Phraseology in children's literature: a contrastive analysis

Verkade, S.A.

4 METHODOLOGY

The scope of this dissertation is to analyse Dutch and Italian phraseological units in their pragmatic context, using a corpus of Children's Literature. The reasons to carry out such a detailed contrastive analysis have been discussed in the preceding chapters. In this chapter we will outline the process we have followed to carry out our analyses.

The corpus of this research (see §4.1.) consists in the Dutch children's book *Wiplala*, written by Annie M.G. Schmidt, and its Italian translation, which will be compared in a bidirectional way. Extant studies mostly adopt a unidirectional approach, in which (some characteristics of) one language function(s) as a starting point to describe the differences and similarities of (those characteristics in) the other language. Yet, the findings of these studies are not necessarily reversible, as only one point of view has been adopted[51]. Bidirectional (and multidirectional) studies, like the present one, overcome this limitation by confronting the languages as autonomous systems. Hence, the *tertium comparationis* is not one of the languages involved in the analysis, but rather a set

[51] This can be illustrated by recalling Saussure's (1916: 166) famous example of *sheep* – *mutton* in English and *mouton* in French: while *sheep* can be used as a (partially) equivalent translatant for *mouton*, the contrary is not necessarily true, as *mouton* can be used in significantly different contexts and can indicate both *sheep* and *mutton*. See Koesters Gensini (2020: 31–32) for a detailed discussion of an Italian-German example regarding unidirectional analysis.

of predefined parameters. The Dutch "starting text" will thus be confronted with the Italian "arrival text", and vice versa. Aware of this uncommon terminology, let it be clear it is a conscious choice not to refer to our corpus in terms of "source text" and "target text". In the case of bi- or multidirectional analyses the translation (i.e. the former target text) also becomes the starting text and the former source text becomes the arrival text (which is thus not always a "target text" in the pure sense).

As a first step, we have read both the Dutch and the Italian text, to get a full understanding of the story and to be able to recognise any foreshadowing. Next, we have gone through the starting text again, highlighting all phraseological units. The following step has been the insertion and annotation of the phraseological units present in the starting text on the CREAMY platform (see §4.2.), followed by their respective "translatants" (i.e. the portion of an arrival text that corresponds to the PU present in the starting text[52]; TLs). In the second phase, we have followed these steps again for the Italian starting text and Dutch arrival text.

We have prepared a small parallel corpus by aligning both texts (see §4.3.), an extremely helpful tool when double checking if every single occurrence of all phraseological units had been inserted and annotated. As CREAMY does not yet dispose of advanced search and analysis options (see §4.2.3. for the options it offers), it was necessary to prepare Excel documents in which phraseological units and translatants remained linked. This linkage, in fact, is one of the big advantages of CREAMY. For each Excel file, numerous sheets were prepared to carry out the quantitative analysis summarised in Chapter 5.

[52] The Dutch portions of text corresponding to Italian phraseological units (hence in the inverted perspective where the Dutch original text becomes the arrival text), will also be referred to as "translatants" – even if they are not truly "translations". While other terms like "original construction" or "source construction" have been debated, these could have led to confusion regarding the perspective of the analysis.

These steps will be highlighted in the following paragraphs, starting from the motivations for choosing this particular corpus (§4.1.). The CREAMY platform will be thoroughly discussed in §4.2.; in this paragraph, the general functioning of the platform, the description of new PUs and translatants (including the classification implemented in this dissertation), and the search and analysis options will be described. In the last paragraph (§4.3.) other research tools will be discussed, including the method used for the alignment of the texts, and the various Excel sheets.

4.1. Corpus

The peculiarities of Children's Literature and its importance in providing opportunities for phraseological analysis have already been discussed in Chapter 3. The corpus of this research is a Dutch children's book, *Wiplala*, and its Italian translation. While the corpus is small and obviously inadequate to provide a basis for the identification of a "core" of a phraseological inventory, it can be a stepping stone for further research. Several reasons came into play in our decision not to add an Italian source text (and its Dutch translation) to our corpus. First, the detailed analysis of each occurrence of every phraseological unit and respective translatant, is a very time-consuming process – especially if the corpus is to be studied bidirectionally. Analysing a larger corpus would have meant spending less time on the detailed annotation of the PUs present in all texts (both original(s) and translations), which is pivotal for this research as a whole. Leaving aside any of the parameters implemented in the analysis, would have meant abandoning the goal of describing the full denotative and connotative meaning of PUs in their co-text, and their use. A research limited to a select type of PU (e.g. only idioms) or a structural composition (e.g. only light verb constructions) would have had a completely different scope. On the other hand, the selected corpus is large enough to be able to make a contribution to both

contrastive phraseology and Translation Studies, offering a first outlook on further research possibilities.

In the following paragraphs we will first discuss the original text, including the author, plot and different editions (§4.1.1.) and then the Italian translation (§4.1.2.).

4.1.1. *Wiplala* – Annie M.G. Schmidt

Wiplala is a children's book written by the Dutch author Anna Maria Geertruida Schmidt, commonly known as Annie M.G. Schmidt (1911 – 1995). Her works have accompanied (and continue to accompany) generations of both Dutch-speaking children and adults. Besides children's books, the author has also written short stories, poems, plays, songs, musicals, radio and television scripts; she has won several prizes in different genres[53]. She is included in the *Canon van Nederland* and referred to as Poet Laureate ('*Dichteres des Vaderlands*') *avant la lettre*[54]. Her oeuvre is considered an important contribution to the development of the Dutch language.

The edition used for this research was published by Em. Querido's Uitgeverij in 1991 and is part of the Querido junior series. The illustrations by Jenny Dalenoord have been adopted from the first edition published by De Arbeiderspers in 1957. The illustrations in the first pages depict the Blom family on whom the plot is centred. Chapter one starts on page 8. The short novel proper is 157 pages long.

Another resource has been the e-book (2014, 43rd reprint, based on the 42nd reprint), that has made it very easy to search the text and select specific

[53] For instance, in the field of CL: Schmidt won the Hans Christian Andersen Award for her important and long lasting contribution to Children's Literature. *Wiplala* won the award for best Dutch children's book of 1957.

[54] The Canon of Dutch History is a list of the fifty "themes" that summarise the history of the Netherlands, and ranges from Charlemagne to Erasmus, Aletta Jacobs, slavery, the world wars and the advent of television. See *Canon of the Netherlands* (2020) and *Annie M.G. Schmidt: Dichteres des Vaderlands avant la lettre* accessed 14-01-2023).

parts. The extraction of PUs and TLs, however, is based entirely on the 1991 print, as the e-book often deviates from the original text. The 2014 digital version, for example, reads "*Nou, ik zal maar eens gauw aan de slag gaan,' zei juffrouw Dingemans* [...]." while the 1991 paper reprint reads *redderen* 'to clean, to tidy up' instead of *aan de slag gaan* 'to start working on something' (Schmidt 1991: 53, 2014: 42/125). In this case the editors might have decided that the verb *redderen* was not accessible for children anymore.

