.625	0.847	0.875	0.958
Sweets	0.375	0.806	0.625	0.958	0.625	0.958
Temperature_Features	0.000	0.639	0.750	0.861	0.500	0.750
Textile_Fibres	0.375	0.917	0.875	0.986	0.875	0.958
Touch_Features	0.375	0.861	1.000	1.000	1.000	1.000
Trees	1.000	1.000	1.000	1.000	0.875	0.986
Units_of_Time	0.375	0.903	0.750	0.889	0.750	0.889
Vegetables	0.625	0.889	0.875	0.986	0.875	0.986
Verbs_Animal_Sounds	0.250	0.889	0.375	0.861	0.375	0.819

Verbs_Cognition	0.125	0.833	0.500	0.889	0.500	0.889
Verbs_Communication_1	0.625	0.917	0.500	0.792	0.750	0.889
Verbs_Communication_2	0.375	0.889	0.125	0.819	0.375	0.931
Verbs_Cooking_1	0.125	0.889	1.000	1.000	1.000	1.000
Verbs_Cooking_2	1.000	1.000	1.000	1.000	1.000	1.000
Verbs_Crime	0.000	0.639	0.500	0.792	0.500	0.819
Verbs_Destroy	0.125	0.819	0.000	0.764	0.000	0.819
Verbs_Dog	0.250	0.764	0.625	0.764	0.625	0.750
Verbs_Driving	0.125	0.681	0.375	0.847	0.375	0.833
Verbs_Eating	0.125	0.903	0.375	0.875	0.250	0.833
Verbs_Economics	0.625	0.917	0.875	0.986	0.875	0.986
Verbs_Farming	0.500	0.889	0.500	0.917	0.500	0.944
Verbs_Hair	0.750	0.903	0.625	0.931	0.625	0.931
Verbs_Human_Sounds	0.125	0.625	0.500	0.819	0.500	0.750
Verbs_Killing	0.250	0.722	0.250	0.750	0.250	0.778
Verbs_Measure	0.750	0.903	0.875	0.958	0.750	0.944
Verbs_Motion	0.125	0.639	0.750	0.972	0.875	0.986
Verbs_Mouth	0.625	0.944	0.625	0.875	0.625	0.847
Verbs_Music	0.000	0.611	0.125	0.806	0.250	0.903
Verbs_Perception	0.125	0.750	0.625	0.861	0.500	0.903
Verbs_Plants	0.000	0.833	0.875	0.986	0.875	0.986
Verbs_Psych	0.000	0.792	0.125	0.736	0.250	0.861
Verbs_Religion	0.250	0.847	0.375	0.917	0.375	0.931
Verbs_School	0.000	0.694	0.250	0.764	0.125	0.792
Verbs_Smell	0.000	0.639	0.500	0.833	0.625	0.917
Verbs_Sport	0.125	0.806	0.625	0.931	0.750	0.958
Verbs_Telephone	0.250	0.681	0.250	0.833	0.250	0.778
Verbs_Touch	0.625	0.944	0.000	0.653	0.375	0.806
Verbs_Weather	0.250	0.819	0.125	0.764	0.250	0.792
War	0.375	0.764	0.750	0.903	0.750	0.944
Water_Means_of_Transport	0.500	0.833	0.750	0.917	0.750	0.903
Weapons	0.000	0.847	1.000	1.000	1.000	1.000
Weather_Conditions	1.000	1.000	1.000	1.000	1.000	1.000
Weather_Events	0.375	0.778	0.750	0.931	0.750	0.931
Wild_Animals	0.000	0.722	0.500	0.931	0.500	0.778
Zodiac_Signs	1.000	1.000	1.000	1.000	1.000	1.000

	ET	ET	ET	ET	ET	ET
set name	SkeThe	SkeThe	WE_	WE_	WE_	WE_
	accuracy	OPP	word	word	lemma	lemma
			accuracy	OPP	accuracy	OPP
Art	0.000	0.708	0.625	0.917	0.875	0.97
Astronomical_Objects	1.000	1.000	0.250	0.736	0.250	0.72
Biomes	1.000	1.000	1.000	1.000	1.000	1.00
Birds	1.000	1.000	0.625	0.903	0.750	0.97
Bodies_of_Water	0.750	0.861	0.250	0.903	1.000	1.00
Book_Genres	0.000	0.625	0.625	0.847	0.750	0.84
Bugs	0.500	0.917	0.625	0.819	0.750	0.90
Building_Materials	1.000	1.000	0.750	0.944	0.750	0.79
Buildings	1.000	1.000	0.625	0.903	0.125	0.86
Car_Components	0.000	0.889	0.500	0.944	0.750	0.97
Chemical_Elements	1.000	1.000	0.625	0.917	0.500	0.83
Clothes	0.500	0.889	1.000	1.000	0.000	0.68
Colours	0.625	0.875	1.000	1.000	0.875	0.98
Computer_Components	0.875	0.903	0.750	0.972	0.500	0.91
Containers	0.625	0.903	0.750	0.958	0.750	0.91
Cooking	1.000	1.000	0.875	0.986	0.625	0.94
Dairy_Products	0.875	0.986	0.125	0.819	0.250	0.81
Dances	0.875	0.972	0.125	0.889	0.375	0.88
Dimensional_Features_1	0.625	0.875	1.000	1.000	1.000	1.00
Dimensional_Features_2	0.875	0.958	0.875	0.986	0.875	0.98
Dishes_and_Cutlery	0.625	0.944	0.750	0.931	0.625	0.87
Economics	0.750	0.819	1.000	1.000	1.000	1.00
Electronics	0.500	0.917	1.000	1.000	1.000	1.00
External_Body_Parts	1.000	1.000	0.625	0.903	0.875	0.98
Extreme_Natural_Events	0.875	0.986	0.500	0.931	0.875	0.97
Family_Members	1.000	1.000	1.000	1.000	1.000	1.00
Fantasy_Characters	0.750	0.861	0.875	0.944	0.375	0.87
Farm_Animals	0.750	0.931	0.750	0.889	0.500	0.72