Wiplala is classified as a B-type book in Dutch libraries: fit for children from approximately nine to twelve years old (see §3.1.). This label takes the average social-emotional development of children and their reading level into account. The novel, however, is clearly fit for younger children as well: some editions state that it can be read to children from approximately five years of age.

The book is named after one of its main characters, a gnome of a kind referred to as a *wiplala*, whose name is also Wiplala. He ends up in the house of the Blom family, where mister Blom and his children Nelly Delly and Johannes live, and gets caught by their cat Fly. Fearing that the cat will kill him, Wiplala turns it to stone. Nelly Delly then finds Wiplala, who tells the family that he has been sent away by the other wiplalas because he cannot "pixilate" (do magic) well enough. Wiplala stays with the Blom family and does all kinds of magic tricks the children thoroughly enjoy. When the poor neighbour poet walks in to have dinner with the family, he sees Wiplala and tries to pick him up. The little wiplala is scared of him and thus pixilates him to stone, causing quite some worry for the family. But the real trouble starts when the family, including Wiplala hidden away in a bag, go out for dinner in town. Not expecting such high prices, mister Blom is not able to pay the bill and all of them get locked up in an office until the police arrives. Wiplala then shrinks the others to his own size so they can all escape together – the start of a true adventure. Hordes of people coming to look for them in their home, a flight on the back of a pigeon, a stay in the Royal Palace of Amsterdam, eating their bellies full in a delicatessen shop and ending up in the

hospital, where they finally seem to find someone, a doctor, who can help them. But when the happy ending is in sight, the bag they are hiding in gets stolen and thrown in the canal. Hidden in the big house of two elderly ladies, they manage to call the doctor for help. At last, Wiplala finds the special berries the family needs to eat to return to their human size. This is the end of their adventures. In the very end, Wiplala realises he can now pixilate well enough to go back to the other wiplalas, leaving the family behind with beautiful memories.

Annie M.G. Schmidt has not only has had – and still has – a great influence in the Netherlands and in the Dutch children's books industry, but has also travelled far across borders. Her books have been translated into at least fifty four languages (*Vertalingendatabase - Annie M.G. Schmidt* accessed 14-01-2023), from Vietnamese to Latin, from Gaelic to Persian. *Wiplala* has been translated into Afrikaans, Bulgarian, Catalan, Chinese, Czech, English, Estonian, Finnish, French, German, Greek, Hebrew, Hungarian, Icelandic, Italian, Japanese, Korean, Lithuanian, Russian, Spanish, Swedish, and Ukrainian. This makes the corpus easily accessible and expandable to other languages and language families.

4.1.2. *Uiplalà* – translated by Laura Pignatti

There is only one translation of *Wiplala* in Italian. The Italian public had to wait for some decennia, before they could discover Schmidt's *Uiplalà*: only in 1995 Arnoldo Mondadori Editore inserted it in its juvenile collection, in Laura Pignatti's translation. Pignatti is still active as translator from Dutch into Italian with over one hundred and sixty translated works (*Laura Pignatti* accessed 14-01-2023). No illustrations have been printed in the Italian version, which only partially explains the length of the book. Only 114 pages long (chapter one starts on page 3, the story ends on page 116), the Italian version seems significantly shorter than the Dutch original.

4.2. CREAMY: a platform for the analysis of multilingual phraseology

CREAMY (*Calvino REpertoire for the Analysis of Multilingual PhraseologY*) is a platform ideated by Paolo Bottoni and Sabine E. Koesters Gensini and built with the help of Filippo Mazzei. Bottoni et al. (2020) describe both the theoretical considerations at the basis of the platform and the technical construction of the platform itself. CREAMY is an instrument that gives researchers the opportunity to annotate the complexity of phraseological units in their co-text, while still being simple and intuitive enough to guarantee a user-friendly environment. One of the major advantages of the platform is the possibility to link phraseological units to their translatants in multiple languages; the translatants can be annotated using the same detailed parameters used for the starting text.

In §4.2.1. the functioning of the platform will be described, to shed light on the process of inserting and annotating phraseological units and their translatants. The description of new phraseological units and translatants, and hence all predefined parameters and the classification implemented in this dissertation, are discussed in §4.2.2. The last subparagraph (§4.2.3.) highlights the currently existing search and analysis options available on the platform.

4.2.1. The functioning of the platform

The platform is not accessible for external users at present[55]. Once you have entered your credentials, if applicable, you have to select the role you want to work in. Access to the platform is scaled: annotators, for example, can only work on the text(s) and in the language(s) they have been assigned to, while linguistic supervisors can also add or modify information about texts or language

[55] Access to the platform can be granted upon registration and authorization by the research manager, Sabine E. Koesters Gensini. The platform itself uses Italian as a metalanguage, i.e. all parts of the platform are in Italian. Furthermore, Italian is also dubbed *linguistichese*, the metalanguage used to connect all language specific categories to ensure mutual understanding.

specific subcategories in the assigned languages. Only the *supervisore umanistico* ('humanistic supervisor'), has complete access to the system and can add new linguistic supervisors. You then have to select the language in which you intend to work[56] from a drop-down menu that contains all the languages you can access. The role and language can be changed in the upper right side of the screen. On the left side of the screen you can find the main menu, divided in three groups: *analisi testo* ('text analysis'), *gestione testi* ('text management') and *impostazioni* ('settings'). The text analysis section hosts different search options, that will be further discussed in §4.2.3.. The text management section has two sub-options: texts and phraseological units.

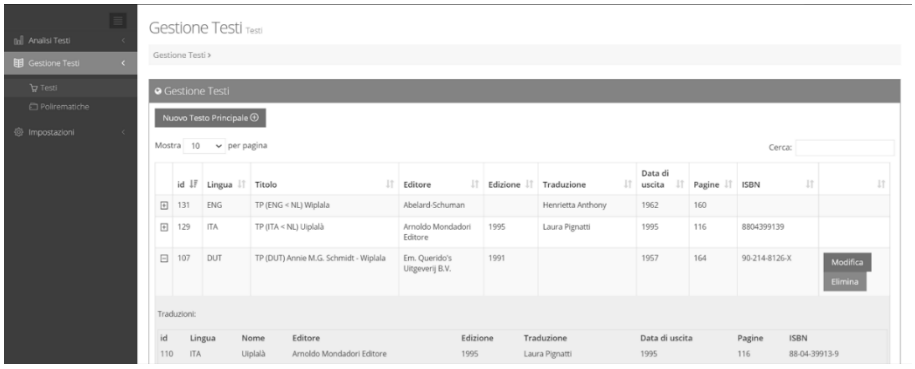

Figure 3 CREAMY: Text management – texts

The first brings you to an interface (see Figure 3) where all starting texts[57] are displayed, with a unique identifier and all relevant metadata (language, title, author, year of first edition, editor, year of edition used, total page number and ISBN of edition used). By clicking on the + symbol on the left side of the internal identification number, a list of linked translations (or rather: arrival texts) appears,

[56] You can only modify or add information for the language you are currently working in. If you choose to work in Italian, for example, you can only work on Italian starting texts or Italian translations.

[57] These texts are marked with "(TP)", *testo di partenza*, 'starting text'. Starting texts that are translations have a field to indicate the translator.

accompanied by the same information as for the starting text and the name(s) of the translator(s). It is required to add a text[58] in the system, before annotators can start working on it.

Figure 4 CREAMY: Text management – phraseological units

By choosing the option *polirematiche* ('phraseological units') in the text management section of the main menu, you reach to the section where you can insert and annotate PUs (see Figure 4). You first have to select the starting text you want to focus on from a drop-down menu (Figure 4: 1) – if it is in the language you are currently working in, another drop-down menu will appear from which you can choose the dictionary you want to work with[59]. Usually, only one reference dictionary is chosen per language, but by allowing the use of different lexicographic resources it is possible to evaluate the differences between their inclusion and presentation of phraseological units. If you are working in a different language (Figure 4: 2) than the one of the starting text you selected, you can choose the arrival text you want to analyse (Figure 4: 3)[60], and select the reference lexicographic resource (Figure 4: 4).

[58] First a starting text, then its arrival text(s) if a contrastive analysis is the scope of the research. It is also possible to work exclusively on phraseological units in one text, without confronting them with their translatants.

[59] In order to avoid too many people having access to it and possibly change data, a language supervisor has to add one or more annotator(s) to a specific text (starting or arrival), and add the reference dictionary they will be working with, before the annotator can start working on a text.

[60] This is a mandatory step, since a starting text can have multiple arrival texts (translations) in the same language.

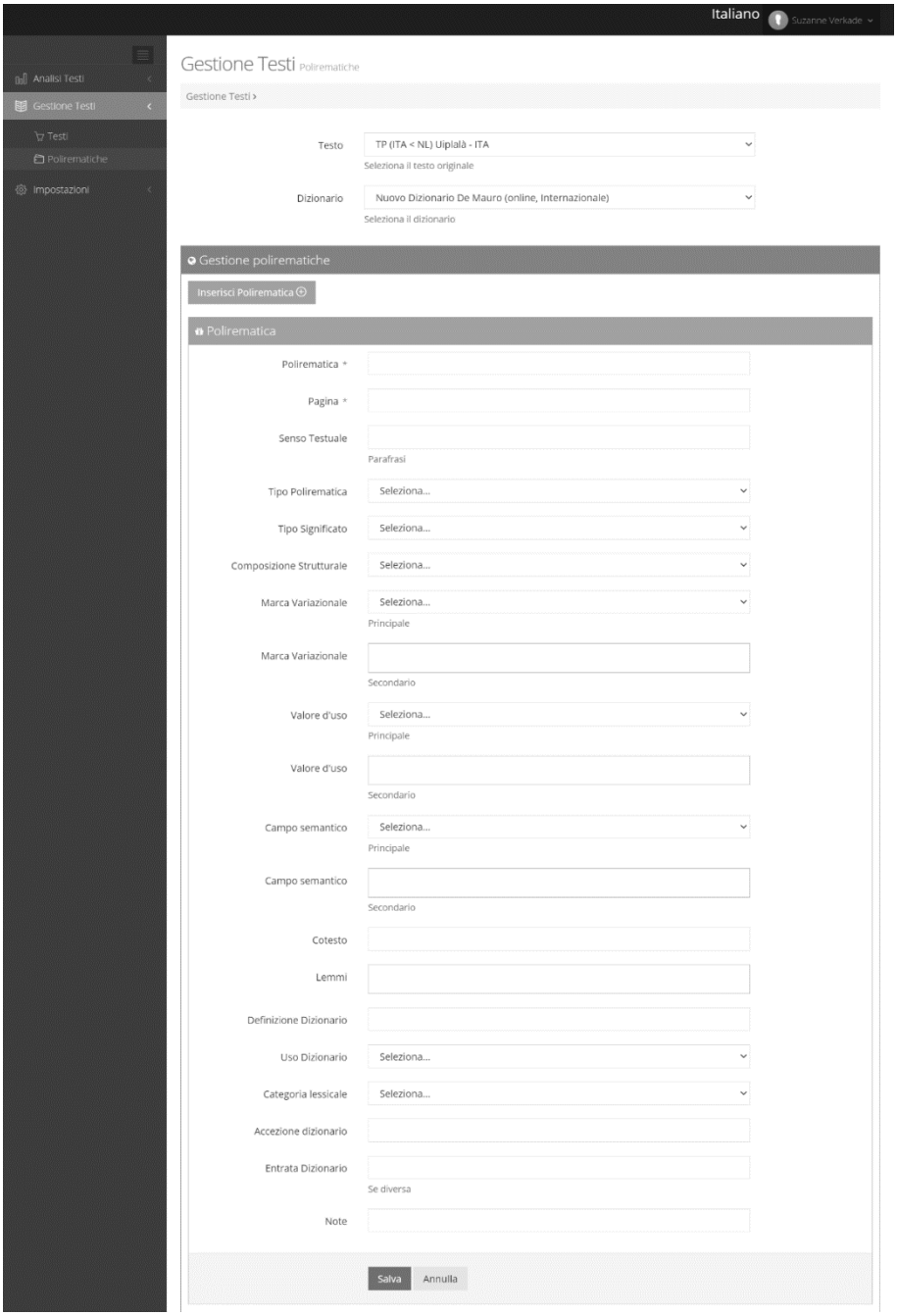

Figure 5 CREAMY: Text management – phraseological units – insert new phraseological unit

After selecting these options, the annotator can add new phraseological units or translatants. When working on the starting text, it is possible to add a new PU straight away; when working on an arrival text, it is necessary to first select the PU you want to add a translatant to. The page in Figure 5 shows the fields available for the annotation of a new phraseological unit[61].