0.000

0.875

0.875

0.000

0.750

0.875

0.875

0.875

0.875

0.625

1.000

0.500

0.000

0.500

0.667

0.986

0.875

0.569

0.972

0.944

0.931

0.986

0.972

0.847

1.000

0.903

0.556

0.944

0.875

0.875

0.375

1.000

0.750

0.500

0.875

0.875

0.750

0.875

0.625

0.250

0.500

0.500

0.986

0.972

0.861

1.000

0.958

0.903

0.986

0.958

0.958

0.875

0.931

0.903

0.653

0.889

0.875

0.750

0.250

1.000

0.000

0.500

0.750

0.500

0.500

1.000

0.500

0.875

0.000

0.500

0.986

0.847

0.792

1.000

0.514

0.931

0.958

0.944

0.903

1.000

0.917

0.972

0.625

0.903

Firearms

Flowers

Food\_Features

Fruit\_Trees

Furniture

Grain

Herbs

Gemstones

Hair\_Features

Free\_Time\_Activities

Flying\_Means\_of\_Transport

Fish

Food

Fruit

Table 4. Sketch Engine Thesaurus and Word Embeddings evaluation; accuracy and OPP computed per set, for Estonian

Human_Features_Negativity	0.750	0.958	0.750	0.972	0.875	0.986
Human_Features_Positivity	0.250	0.722	0.500	0.917	0.125	0.861
Human_Moods	0.250	0.778	0.625	0.931	0.625	0.931
Human_Physical_Features	0.500	0.778	0.875	0.986	1.000	1.000
Illnesses	0.875	0.972	0.875	0.986	0.750	0.972
Informatics	1.000	1.000	1.000	1.000	1.000	1.000
Internal_Body_Parts	0.500	0.847	0.250	0.778	0.375	0.861
Kitchenware	1.000	1.000	0.500	0.944	0.375	0.819
Landscape_Features	0.625	0.792	1.000	1.000	1.000	1.000
Languages	0.750	0.861	1.000	1.000	0.875	0.986
Linguistics	0.500	0.694	0.375	0.931	0.625	0.958
Liquid_Containers	0.500	0.667	0.375	0.764	0.375	0.778
Materials	0.625	0.861	0.500	0.889	0.500	0.889
Maths	1.000	1.000	0.375	0.847	0.500	0.847
Means_of_Transport	0.625	0.889	0.750	0.972	0.750	0.958
Medicine	1.000	1.000	0.250	0.917	0.625	0.958
Metals	0.875	0.986	0.625	0.958	0.875	0.986
Music	0.625	0.750	1.000	1.000	1.000	1.000
Music_Genres	1.000	1.000	1.000	1.000	0.875	0.972
Musical_Instruments	1.000	1.000	1.000	1.000	0.000	0.778
Non-alcoholic_Drinks	0.375	0.653	0.375	0.833	0.250	0.833
Nuts	0.000	0.694	0.000	0.722	0.375	0.861
Office_Supplies	0.875	0.972	0.625	0.958	0.875	0.986
Parts_of_Head	0.750	0.903	0.750	0.917	0.625	0.819
Parts_of_House	0.750	0.944	0.250	0.917	0.250	0.917
Parts_of_Skeleton	0.750	0.847	0.375	0.875	0.625	0.944
Parts_of_Speech	1.000	1.000	0.750	0.972	0.500	0.861
Politics	0.750	0.917	0.875	0.986	1.000	1.000
Professions	0.375	0.778	0.625	0.764	0.500	0.792
Reptiles	0.500	0.778	0.500	0.750	0.500	0.833
Road_Means_of_Transport	0.875	0.903	1.000	1.000	0.875	0.875
Rooms_in_the_House	1.000	1.000	0.750	0.944	0.750	0.931
Savanna_Animals	1.000	1.000	0.500	0.847	0.375	0.889
School_Subjects	0.875	0.958	0.625	0.958	0.750	0.931
Shapes	1.000	1.000	0.875	0.972	0.500	0.847
Shoes	0.500	0.875	0.750	0.917	0.000	0.000
Shops	0.000	0.889	0.500	0.944	1.000	1.000
Sources_of_Energy	0.625	0.694	0.375	0.931	0.375	0.875
Spices	1.000	1.000	1.000	1.000	0.625	0.958
Spirits	0.500	0.833	0.750	0.972	0.750	0.972
Sport	1.000	1.000	1.000	1.000	1.000	1.000
Sports	0.000	0.833	1.000	1.000	1.000	1.000
Sweets	0.000	0.722	0.500	0.875	0.375	0.875
Temperature_Features	0.375	0.764	0.750	0.972	0.750	0.972
Textile_Fibres	0.750	0.972	1.000	1.000	0.625	0.819
Touch_Features	0.125	0.833	0.875	0.986	0.875	0.958
Trees	1.000	1.000	1.000	1.000	1.000	1.000
Units_of_Time	0.625	0.903	0.625	0.889	0.375	0.847
Vegetables	0.625	0.958	1.000	1.000	1.000	1.000