Using this template consisting of twenty fields (either a text field or a drop-down menu) that refer to a group of parameters, you can thoroughly describe phraseological units and their translatants in a systematic way, thus quite precisely identifying their "value", i.e. the function of the PU inside the linguistic system it belongs to. These parameters are discussed in detail in the following paragraph.

4.2.2. Description of new phraseological units and translatants

The fields available for the systematic and detailed analysis of phraseological units and translatants in CREAMY are the following[62]:

a) The lemmatized form of the phraseological unit;
b) The page of the edition used in which the PU is present (every occurrence of a PU has a separate entry);
c) The co-text in which it occurs (a portion of text preceding and/or succeeding the PU, needed to determine its value in that specific pragmatic context);
d) The 'textual' meaning of the PU, i.e. a paraphrasis of the meaning of the PU in that precise co-text;

[61] When adding a translatant, all basic information of the source phraseological unit is displayed: the PU itself; the page number; the meaning in its specific co-text; the co-text. All fields available for the annotation of a PU are also available for the annotation of a TL, plus one extra field for determining the equivalence between the starting and arrival text.

[62] To avoid repetition that might cause confusion, the following list will only refer to phraseological units. The same fields apply to translatants.

e) The type of PU, i.e. the type of semantic relation between the single lexical constituents of the PU and the meaning the PU has, as a whole, in that specific co-text (§4.2.2.1.);
f) The type of meaning, i.e. the presence or absence of different kinds of figurative meaning in that specific co-text (§4.2.2.2.);
g) The structural composition of the PU, i.e. the mostly syntactic relation between the single constituents of the PU (§4.2.2.3.);
h) The lexical category of the PU, i.e. the part of speech it belongs to (§4.2.2.4.);
i) The position the PU occupies within the variational system of the language, i.e. indicator(s) of the sociolinguistic-variational value within its linguistic system (e.g. "slang", "bureaucratic") (§4.2.2.5.);
j) The use value(s), i.e. the connotation the PU has within the specific co-text (e.g. "ironic", "derisive") (§4.2.2.6.);
k) The semantic field(s) the PU belongs to (§4.2.2.7.);
l) The individual lemma(ta) that compose the PU (§4.2.2.8.);
m) The full description that the monolingual reference dictionary[63] offers of the PU (if there is no description present, this absence will be noted) (§4.2.2.8.);
n) The number and/or letter of the specific sense of the PU in that context among those present in the reference dictionary and reproduced in the full description (§4.2.2.8.);
o) The usage mark(s) attributed to the PU in the reference dictionary (e.g. "regionalism", "formal"; the field remains empty if no usage mark is present) (§4.2.2.8.);
p) The lemma under which the PU is described if different than the PU (e.g. *wind* for the PU *in de wind slaan*) (§4.2.2.8.);

[63] The dictionary chosen by an annotator working on that specific text in that language.

q) Notes from the annotator if necessary.

There is no default setting for any of these fields (e.g. "standard" for language variety): they need to be filled out singularly for each new phraseological unit or translatant. The parameters of language variety, use value and semantic field all have two fields that can be used to define them; it is obligatory to choose a subcategory from a drop-down menu in the first field, while the second one may be left blank but can hold multiple secondary subcategories to describe all nuances in more detail. Furthermore, a unique identifier is assigned to each PU and TL, in order to guarantee that multiple occurrences of the same PU in the same page, sometimes even within the same co-text, can still be kept apart. To ensure traceability, the platform also keeps track of the creator of each PU and TL, and of the person who last modified it.

Even though filling out some of these fields on the platform may seem quite straightforward, the first one, (a) (the lemmatized form of the PU), already poses some methodological problems. For instance, what is the correct lemmatized form of a light verb construction? In a text, multiple variants of a light verb construction can occur (e.g. "to take a photograph": *een foto nemen* A PHOTO TAKE, *mijn foto nemen* MY PHOTO TAKE, *foto's nemen* PHOTOS TAKE), but being variants of the same construction (*foto* PHOTO + *nemen* TAKE), it is important to insert them all in the same canonical form, so they can be found as a single entry. We have decided to insert the "emptiest" form possible, even if it does not correspond to the use in the specific language (in the case of the example given above: *foto nemen* PHOTO TAKE).

The parameters described up until this point refer to all languages, texts, PUs and TLs; the subcategories, however, are language-dependent and can show rather large differences based on the properties of the language they describe. Both parameters and subcategories aim to be distinct – and thus try to avoid redundancy – in order to give a transparent description and classification of PUs.

In the next section, single parameters will be discussed in more depth, and all relevant subcategories for each language will be presented[64].

4.2.2.1. Type of phraseological unit

The parameter "type of phraseological unit" takes into account semantic criteria. It distinguishes three types of semantic agglutination[65], for both the PUs in the starting texts and the TLs in the arrival texts. When the whole PU is non-compositional, i.e. the single constituents undergo a modification of their autonomous semantic value resulting in a PU's meaning that cannot be deduced from those constituents, it is classified as an "idiom" (e.g. *in de steek laten*; *piantare in asso*). When only one of the constituents is affected by a modification from a semantic point of view, the PU is classified as a "collocation" (e.g. *de hand drukken*; *stringere la mano*). PUs with no semantic agglutination have been classified as "other"[66] (e.g. *lawaai maken*; *fare rumore*). If we imagine placing these three types of PUs on a hypothetical continuum of semantic transparency, which becomes more and more transparent as we move from left to right, idioms would occupy the left-hand side and "other" PUs the right-hand side, with collocations somewhere in the middle (Figure 6).

---

[64] The following paragraphs (§4.2.2.1 – §4.2.2.9.) are partially based on general and language specific drafts (later modified, and newly modified for this dissertation) presented in Koesters Gensini & Berardini (2020). Especially useful for the following paragraphs has been the chapter written on Italian by Piattelli (2020).

[65] There are two more types of phraseological units, that, however, do not only take semantic criteria into account. This remains an issue to be resolved. One of these is "proverbs, sayings and aphorisms" and has only be used once in this dissertation, for an Italian saying (*gatta ci cova*). While in theory this saying could have been classified as an idiom from a semantic point of view, it was deemed best to keep paremiology and phraseology separate from the start. The second type of phraseological unit not included in these three types of semantic agglutination, are compounds (see below).

[66] These PUs are characterised by other kinds of agglutination or restrictions, mostly on a morphosyntactic level or because their constituents have a particular co-occurrence.