Verbs_Animal_Sounds	0.000	0.597	0.375	0.889	0.750	0.944
Verbs_Cognition	0.500	0.792	0.625	0.903	0.500	0.917
Verbs_Communication_1	0.750	0.903	0.750	0.833	0.875	0.931
Verbs_Communication_2	0.375	0.764	0.375	0.792	0.250	0.861
Verbs_Cooking_1	0.000	0.778	1.000	1.000	1.000	1.000
Verbs_Cooking_2	0.875	0.875	0.750	0.944	1.000	1.000
Verbs_Crime	0.000	0.778	0.750	0.931	0.500	0.889
Verbs_Destroy	0.125	0.625	0.625	0.847	0.500	0.819
Verbs_Dog	0.625	0.819	0.500	0.889	0.500	0.903
Verbs_Driving	0.125	0.736	0.875	0.931	0.750	0.958
Verbs_Eating	0.250	0.764	0.875	0.958	0.875	0.972
Verbs_Economics	0.125	0.611	0.500	0.944	0.875	0.986
Verbs_Farming	0.375	0.861	1.000	1.000	1.000	1.000
Verbs_Hair	0.000	0.514	0.625	0.819	0.625	0.889
Verbs_Human_Sounds	0.000	0.542	0.500	0.847	0.375	0.778
Verbs_Killing	0.250	0.569	0.625	0.958	0.625	0.875
Verbs_Measure	0.000	0.694	0.375	0.819	0.750	0.944
Verbs_Motion	0.375	0.778	0.875	0.972	1.000	1.000
Verbs_Mouth	0.250	0.778	0.375	0.889	0.500	0.847
Verbs_Music	0.000	0.597	0.875	0.986	0.875	0.972
Verbs_Perception	0.250	0.903	0.500	0.806	0.875	0.986
Verbs_Plants	0.000	0.653	0.875	0.986	0.875	0.986
Verbs_Psych	0.000	0.611	0.500	0.875	0.500	0.903
Verbs_Religion	0.000	0.625	0.875	0.958	0.625	0.931
Verbs_School	0.125	0.750	0.625	0.875	0.125	0.653
Verbs_Smell	0.125	0.764	0.750	0.972	0.875	0.986
Verbs_Sport	0.125	0.847	0.875	0.958	0.625	0.931
Verbs_Telephone	0.000	0.681	0.000	0.597	1.000	1.000
Verbs_Touch	0.000	0.361	0.500	0.875	0.250	0.639
Verbs_Weather	0.000	0.500	1.000	1.000	0.875	0.972
War	0.000	0.708	0.000	0.861	1.000	1.000
Water_Means_of_Transport	0.750	0.847	0.500	0.917	0.750	0.917
Weapons	0.750	0.819	0.875	0.972	0.625	0.944
Weather_Conditions	0.875	0.986	0.875	0.986	0.875	0.986
Weather_Events	1.000	1.000	1.000	1.000	1.000	1.000
Wild_Animals	0.875	0.944	0.875	0.931	0.750	0.833
Zodiac_Signs	0.750	0.917	1.000	1.000	0.000	0.778

	FR	FR	FR	FR	FR	FR
set name	SkeThe	SkeThe	WE_	WE_	WE_	WE_
	accuracy	OPP	word	word	lemma	lemma
			accuracy	OPP	accuracy	OPP
Art	0.500	0.778	0.750	0.903	0.875	0.931
Astronomical_Objects	0.875	0.972	0.250	0.667	0.250	0.625
Biomes	0.500	0.875	1.000	1.000	0.875	0.986
Birds	1.000	1.000	0.875	0.931	1.000	1.000
Bodies_of_Water	0.250	0.708	0.750	0.944	0.625	0.875
Book_Genres	0.375	0.611	0.375	0.694	0.250	0.667
Bugs	0.625	0.889	0.625	0.917	0.875	0.986
Building_Materials	1.000	1.000	0.875	0.986	0.750	0.931
Buildings	0.875	0.986	0.250	0.792	0.250	0.792
Car_Components	0.750	0.931	0.500	0.931	0.500	0.944
Chemical_Elements	0.750	0.958	0.500	0.917	0.500	0.833
Clothes	1.000	1.000	1.000	1.000	1.000	1.000
Colours	0.500	0.931	1.000	1.000	0.875	0.875
Computer_Components	0.375	0.653	0.750	0.861	0.625	0.903
Containers	0.500	0.819	0.625	0.833	0.375	0.764
Cooking	0.250	0.875	0.000	0.889	0.000	0.889
Dairy_Products	0.625	0.792	0.375	0.903	0.375	0.889
Dances	0.125	0.597	0.625	0.931	0.500	0.889
Dimensional_Features_1	0.500	0.861	1.000	1.000	1.000	1.000
Dimensional_Features_2	0.125	0.833	1.000	1.000	0.875	0.972
Dishes_and_Cutlery	0.250	0.903	0.625	0.833	0.625	0.861
Economics	0.000	0.750	0.625	0.958	0.875	0.986
Electronics	0.000	0.806	1.000	1.000	1.000	1.000
External_Body_Parts	0.750	0.958	1.000	1.000	0.875	0.986
Extreme_Natural_Events	1.000	1.000	0.625	0.889	0.750	0.917
Family_Members	0.000	0.889	0.000	0.889	0.000	0.889
Fantasy_Characters	0.750	0.903	0.250	0.889	0.375	0.833
Farm_Animals	0.500	0.889	0.375	0.806	0.625	0.917
Firearms	0.750	0.958	1.000	1.000	1.000	1.000
Fish	1.000	1.000	0.750	0.958	0.625	0.917
Flowers	0.625	0.889	0.500	0.861	0.250	0.847
Flying_Means_of_Transport	0.000	0.792	0.500	0.861	0.375	0.889
Food	0.500	0.931	0.750	0.944	0.875	0.944
Food_Features	0.625	0.806	0.625	0.931	0.000	0.611
Free_Time_Activities	0.125	0.667	0.625	0.931	0.625	0.931
Fruit	0.250	0.778	0.250	0.764	0.250	0.736
Fruit_Trees	0.000	0.806	0.000	0.875	0.000	0.819
Furniture	0.625	0.875	1.000	1.000	1.000	1.000
Gemstones	1.000	1.000	1.000	1.000	0.750	0.972
Grain	0.750	0.972	1.000	1.000	0.250	0.917
Hair_Features	0.000	0.708	0.500	0.819	0.000	0.417
Herbs	0.750	0.972	1.000	1.000	0.500	0.931
Human_Features_Negativity	0.750	0.958	0.750	0.931	0.875	0.972

Table 5. Sketch Engine Thesaurus and Word Embeddings evaluation; accuracy and OPP computed per set, for French

	0.105	0.770	0.050	0.002	0.000	0 (11
Human_Features_Positivity	0.125	0.778	0.250	0.903	0.000	0.611
Human_Moods	0.125	0.597	0.500	0.861	0.000	0.667
Human_Physical_Features	0.125	0.653	0.625	0.903	0.625	0.819
Illnesses	0.875	0.986	0.625	0.958	0.750	0.944
Informatics	0.125	0.542	0.750	0.972	0.750	0.972
Internal_Body_Parts	0.625	0.833	0.750	0.847	0.500	0.819
Kitchenware	0.000	0.611	0.625	0.875	0.125	0.764
Landscape_Features	0.500	0.889	0.500	0.944	1.000	1.000
Languages	1.000	1.000	1.000	1.000	0.750	0.750
Linguistics	0.125	0.625	1.000	1.000	1.000	1.000
Liquid_Containers	0.375	0.819	0.375	0.917	0.625	0.958
Materials	0.500	0.903	0.500	0.931	0.375	0.917
Maths	0.000	0.667	0.750	0.931	0.875	0.986
Means_of_Transport	0.250	0.806	0.250	0.764	0.250	0.792
Medicine	0.250	0.583	0.875	0.986	1.000	1.000
Metals	0.750	0.903	0.750	0.972	0.375	0.931
Music	0.125	0.764	0.125	0.847	0.375	0.875
Music_Genres	0.250	0.903	1.000	1.000	1.000	1.000
Musical_Instruments	0.500	0.917	0.875	0.986	0.750	0.958
Non-alcoholic_Drinks	0.000	0.569	0.125	0.819	0.250	0.833
Nuts	0.125	0.847	0.375	0.764	0.250	0.861
Office_Supplies	0.000	0.583	0.125	0.819	0.125	0.819
Parts_of_Head	0.500	0.903	0.125	0.764	0.250	0.861
Parts_of_House	0.125	0.903	0.875	0.986	0.875	0.944
Parts_of_Skeleton	1.000	1.000	0.750	0.972	0.625	0.958
Parts_of_Speech	0.125	0.708	0.375	0.847	0.375	0.861
Politics	0.125	0.903	1.000	1.000	1.000	1.000
Professions	0.000	0.667	0.875	0.944	0.875	0.986
Reptiles	0.625	0.833	0.625	0.833	0.625	0.861
Road_Means_of_Transport	0.125	0.792	0.750	0.944	0.750	0.944
Rooms_in_the_House	0.250	0.917	0.625	0.917	0.750	0.889
Savanna_Animals	0.750	0.972	0.875	0.986	0.875	0.958
School_Subjects	0.750	0.819	0.875	0.972	0.875	0.972
Shapes	0.500	0.931	0.500	0.861	0.625	0.861
Shoes	0.750	0.806	0.625	0.861	0.625	0.889
Shops	0.875	0.986	0.875	0.972	0.875	0.986
Sources_of_Energy	0.375	0.625	0.750	0.889	0.625	0.847
Spices	0.500	0.931	0.500	0.903	0.500	0.861
Spirits	0.500	0.847	1.000	1.000	0.625	0.958
Sport	0.750	0.875	1.000	1.000	1.000	1.000
Sports	0.500	0.931	0.625	0.944	0.625	0.903
Sweets	0.375	0.861	0.375	0.806	0.375	0.819
Temperature_Features	0.125	0.764	0.250	0.819	0.000	0.569
Textile_Fibres	0.500	0.819	1.000	1.000	0.500	0.944
Touch_Features	0.000	0.861	0.875	0.972	0.000	0.889
Trees	1.000	1.000	1.000	1.000	0.875	0.986
Units_of_Time	0.125	0.889	0.500	0.917	0.625	0.889
Vegetables	0.625	0.903	0.625	0.931	0.500	0.875
Verbs_Animal_Sounds	0.250	0.819	0.875	0.986	0.500	0.889

Verbs_Cognition	0.500	0.806	0.750	0.958	0.375	0.847
Verbs_Communication_1	0.125	0.833	0.750	0.861	0.750	0.847
Verbs_Communication_2	0.875	0.986	0.375	0.792	0.875	0.958
Verbs_Cooking_1	0.375	0.861	1.000	1.000	1.000	1.000
Verbs_Cooking_2	0.875	0.986	1.000	1.000	0.750	0.972
Verbs_Crime	0.000	0.556	0.625	0.917	0.500	0.806
Verbs_Destroy	0.125	0.736	0.375	0.764	0.375	0.750
Verbs_Dog	0.000	0.611	0.500	0.833	0.250	0.778
Verbs_Driving	0.375	0.847	0.125	0.847	0.250	0.903
Verbs_Eating	0.125	0.847	0.875	0.917	0.750	0.889
Verbs_Economics	0.375	0.806	0.250	0.917	1.000	1.000
Verbs_Farming	0.250	0.806	1.000	1.000	1.000	1.000
Verbs_Hair	0.000	0.736	0.875	0.917	0.750	0.972
Verbs_Human_Sounds	0.250	0.653	0.625	0.931	0.125	0.847
Verbs_Killing	0.125	0.736	0.375	0.903	0.375	0.819
Verbs_Measure	0.375	0.861	0.625	0.931	0.125	0.889
Verbs_Motion	0.750	0.931	0.125	0.903	0.875	0.986
Verbs_Mouth	0.000	0.875	0.500	0.931	1.000	1.000
Verbs_Music	0.125	0.639	0.500	0.861	0.375	0.833
Verbs_Perception	0.125	0.833	0.375	0.653	0.250	0.722
Verbs_Plants	0.000	0.417	0.875	0.972	1.000	1.000
Verbs_Psych	0.250	0.819	0.250	0.833	0.250	0.833
Verbs_Religion	0.500	0.847	0.500	0.903	0.500	0.819
Verbs_School	0.125	0.847	0.125	0.750	0.500	0.792
Verbs_Smell	0.125	0.431	0.250	0.736	0.250	0.792
Verbs_Sport	0.250	0.722	0.750	0.958	0.750	0.875
Verbs_Telephone	0.000	0.611	0.250	0.903	0.250	0.889
Verbs_Touch	0.500	0.778	0.125	0.806	0.250	0.764
Verbs_Weather	0.000	0.542	0.875	0.972	0.750	0.931
War	0.000	0.653	0.625	0.944	0.875	0.986
Water_Means_of_Transport	0.500	0.792	0.250	0.875	0.250	0.819
Weapons	0.500	0.875	0.875	0.986	0.750	0.931
Weather_Conditions	0.250	0.917	0.875	0.986	0.000	0.764
Weather_Events	0.500	0.931	1.000	1.000	1.000	1.000
Wild_Animals	0.250	0.847	0.375	0.931	0.375	0.917
Zodiac_Signs	0.375	0.792	0.125	0.806	0.625	0.931