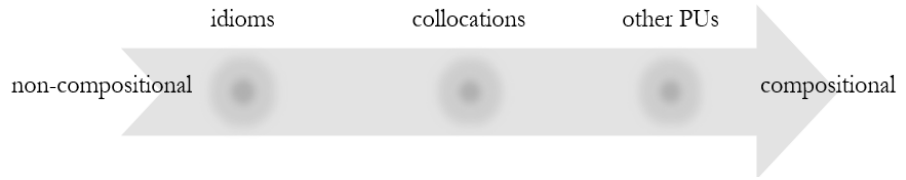

Figure 6 Continuum of semantic transparency

Another type of phraseological unit that, however, only partially relies on a semantic criterion, is compounds. Although we are aware of the fact that compounds are at present often not included in phraseological research (see §2.2.3 and specifically the polylexicality criterion), we have decided to focus on them for three reasons:

1) compounds are often translatants of a multiword expression[67] in another language (e.g. *battlefield* (or *slagveld* BATTLE-FIELD in Dutch) = *campo di battaglia* FIELD OF BATTLE in Italian), and as such they are challenging for language users;
2) orthographic rules tend to change, allowing locutions to become single graphic words (e.g. Dutch *dagen lang* DAYS LONG 'going on for days' became *dagenlang*) – therefore, orthography cannot be a criterion;
3) when not completely transparent and compositional, compounds pose a similar difficulty for language learners as multiword expressions do.

A single graphic word composed of two or more lexical morphemes is thus classified as a "compound". Compounds that have a literal overall meaning ("constituent 1 + constituent 2"), however, have not been taken into consideration. These compounds, although bound by their composition in a single graphic word, can in fact be seen as constructions similar to free combinations of words, opposed to compounds with a clear overall meaning

[67] Here the term "multiword expression" is purposely used and not "phraseological unit", because the first focusses on the composition in multiple words, while the latter is neutral and is used throughout this dissertation to include all types of phraseologisms, including compounds.

(even if the constituents are quite transparent). Compounds like *poppenstoeltje* DOLL-CHAIR-DIM 'little chair intended for dolls' have not been taken into consideration, while a relatively transparent compound like *schrijfmachine* WRITING-MACHINE 'typewriter' has entered our analysis – a 'machine intended for writing' is a typewriter, but the overall meaning is not literally *schrijf* + *machine*.

Compounds can be positioned on a continuum of semantic agglutination as well, just as multiword expressions, but we have decided not to include these possibilities directly as different types of phraseological units. Adding specific subcategories for compounds would have undermined the attempt at keeping the classification as simple as possible, without any (partially) overlapping subcategories. Including compounds directly as idioms, collocations or "other" PUs would not have done justice to the specific aspects of both multiword expressions and compounds. In a future stage of this research, an attempt could be made at classifying compounds more precisely from the semantic point of view following Libben et al. (2003).

The same types of PUs apply to translatants. However, not all TLs are phraseological units. For this reason, more subcategories are available for their classification, namely:

a) free combination of words;
b) monorematic word;
c) too freely translated to identify a direct translatant;
d) not translated.

The first two subcategories describe the two cases in which we do find a TL in the arrival text corresponding to a PU in the starting text, resulting in: a) a free combination of words, when the TL consists in multiple words (e.g. *te voorschijn komen* translated by *abbandonare il nascondiglio*) and b) a monorematic word, when it is a single graphic word with one lexical morpheme (either a simple or a complex word, but not a compound word, which has at least two lexical morphemes) (e.g. *foto nemen* translated by *fotografare*). The last two subcategories

describe the cases in which we cannot identify a TL (marked as "--" in CREAMY in the field where the lemmatised form of the TL would normally be annotated): either because there is no clear translatant for the original PU (c) or because the PU is not translated at all (d).

4.2.2.2. Type of meaning

The parameter "type of meaning" refers to the figurativeness (or lack thereof) of a PU. In an attempt to describe PUs as accurately as possible, a total of five subcategories are applied:

a) Generically figurative (e.g. *in een oogwenk*; *in un batter d'occhio*);
b) Metaphorically figurative (e.g. *broodmager*; *magro come un chiodo*);
c) Metonymically figurative (e.g. *naar bed gaan*; *andare a letto*);
d) Nor figurative, nor compositional (overall agglutinated; e.g. *pindakaas*; *burro d'arachidi*);
e) Not-figurative and compositional (e.g. *boodschappen doen*; *fare la spesa*).

The first three subcategories describe cases in which the PU in question has a figurative meaning in its specific co-text, in the last two subcategories no figurativeness is present. It is also possible that a PU has an agglutinated (i.e. non-compositional) meaning, while not having any figurative meaning: in that case it is classified as having a "nor figurative, nor compositional" meaning.

4.2.2.3. Structural composition

The parameter of structural composition aims to classify the PUs from a syntactic point of view. The classification of the internal structure of PUs is a very complex task, as it entails organising the PUs in distinct, non-redundant categories, based on purely lexical-syntactic criteria, while still being exhaustive (Koesters Gensini 2020a: 332). It is evident why idioms and collocations are considered PUs – based on their semantic agglutination – whereas this is not the case for PUs classified in the subcategory "other" of the parameter "type of PU".

Especially this group of PUs benefits from a classification based on their structural composition.

In the following, we will first introduce the subcategories used for PUs that are shared by both languages included in this research. Next, the subcategories added for the analysis of TLs will be presented.

Structural compositions of phraseological units

The subcategories shared by Dutch and Italian are the following:

a) co-occurrence of lexical morphemes (CLM; e.g. *proef afleggen*; *sostenere una prova*);
b) irreversible binomial (IB; e.g. *heen en weer*; *avanti e indietro*);
c) light verb construction (LVC; e.g. *herrie maken*; *fare confusione*);
d) verb-particle construction (VPC; e.g. *correre via*);
e) expression with one or more prepositions (EP; e.g. *in plaats van*; *al posto di*);
f) compound (e.g. *wegrennen*, *pijlsnel*; *francobollo*);
g) simile (e.g. *zo bang als een muis*; *rosso come un gambero*);
h) other (e.g. *'s nachts*; *zitto zitto*).