	IT	IT	IT	IT	IT	IT
set name	SkeThe	SkeThe	WE_	WE_	WE_	WE_
	accuracy	OPP	word	word	lemma	lemma
			accuracy	OPP	accuracy	OPP
Art	0.375	0.806	0.500	0.931	0.750	0.958
Astronomical_Objects	0.250	0.875	0.375	0.819	0.375	0.806
Biomes	0.500	0.903	0.875	0.986	0.625	0.958
Birds	1.000	1.000	0.750	0.861	0.750	0.972
Bodies_of_Water	0.250	0.708	0.875	0.986	0.875	0.986
Book_Genres	0.375	0.792	0.500	0.806	0.625	0.778
Bugs	0.625	0.903	0.500	0.792	0.500	0.931
Building_Materials	0.625	0.958	1.000	1.000	0.875	0.944
Buildings	0.625	0.875	0.375	0.681	0.375	0.708
Car_Components	1.000	1.000	0.875	0.986	0.875	0.986
Chemical_Elements	0.500	0.861	0.250	0.819	0.000	0.681
Clothes	1.000	1.000	0.875	0.986	0.000	0.681
Colours	0.500	0.903	0.875	0.958	0.625	0.819
Computer_Components	0.375	0.611	0.375	0.819	0.375	0.806
Containers	0.625	0.875	0.375	0.875	0.250	0.861
Cooking	0.500	0.903	0.750	0.972	0.875	0.986
Dairy_Products	0.875	0.958	0.750	0.931	0.750	0.875
Dances	0.125	0.569	0.375	0.833	0.375	0.806
Dimensional_Features_1	0.625	0.944	0.500	0.944	1.000	1.000
Dimensional_Features_2	1.000	1.000	1.000	1.000	0.875	0.986
Dishes_and_Cutlery	1.000	1.000	0.750	0.944	0.750	0.931
Economics	0.250	0.806	1.000	1.000	1.000	1.000
Electronics	0.125	0.819	1.000	1.000	1.000	1.000
External_Body_Parts	0.625	0.944	0.625	0.958	1.000	1.000
Extreme_Natural_Events	0.875	0.944	0.625	0.944	0.625	0.931
Family_Members	1.000	1.000	1.000	1.000	1.000	1.000
Fantasy_Characters	0.375	0.806	0.250	0.819	0.375	0.875
Farm_Animals	0.500	0.875	0.500	0.931	0.250	0.917
Firearms	0.750	0.972	1.000	1.000	0.875	0.958
Fish	0.750	0.972	0.875	0.986	0.875	0.986
Flowers	0.250	0.917	0.500	0.847	0.250	0.847
Flying_Means_of_Transport	0.125	0.875	0.750	0.917	0.750	0.931
Food	0.875	0.986	0.875	0.958	0.000	0.875
Food_Features	0.500	0.736	0.375	0.764	0.375	0.667
Free_Time_Activities	0.000	0.486	0.250	0.861	0.375	0.806
Fruit	0.625	0.958	0.375	0.931	0.250	0.792
Fruit_Trees	0.000	0.792	0.625	0.944	0.000	0.833
Furniture	0.750	0.917	1.000	1.000	0.875	0.986
Gemstones	0.250	0.736	0.875	0.986	0.875	0.986
Grain	0.125	0.792	0.625	0.944	0.000	0.681
Hair_Features	0.000	0.708	0.625	0.875	0.000	0.764
Herbs	0.250	0.903	0.625	0.958	0.500	0.889
Human_Features_Negativity	1.000	1.000	1.000	1.000	1.000	1.000

Table 6. Sketch Engine Thesaurus and Word Embeddings evaluation; accuracy and OPP computed per set, for Italian

	0.105	0.700	0.050	0.007	0.105	0.700
Human_Features_Positivity	0.125	0.708	0.250	0.806	0.125	0.722
Human_Moods	0.625	0.903	0.500	0.944	0.000	0.847
Human_Physical_Features	0.000	0.583	0.625	0.917	0.750	0.917
Illnesses	0.750	0.931	0.750	0.958	0.750	0.944
Informatics	0.250	0.750	1.000	1.000	1.000	1.000
Internal_Body_Parts	0.750	0.972	0.750	0.944	0.750	0.958
Kitchenware	0.500	0.861	0.250	0.778	0.000	0.611
Landscape_Features	0.500	0.861	0.750	0.972	1.000	1.000
Languages	1.000	1.000	1.000	1.000	1.000	1.000
Linguistics	0.125	0.611	1.000	1.000	1.000	1.000
Liquid_Containers	0.500	0.917	0.750	0.944	0.625	0.931
Materials	0.375	0.847	1.000	1.000	0.625	0.944
Maths	0.500	0.806	0.625	0.903	0.500	0.819
Means_of_Transport	0.375	0.819	0.625	0.889	0.500	0.917
Medicine	0.250	0.694	1.000	1.000	1.000	1.000
Metals	0.375	0.792	0.625	0.944	0.000	0.722
Music	0.125	0.750	0.375	0.931	1.000	1.000
Music_Genres	0.125	0.903	0.750	0.972	0.750	0.931
Musical_Instruments	0.875	0.972	0.875	0.986	0.625	0.847
Non-alcoholic_Drinks	0.125	0.681	0.625	0.806	0.375	0.889
Nuts	0.250	0.764	0.375	0.819	0.000	0.750
Office_Supplies	0.000	0.653	0.750	0.917	0.000	0.819
Parts_of_Head	0.875	0.958	0.125	0.889	0.125	0.889
Parts_of_House	0.750	0.972	0.625	0.875	0.000	0.819
Parts_of_Skeleton	1.000	1.000	0.625	0.917	0.750	0.944
Parts_of_Speech	0.250	0.736	0.375	0.861	0.375	0.861
Politics	0.250	0.694	1.000	1.000	1.000	1.000
Professions	0.250	0.736	0.750	0.972	0.750	0.958
Reptiles	0.625	0.819	0.750	0.875	0.875	0.931
Road_Means_of_Transport	0.375	0.931	1.000	1.000	0.500	0.722
Rooms_in_the_House	0.125	0.833	0.375	0.847	0.375	0.819
Savanna_Animals	0.125	0.819	0.500	0.889	0.875	0.944
School_Subjects	0.375	0.792	0.875	0.958	0.000	0.639
Shapes	0.250	0.917	0.375	0.806	0.375	0.764
Shoes	0.125	0.653	0.625	0.806	0.000	0.444
Shops	0.250	0.889	0.750	0.972	0.375	0.931
Sources_of_Energy	0.500	0.708	0.625	0.861	0.625	0.861
Spices	0.125	0.736	0.625	0.958	0.500	0.944
Spirits	0.625	0.889	1.000	1.000	0.875	0.875
Sport	0.500	0.833	0.875	0.986	0.875	0.875
Sports	0.125	0.861	0.625	0.958	0.000	0.819
Sweets	0.625	0.847	0.625	0.958	0.875	0.986
Temperature_Features	0.375	0.833	0.750	0.903	0.875	0.958
Textile_Fibres	0.125	0.764	0.500	0.944	0.000	0.667
Touch_Features	0.000	0.653	0.625	0.958	0.000	0.847
Trees	0.250	0.917	1.000	1.000	1.000	1.000
Units_of_Time	0.125	0.792	0.750	0.889	0.625	0.875
Vegetables	0.000	0.847	0.875	0.986	0.000	0.861
Verbs_Animal_Sounds	0.250	0.681	0.500	0.944	0.625	0.875

Verbs_Cognition	0.625	0.903	0.875	0.972	0.500	0.931
Verbs_Communication_1	0.875	0.972	0.875	0.972	0.875	0.972
Verbs_Communication_2	0.750	0.958	0.875	0.986	0.750	0.972
Verbs_Cooking_1	1.000	1.000	0.750	0.972	1.000	1.000
Verbs_Cooking_2	0.875	0.986	1.000	1.000	1.000	1.000
Verbs_Crime	0.000	0.667	0.875	0.944	0.625	0.958
Verbs_Destroy	0.000	0.764	0.750	0.903	0.375	0.861
Verbs_Dog	0.500	0.833	0.500	0.861	0.500	0.847
Verbs_Driving	0.125	0.694	0.750	0.917	0.750	0.903
Verbs_Eating	0.125	0.833	0.875	0.903	0.000	0.792
Verbs_Economics	0.375	0.806	0.625	0.958	0.875	0.972
Verbs_Farming	0.250	0.708	0.875	0.986	0.375	0.889
Verbs_Hair	0.625	0.847	0.375	0.903	0.375	0.917
Verbs_Human_Sounds	0.125	0.667	0.625	0.903	0.500	0.917
Verbs_Killing	0.125	0.833	0.375	0.875	0.375	0.750
Verbs_Measure	0.000	0.764	0.875	0.986	1.000	1.000
Verbs_Motion	0.875	0.944	0.375	0.903	0.875	0.958
Verbs_Mouth	0.000	0.847	0.875	0.958	0.500	0.847
Verbs_Music	0.000	0.569	0.375	0.903	0.125	0.819
Verbs_Perception	0.250	0.778	0.750	0.903	0.625	0.861
Verbs_Plants	0.000	0.597	1.000	1.000	0.000	0.889
Verbs_Psych	0.750	0.931	0.375	0.833	0.375	0.903
Verbs_Religion	0.500	0.875	0.500	0.931	0.375	0.889
Verbs_School	0.375	0.833	0.250	0.847	0.375	0.819
Verbs_Smell	0.250	0.819	0.625	0.917	0.250	0.861
Verbs_Sport	0.125	0.736	0.875	0.958	1.000	1.000
Verbs_Telephone	0.000	0.556	0.375	0.833	0.250	0.833
Verbs_Touch	0.375	0.861	0.375	0.819	0.625	0.903
Verbs_Weather	0.000	0.444	0.500	0.917	0.375	0.931
War	0.000	0.625	0.875	0.986	1.000	1.000
Water_Means_of_Transport	0.375	0.764	0.750	0.972	0.875	0.972
Weapons	0.625	0.875	0.500	0.875	0.875	0.931
Weather_Conditions	0.875	0.986	1.000	1.000	1.000	1.000
Weather_Events	0.875	0.972	0.250	0.903	0.750	0.972
Wild_Animals	0.125	0.903	0.750	0.972	0.000	0.694
Zodiac_Signs	0.375	0.681	0.250	0.681	0.000	0.681

	SK	SK	SK	SK	SK	SK
set name	SkeThe	SkeThe	WE_	WE_	$WE_{-}$	WE_
	accuracy	OPP	word	word	lemma	lemma
			accuracy	OPP	accuracy	OPP
Art	0.375	0.750	0.750	0.903	0.750	0.861
Astronomical_Objects	0.125	0.792	0.375	0.819	0.250	0.792
Biomes	0.000	0.722	0.750	0.972	1.000	1.000
Birds	1.000	1.000	0.625	0.861	0.875	0.986
Bodies_of_Water	0.375	0.819	0.875	0.931	0.875	0.944
Book_Genres	0.750	0.944	0.375	0.833	0.250	0.847
Bugs	0.625	0.806	0.750	0.917	0.625	0.917
Building_Materials	0.875	0.986	0.875	0.972	0.875	0.958
Buildings	0.250	0.903	0.625	0.819	0.125	0.778
Car_Components	0.625	0.875	0.875	0.986	0.750	0.944
Chemical_Elements	0.750	0.972	1.000	1.000	0.750	0.944
Clothes	0.875	0.986	1.000	1.000	1.000	1.000
Colours	0.875	0.986	1.000	1.000	1.000	1.000
Computer_Components	0.000	0.542	0.625	0.861	0.625	0.792
Containers	0.500	0.833	0.250	0.806	0.625	0.806
Cooking	0.375	0.750	0.375	0.931	0.625	0.931
Dairy_Products	1.000	1.000	0.750	0.806	0.625	0.819
Dances	0.500	0.667	0.375	0.917	0.500	0.903
Dimensional_Features_1	0.875	0.986	1.000	1.000	1.000	1.000
Dimensional_Features_2	0.375	0.889	0.625	0.875	0.875	0.944
Dishes_and_Cutlery	0.875	0.986	0.500	0.917	0.625	0.931
Economics	0.375	0.778	1.000	1.000	0.625	0.958
Electronics	0.125	0.903	1.000	1.000	1.000	1.000
External_Body_Parts	0.750	0.861	1.000	1.000	1.000	1.000
Extreme_Natural_Events	0.750	0.972	0.750	0.958	0.625	0.944
Family_Members	1.000	1.000	1.000	1.000	1.000	1.000
Fantasy_Characters	0.500	0.903	0.375	0.861	0.500	0.847
Farm_Animals	0.625	0.944	0.750	0.847	0.375	0.847
Firearms	0.000	0.861	0.875	0.986	0.750	0.972
Fish	0.000	0.889	1.000	1.000	0.875	0.986
Flowers	0.000	0.722	0.625	0.931	0.500	0.903
Flying_Means_of_Transport	0.375	0.889	0.875	0.986	0.875	0.986
Food	0.875	0.986	0.750	0.931	0.750	0.958
Food_Features	0.250	0.875	0.500	0.819	0.500	0.875
Free_Time_Activities	0.125	0.764	0.750	0.958	0.750	0.903
Fruit	0.750	0.972	0.375	0.778	0.500	0.889
Fruit_Trees	0.000	0.889	0.875	0.986	1.000	1.000
Furniture	0.250	0.917	0.875	0.986	1.000	1.000
Gemstones	1.000	1.000	0.750	0.972	1.000	1.000
Grain	0.500	0.833	0.750	0.972	0.500	0.944
Hair_Features	0.000	0.403	0.375	0.653	0.500	0.806
Herbs	1.000	1.000	1.000	1.000	1.000	1.000
Human_Features_Negativity	0.500	0.917	1.000	1.000	0.875	0.986