According to Piattelli's (2020a: 142) excellent definition, phraseological units that are characterised by a recurring (but not mandatory) association of its constituents, are classified as a co-occurrence of lexical morphemes (a). Both PUs with a semantic modification (idioms and collocations) and semantically transparent expressions are included ("other" PUs). Piattelli (2020a: 142–143) then goes on to illustrate the fine line between syntactic and semantic criteria in this subcategory. A PU can be classified as a CLM, if at least one of its constituents can be substituted, regardless of the presence or absence of a semantic surplus. As a result

> [e]spressioni semanticamente non marcate come "fronte corrugata", "momento passeggero", "pericolo scampato", "pioggia

> scrosciante" e altre sono state considerate come co-occorrenze di morfi lessicali dal momento che – pur con risultati discutibili a livello stilistico – è teoricamente consentito dire "fronte raggrinzata", "momento fuggevole", "pericolo evitato", "pioggia fiottante", senza che la modifica incida significativamente sul significato dell'espressione. Anche nei casi di espressioni idiomatiche, una polirematica come "reggere il cuore" si configura a livello sintattico come co-occorrenza di morfi lessicali in quanto il verbo "reggere" potrebbe essere sostituito da un altro mantenendo il medesimo significato (es. "tenere il cuore"). Al contrario, casi come "se stesso", "poco di buono", "farsi largo", ecc. non sono stati considerati come co-occorrenze in quanto l'associazione tra i lessemi si configura come una vera e propria agglutinazione, in cui eventuali prove di commutazione porterebbero alla perdita del significato dell'espressione.[68]

Irreversible binomials (b) according to Malkiel (1959) are constructions consisting of two lexemes, belonging to the same lexical category and joined by a conjunction, in a fixed conventional order, e.g. "salt and pepper", "bed and breakfast", "cut and paste", "now and then", "double or quits/nothing", "good or bad", "make or break", "sink or swim".

---

[68] "Semantically unmarked expressions like "fronte corrugata", "momento passeggero", "pericolo scampato", "pioggia scrosciante" and others have been considered co-occurrences of lexical morphemes since – even if with questionable results on a stylistic level – it is theoretically allowed to say "fronte raggrinzata", "momento fuggevole", "pericolo evitato", "pioggia fiottante", without that modification significantly impacting on the meaning of the expression. Also where idioms are concerned, a phraseological unit like "reggere il cuore" is considered a co-occurrence of lexical morphemes from a syntactic point of view, as the verb "reggere" could be substituted by another while maintaining the same meaning (e.g. "tenere il cuore"). On the contrary, cases like "se stesso", "poco di buono", "farsi largo", etc. have not been considered co-occurrences of lexical morphemes because the association between the lexemes is truly agglutinated, and any commutation test would lead to the loss of the meaning of the expression." Cf. for instance Dutch *proef afleggen* (*proef ondergaan*), Italian *sostenere una prova* (*fare una prova*).

Light verb constructions (c) are intended as expressions consisting in an NP and a light verb (Jespersen 1942: 117–118) that has a supporting function. The whole construction can often be reformulated with a simple verb (e.g. *to make a call* > *to call*, but not *to make an appointment* > **to appointment*) (Ježek 2011: 198; Bonial 2014: 181), although this does not imply that they are completely interchangeable from a semantic and pragmatic point of view (see Wierzbicka 1982 for a discussion). Literature on the subject varies not only within a specific linguistic tradition, but also between different languages. For an overview of the treatment of LVCs in different linguistic traditions including Dutch and Italian, see Koesters Gensini et al. (2022). Extremely useful from this point of view are the very detailed PARSEME (PARSing and Multi-word Expressions) annotation guidelines for light verb constructions (2018; 2020, also cf. Cordeiro & Candito 2019; Ramisch et al. 2018, 2020). Everaert & Hollebrandse (1995: 95–100) will be followed for Dutch; Ježek (2011: 195–198) for Italian.

Dutch and Italian both have peculiar verbal expressions. These could be generally classified in the structural composition "verb-particle construction" (d), especially when other languages are involved[69]. Nevertheless, in this contrastive analysis of Dutch and Italian a different approach has been chosen.

In the case of Dutch, we have separable complex verbs (SCVs): "combinations of a verb and another word that function as lexical units" (Booij 2019: 223). That word can either be a noun (e.g. *pianospelen* PIANO-PLAY 'to play the piano'), an adposition (e.g. *opbellen* ON/AT-CALL 'to call'), an adjective (*schoonmaken* CLEAN-MAKE 'to clean'), an adverb (*neerstorten* DOWN-COLLAPSE transitive 'to dump', intransitive 'to crash'), or a word that occurs only when combined with a verb (e.g. *teleurstellen* 'to let down') (Booij 1998: 6). When first working on the classification of SCVs in CREAMY in a previous research

[69] For instance, English has verb-particle constructions that are often referred to as particle verbs or, more specifically, as phrasal verbs or prepositional verbs. As in Italian, they can either consist in verb + preposition (e.g. *to pick up*) or verb + adverb (e.g. *to come back*), but also in verb + adverb + preposition (*to put up with*).

project, it seemed most fitting to analyse them as "other" phraseological units and to divide them into three different kinds of structural compositions: transparent, semi-transparent and opaque separable complex verbs. This practice later proved incorrect for two reasons:

1) it introduced a semantic criterion in a parameter only meant for structural, syntactic classification[70];
2) as SCVs are (separable) compounds, this resulted in an overlap with the structural composition "compound".

Therefore it is preferable to classify SCVs as compounds, that can be easily filtered out thanks to the lexical category "separable complex verb" (thus keeping the opposition with non-separable verb compounds). This solution is also not fully satisfactory because of the general similarity of SCVs with Italian verb-particle constructions. While this similarity cannot be overlooked, the empirical analyses presented in Chapters 5, 6 and 7 show that Dutch SCVs and Italian VPCs are not frequently translatants of each other – less than expected, in fact. SCVs are thus classified among compounds because of their peculiar form, but can be seen as an intermediate category between multiword units and compounds.

Italian verb-particle constructions are usually referred to as *verbi sintagmatici* 'syntagmatic verbs' and comprise verb + preposition constructions (e.g. *tirare su* PULL UP 'to pull up' or idiomatic 'to raise [children]') and verb + adverb constructions (e.g. *buttare fuori* THROW OUT). VPCs in Italian can have both compositional and non-compositional, idiomatic meanings, and can also be used figuratively. In this dissertation both "verb-particle construction" and "syntagmatic verb" will be used to refer to the same Italian phenomenon.

The subcategory "expression with one or more prepositions" (e) has been used to classify expressions characterised by the presence of one or more

[70] The semantic transparency of SCVs is analysed in "type of meaning" (as for all phraseological units).

lexical morphemes and a specific preposition that expresses a certain syntactic relationship (often space- or time-related) that could not be expressed in absence of that preposition or by substituting it with another, thus revealing some degree of agglutination (Piattelli 2020: 143).

As already stated with regard to the types of phraseological unit, single graphic words composed of two or more lexical morphemes are classified as a "compound" (f). This means that all phraseological units that have been classified as a compound in "type of phraseological unit", will also be classified as a compound in "Structural composition". In future research, an attempt could be made to further investigate the different internal structures of non-compositional compounds, and what we can learn from them.

Similes (g) have been added as a structural composition when a first annotation of the phraseological units had already been completed. They are a peculiar aspect of *Wiplala* and had not yet occurred – or at least not frequently enough – in other research carried out on the CREAMY platform, hence the lack of the category.

In the last subcategory, "other" (h), we find all PUs that do not fit into another structural composition. It is clear from the mere existence of this subcategory, that the structural classification of phraseological units remains extremely complex, and that this attempt at classification is far from satisfactory.

Equally clear is the fact that the structural compositions illustrated up to this point are not exclusive subcategories. Especially the co-occurrence of lexical morphemes overlaps with other categories, as most PUs (compounds included) are also, to a certain extent, co-occurrences of lexical morphemes. It seems necessary, though, to maintain these subcategories, although not satisfactory, in order to distinguish as much as possible between certain internal structures. Needless to say, the parameter of structural composition and especially the subcategories "co-occurrence of lexical morphemes" and "other" will be subject to further research.

Structural compositions for non-phraseological translatants

Besides the structural compositions presented above used to classify PUs and phraseological TLs, more subcategories are needed to classify non-phraseological TLs:

h) free combination of words;

i) monorematic word.

If a TL is classified as a free combination of words in the category "type of PU", it will then automatically have to be classified as a free combination of words in the parameter "structural composition" [71]. The same applies to monorematic words[72]. Phraseological units that do not have a translatant, i.e. they are either too freely translated to identify a precise translatant or they have not been translated at all, are obviously not assigned a specific structural composition, as they do not offer material for analysis. Except for "--" in the translatant field to mark its absence, the page number, the co-text, and the "type of phraseological unit", all other fields in these cases are empty.

4.2.2.4. Lexical category

Another parameter for the description of PUs in CREAMY is that of the lexical category, which refers to the function of the entire phraseological unit (not the part of speech of its single constituents). Especially in analysis this is a very useful to be able to filter out specific phraseological units, for example only those that function as an adverbial phrase. The lexical categories are:

---

[71] A free combination of words does not exclude semantic solidarity between its constituents. The fact that a lexeme is combined more often with some lexemes than with others, does not necessarily make such a combination a phraseological unit in general, or more specifically a co-occurrence of lexemes. In this case as well, as we have seen in many aspects of phraseological units, there is a continuum.

[72] A complex issue regards reflexive verbs. In an effort to keep the number of the subcategories of the structural composition to a minimum, the decision has been made to classify reflexive verbs that might need to be analysed as translatants of a phraseological unit (e.g. *in orde komen* translated into Italian with *risolversi*) among monorematic words. In future research we would recommend adding a separate subcategory for reflexive verbs.

a) adjective
b) adjectival phrase
c) adverb
d) adverbial phrase
e) conjunction
f) conjunctional phrase
g) noun
h) noun phrase
i) preposition
j) prepositional phrase
k) pronoun
l) pronominal phrase
m) verb
n) verb phrase
o) separable complex verb
p) formula
q) other

There are separate categories for single graphic words (either compounds or monorematic words) and phrases (multiword expressions and free combinations of words). There is a separate category for formulae (e.g. *dames en heren*, *tot ziens*; *signore e signori*, *a presto*) and an "other" category for very rare PUs, but mostly for TLs that do not fit in any other category[73].

4.2.2.5. Language variety

The aim of the "language variety" parameter is to describe the PU's position in its language-specific sociolinguistic-variational system. When this sociolinguistic positioning deviates from the standard, it often represents a distinct characteristic for the PU in question and its broad co-text. Therefore, it is vital for a good translation to maintain an equivalent position in the sociolinguistic system of the target language.

The first variational labels that have been identified for research on the CREAMY platform are those of Italian, based on the model of the diasystem of Italian varieties elaborated by Berruto (1987/2012: 24). The diaphasic continuum is illustrated with a diagonal axis going from the lower right (highly informal) to

[73] The "other" subcategory is mostly needed in cases where there has been quite a modification between starting and arrival text. E.g. the Italian *ora come ora* has *als hij dit beleefd had* as a Dutch translatant.

the upper left (highly formal), the diastratic continuum with a vertical axis from bottom (lower social class) to top (higher social class), and the diamesic continuum with a horizontal axis from left (written) to right (spoken). In this model, Italian standard language (literary standard and neo-standard) is positioned slightly off-centre, and stretched upwards in order to occupy a position closer to the high end of the diaphasic continuum and of the diastratic continuum, and more to the left of the diamesic axis (more written than spoken). The Italian varieties available for the classification of PUs and TLs on CREAMY are:

a) standard;
b) substandard;
c) colloquial (more spoken, informal);
d) highly informal;
e) 'popular' (diastratically and diafasically low variety);
f) regional;
g) 'popular' regional (diatopically marked, diastratically and diafasically low)
h) spoken;
i) formal;
j) highly formal;
k) slang;
l) technical-specialist language, jargon;
m) archaic;
n) obsolete;
o) idiolectal;
p) dialectal;
q) bureaucratic;
r) other.

Most of these have not been used in the annotation of *Uiplalà*. Only standard, colloquial, spoken and very rarely technical-specialist language, formal, and "other" characterise the Italian corpus.

Extant studies on sociolinguistic variation in Dutch to our knowledge do not give a comprehensive overview or model of the Dutch sociolinguistic-variational system, and tend to focus on a specific variety (e.g. Smakman 2006) or on the opposition between registers or varieties (e.g. Impe et al. 2009), often focussing on the differences and similarities between Belgian and Netherlandic Dutch (e.g. Tummers et al. 2011; Van de Velde et al. 1997; van Halteren & Oostdijk 2018). The usage labels in dictionaries did not prove very useful as they often lack consistent and exhaustive application. Stachurska (2018) discusses the issues of codifying usage in lexicographic resources by the use of labels, and highlights some of the many diverging classificatory schemes that have been proposed. She then analyses the usage labels in five lexicographic resources for English as a foreign language, shedding light on the divergences of the labelling and the problems this causes. Janssen et al. (2003) also discuss the codification of usage labels, but do so with the help of Dutch examples.

Given the lack of a steady theoretical basis for the implementation of a specific Dutch variational classification, a possible solution is to apply Berruto's model for Italian to Dutch as well. The same variety labels as previously listed for Italian ((a)-(r)), have thus been used for the description of Dutch PUs and TLs. Based on the reference dictionary for Dutch, Van Dale, it has been decided to add the following labels:

s) Dutch Dutch (typical for language used in the Netherlands);
t) Belgian Dutch (typical for language used in the Flanders);
u) Literary Dutch.

Neither of these added labels has been applied in the annotation of *Wiplala*.

Language varieties, too, are a continuum. After selecting one main variety, other, secondary varieties can be added to fully describe the PU, in order to attain a more complete annotation. There is no default language variety: for every new PU and TL at least one main variety needs to be selected.