Table 7. Sketch Engine Thesaurus and Word Embeddings evaluation; accuracy and OPP computed per set, for Slovak

Human_Features_Positivity	0.375	0.889	0.500	0.917	0.375	0.931
Human_Moods	0.375	0.833	0.500	0.931	0.625	0.903
Human_Physical_Features	0.000	0.444	0.750	0.931	0.500	0.931
Illnesses	0.375	0.931	0.625	0.958	0.625	0.958
Informatics	0.375	0.681	0.750	0.944	0.875	0.972
Internal_Body_Parts	0.000	0.889	0.875	0.972	0.875	0.972
Kitchenware	0.000	0.736	1.000	1.000	0.375	0.875
Landscape_Features	0.750	0.958	0.625	0.958	1.000	1.000
Languages	0.000	0.889	1.000	1.000	0.750	0.750
Linguistics	0.375	0.653	0.875	0.986	0.875	0.986
Liquid_Containers	0.500	0.903	0.625	0.944	0.625	0.931
Materials	1.000	1.000	0.750	0.958	0.750	0.931
Maths	0.500	0.694	0.625	0.931	0.750	0.944
Means_of_Transport	0.500	0.903	0.750	0.972	0.750	0.972
Medicine	0.375	0.778	0.875	0.986	1.000	1.000
Metals	0.875	0.986	0.750	0.972	0.750	0.972
Music	0.125	0.625	1.000	1.000	1.000	1.000
Music_Genres	0.500	0.944	1.000	1.000	0.875	0.958
Musical_Instruments	1.000	1.000	1.000	1.000	0.750	0.861
Non-alcoholic_Drinks	0.250	0.736	0.750	0.889	0.625	0.875
Nuts	0.000	0.556	0.375	0.861	0.750	0.944
Office_Supplies	0.000	0.597	0.625	0.875	0.625	0.889
Parts_of_Head	0.625	0.778	0.375	0.708	0.375	0.819
Parts_of_House	0.625	0.944	0.250	0.889	0.375	0.931
Parts_of_Skeleton	0.125	0.431	0.375	0.569	0.375	0.667
Parts_of_Speech	0.750	0.875	0.375	0.833	0.375	0.778
Politics	0.250	0.847	1.000	1.000	1.000	1.000
Professions	0.500	0.861	0.875	0.972	0.875	0.972
Reptiles	0.000	0.722	0.625	0.875	0.750	0.903
Road_Means_of_Transport	0.500	0.778	0.000	0.889	0.125	0.903
Rooms_in_the_House	0.250	0.917	0.875	0.972	0.625	0.944
Savanna_Animals	0.000	0.750	0.500	0.708	0.750	0.903
School_Subjects	1.000	1.000	1.000	1.000	0.750	0.972
Shapes	0.375	0.875	0.875	0.958	0.875	0.972
Shoes	0.000	0.764	0.500	0.833	0.000	0.556
Shops	0.875	0.972	1.000	1.000	1.000	1.000
Sources_of_Energy	0.375	0.778	0.500	0.903	0.625	0.861
Spices	0.500	0.861	0.875	0.986	0.750	0.944
Spirits	0.500	0.944	1.000	1.000	1.000	1.000
Sport	0.625	0.847	0.750	0.958	0.875	0.986
Sports	0.750	0.931	0.750	0.903	0.750	0.958
Sweets	0.250	0.847	0.625	0.917	0.750	0.972
Temperature_Features	0.000	0.694	0.375	0.847	0.375	0.847
Textile_Fibres	0.750	0.972	0.375	0.931	0.375	0.889
Touch_Features	0.125	0.667	0.625	0.819	0.625	0.944
Trees	1.000	1.000	1.000	1.000	1.000	1.000
Units_of_Time	0.250	0.917	0.750	0.931	0.875	0.958
Vegetables	0.875	0.972	1.000	1.000	0.875	0.958
Verbs_Animal_Sounds	0.000	0.361	0.000	0.722	0.375	0.889

Verbs_Cognition	0.375	0.736	0.625	0.875	0.375	0.833
Verbs_Communication_1	0.250	0.667	0.500	0.861	0.625	0.889
Verbs_Communication_2	0.875	0.958	0.250	0.806	0.750	0.972
Verbs_Cooking_1	0.875	0.972	1.000	1.000	1.000	1.000
Verbs_Cooking_2	1.000	1.000	0.875	0.986	1.000	1.000
Verbs_Crime	0.375	0.847	0.125	0.806	0.250	0.833
Verbs_Destroy	0.250	0.694	0.750	0.931	0.500	0.861
Verbs_Dog	0.000	0.722	0.000	0.694	0.500	0.847
Verbs_Driving	0.375	0.861	0.250	0.750	0.500	0.792
Verbs_Eating	0.625	0.778	0.500	0.917	0.250	0.847
Verbs_Economics	0.375	0.653	0.500	0.944	0.500	0.917
Verbs_Farming	0.375	0.875	1.000	1.000	1.000	1.000
Verbs_Hair	0.000	0.667	0.250	0.792	0.250	0.819
Verbs_Human_Sounds	0.000	0.222	0.375	0.708	0.375	0.736
Verbs_Killing	0.875	0.931	0.375	0.875	0.375	0.931
Verbs_Measure	0.875	0.972	0.625	0.958	0.625	0.944
Verbs_Motion	0.250	0.889	1.000	1.000	0.875	0.958
Verbs_Mouth	0.000	0.486	1.000	1.000	0.750	0.944
Verbs_Music	0.125	0.528	0.500	0.944	0.250	0.917
Verbs_Perception	1.000	1.000	0.875	0.986	1.000	1.000
Verbs_Plants	0.625	0.931	1.000	1.000	1.000	1.000
Verbs_Psych	0.000	0.375	0.500	0.764	1.000	1.000
Verbs_Religion	0.500	0.778	0.375	0.917	0.375	0.917
Verbs_School	0.375	0.722	0.375	0.764	0.375	0.792
Verbs_Smell	0.125	0.625	0.125	0.736	0.125	0.736
Verbs_Sport	0.750	0.944	1.000	1.000	0.750	0.917
Verbs_Telephone	0.000	0.708	0.125	0.625	0.250	0.750
Verbs_Touch	0.000	0.736	0.250	0.806	0.500	0.903
Verbs_Weather	0.750	0.958	0.625	0.944	0.750	0.972
War	0.250	0.556	0.625	0.903	0.375	0.931
Water_Means_of_Transport	0.625	0.875	0.750	0.958	0.750	0.903
Weapons	0.000	0.847	0.750	0.958	0.750	0.931
Weather_Conditions	0.000	0.778	0.875	0.986	0.875	0.986
Weather_Events	0.750	0.958	0.875	0.931	0.875	0.917
Wild_Animals	0.500	0.889	0.875	0.972	0.875	0.958
Zodiac_Signs	0.625	0.958	0.125	0.847	0.875	0.972

## Acknowledgments

It is time to thank all the – numerous – people who were involved in this thesis project, even with the smallest contribution.

I would like to thank Professor Elisabetta Jezek, who has been my guide through the last four years, since my Bachelor Thesis in Applied Linguistics. Thanks to her, I became passionate about Linguistics and decided to take this route in my university career. Also, through her I had the chance to learn about Lexical Computing and undertake the Erasmus+ Traineeship which led me to this thesis project. I also want to thank her for all her comments and reviews as the supervisor of this thesis, which helped me improve it.

Lexical Computing hosted me for five months. I would like to thank all the people I met there and made me feel welcome in the family, especially those who contributed to this thesis. First, Miloš Jakubíček, this thesis' co-supervisor, who proposed me this project and followed me in its various phases. I would also like to thank him for accepting me at the company and for being my beer-mate in all those months. Second, I'd like to thank Jan Kraus, my desk mate and favourite companion when it came to trying new cuisines and restaurants in Brno. Together with Ota Mikušek and Vlasta Ohlídalová, he was a precious help in the translation of the dataset in Czech. I'd also like to thank Ota Mikušek for his support with the experiment on the Czech students and the processing of the results. Ondřej Herman helped me with the evaluation of the distributional models, providing me the results to analyse. Finally, I would like to mention all the people at Lexical Computing with whom I have collaborated outside this thesis project or were part of my stay in Brno: Pavel Rychlý, Vojtěch Kovář, Tereza Olšanová, Ondřej Matuška, Vít Suchomel, Jan Michelfeit, Jan Bušta, Michal Cukr, Marek Medveď, Tomáš Svoboda, Marek Blahuš, Michal Němec.

Kristina Koppel and Michaela Denisová were a precious help in the translation of the Estonian and Slovak part of the dataset. Sofia Romani, my sister, took care of the German and French parts.

Finally, I would also like to thank all the University of Pavia students attending Professor Jezek course who took part to the human evaluation experiment: Lucia Volpi, Agnese, Elena Fioraliso, Maddalena Bressler, Elena Didoni, Sara Periti, Marco Brandizzi, Luca Briatico, Francesca Torchio, Ilaria Barzon, Petra, Silvia Cangiano, Michelle Carnevale, Luca Dogliani, Eleonora, Federica Fiore, Elena Tamburini.

## Ringraziamenti

Si dice che Peter Mark Roget, l'inventore del primo *thesaurus* nel senso attuale, avesse una ossessione per la compilazione di liste e che questo l'abbia aiutato a gestire la sua depressione. Sicuramente le liste hanno avuto un ruolo fondamentale per me nel portare a termine la stesura di questa tesi e nel riportarmi sulla strada quando ero sul punto di gettare tutto all'aria. Pertanto, ringrazio l'esistenza di Google Tasks e tutte le liste incredibilmente dettagliate che vi ho accuratamente salvato e puntualmente depennato man mano procedevo nella scrittura.

Ringrazio tutta la mia famiglia, per il supporto in questi lunghi anni universitari: i miei genitori, mia sorella Sofia e le nonne, Tatiana ed Enrica, per l'orgoglio nei miei confronti e per la motivazione che mi hanno dato ad andare avanti.

Ringrazio la mia famiglia universitaria di linguisti, sempre più grande: Marta, Luca, Viola, Emma, Valerio. Questi anni di magistrale a distanza non sarebbero stati gli stessi senza la vostra compagnia e il vostro affetto.

Ringrazio tutti i miei amici, che hanno reso più leggero e spensierato questo ultimo periodo di scrittura: Lucia, Elisabetta, Ingrid, Rita, Mattia, Thy Van, Filippo.

Ringrazio i miei due compagni di sventure espatriati con me a Brno, Ilaria ed Edoardo.

Ringrazio Diletta, il cui supporto e ascolto è stato fondamentale per riuscire a trovare la motivazione per non gettare la spugna nei momenti più difficili di questo progetto di tesi.

Infine, ringrazio Andrea, il mio vero *thesaurus*, o meglio, l'unico elemento di ordine nella mia vita caotica, per la forza e la serenità che mi hai trasmesso in tutti i drammi esistenziali che questa tesi mi ha causato. Grazie perché ci sei sempre e perché posso contare sul tuo amore, la tua fiducia e il tuo supporto in ogni momento della mia vita.