4.2.2.6. Use value

An important characteristic of phraseological units is their semantic surplus (cf. Gréciano 1994), that is to say the connotative nuances that the parameter "use value" aims to (partially) describe. The subcategories refer to the way a PU is used in the co-text or the effect it has on the receiver(s)[74]:

a) derisive;
b) derogatory;
c) flattering;
d) hyperbolic;
e) interjectional;
f) ironic;
g) jokingly;
h) neutral;
i) pejorative;
j) sarcastic;
k) sentimental.

The use value thus tries to capture the connotation of the PU or TL in the co-text and broader context. For this parameter as well, CREAMY provides two fields: one for the main use value and one for any secondary use value(s). While a neutral use value is by far the most common subcategory, it is not set as a default.

4.2.2.7. Semantic field

The parameter "semantic field" is designed to classify the phraseological units and translatants in macro-subjects. In its current conception within the CREAMY project, it is rather problematic. The semantic fields identified and available in CREAMY up until this point are the following:

[74] Although the names of these use values give a rather clear indication of the situations they ought to describe, the implementation of these subcategories does depend on the subjective choices of the annotator.

a) adolescence
b) agriculture
c) animals
d) body parts
e) causal relation
f) celestial bodies
g) childhood
h) clothing
i) cognition
j) communication
k) danger
l) death
m) family and relatives
n) fantasy
o) feelings and emotions
p) five senses: hearing
q) five senses: sight
r) five senses: smell
s) five senses: taste
t) five senses: touch
u) food
v) four elements: air
w) four elements: earth
x) four elements: fire
y) four elements: water
z) generic
aa) human activity
bb) human behaviour
cc) human character
dd) illness
ee) jobs
ff) materials – objects
gg) modality of action
hh) modality of event
ii) money
jj) mood
kk) movement
ll) music
mm) nature
nn) negativity/worsening
oo) old age
pp) other
qq) physical action
rr) physical appearance
ss) plant kingdom
tt) politics
uu) positivity/improvement
vv) private life
ww) reflectiveness
xx) religion
yy) social relations
zz) spare time
aaa) spatial relation
bbb) temporal relation
ccc) war
ddd) weather

The semantic fields in CREAMY are not a closed category; new semantic fields can be added by the humanistic supervisor (for any language) or by linguistic supervisors (for specific languages). One main field needs to be selected for each annotated PU or TL, but one or more secondary fields can be added as well. For instance, "to earn one's bread" would be classified as a "human activity", "every now and again" as a "temporal relation", while "on horseback" would be considered a "modality of action", but is also related to "animals".

Some problems arise in the implementation of these fields: they are non-exhaustive and partially overlap. Furthermore, there are no clear annotation guidelines as of yet for the single semantic fields. This leads to open interpretations of the subcategories, due to the lack of limitations on the subjective choices of individual annotators.

The UCREL[75] Semantic Analysis System seems to be very promising: it has been implemented into various research projects and covers multiple languages, including Dutch and Italian (Piao et al. 2015, 2016). This framework, built for automatic semantic tagging of texts, is divided into twenty-one major discourse fields and further subdivided into 232 category labels (Archer et al. 2002; *UCREL Semantic Analysis System (USAS)* accessed 15-01-2023). Implementing a system with a totally different structure, however, would require a preliminary study on the differences and similarities between USAS and the semantic fields in CREAMY. At this moment, the latter guarantee a certain amount of comparability with the other studies conducted on the CREAMY platform, which is why in this dissertation we will continue to use them. Time constrictions and the scope of this project do not allow us to evaluate the possibility of implementing USAS on the whole platform, especially considering that 1) not every language analysed on CREAMY has a specific tagger in USAS and 2) it would need to be implemented not only in future analyses, but also in

[75] UCREL (University Centre for Computer Corpus Research on Language) is a research centre of Lancaster University.

those already present on the platform. Nevertheless, it is necessary to take into account this issue in future research.

4.2.2.8. Lemmata, reference dictionaries and senses

For each language, text, and annotator a monolingual reference dictionary has to be selected. The reference dictionary for Dutch is the *Dikke van Dale Online* (n.d.). A more suitable dictionary for Dutch, with a fuller description of PUs, is not available at the moment. In the future[76], *Woordcombinaties* (accessed 15-01-2023) will be able to fill a crucial gap for phraseological studies in Dutch lexicography, as it provides both a collocation and idiom dictionary, and a pattern dictionary (Colman & Tiberius 2018).

The reference dictionary for Italian is *Il Nuovo De Mauro* (n.d.-a), the online and abridged version of the *Grande Dizionario Italiano dell'Uso* (GRADIT; De Mauro 1999-2007), that has a special section for PUs.

When annotating PUs or TLs, five fields are devoted to lemmata and the description of the PU in the reference dictionary (if present at all; see §4.2.2. list items (l) to (p)). The first step is to fill out the individual lemmata that compose the PU or TL; this makes it possible to search for PUs that contain a specific lemma, e.g. "hand". After that, one has to add the full description of the PU in the chosen reference dictionary (if there is no description present, the absence of it has to be noted). A third field will be filled in with the number and/or letter in the reference dictionary referring to the specific sense of the PU in that specific co-text, in order to guarantee findability. The last two fields are optional: a usage label, if attributed in the dictionary, is included in the annotation, as well as the lemma under which the PU is described, if it's different from the PU itself.

[76] At this moment, the present lemmata are not enough to be able to use *Woordcombinaties* for research purposes.

4.2.2.9. Translational equivalence

After a thorough discussion in §2.3., we have concluded that it is important to measure the translation equivalence between two texts, confronting every PU with its TL. The translation equivalence will be measured on two levels (semantic and formal) and in four grades (absent, low, high and total), resulting in sixteen subcategories. Equivalence can hence be:

a) formally and semantically absent;
b) formally absent, semantically low;
c) formally absent, semantically high;
d) formally absent, semantically total;
e) formally low, semantically absent;
f) formally and semantically low;
g) formally low, semantically high;
h) formally low, semantically total;
i) formally high, semantically absent;
j) formally high, semantically low;
k) formally and semantically high;
l) formally high, semantically total;
m) formally total, semantically absent;
n) formally total, semantically low;
o) formally total, semantically high;
p) formally and semantically total.

4.2.3. Search and analysis options

In the *Analisi testo* 'text analysis' section of CREAMY, there are multiple options available for searching specific characteristics or for the statistical analysis of a specific text, PU or characteristic:

a) *Ricerca per polirematica* 'Search per phraseological unit';
b) *Ricerca per traducente* 'Search per translatant';

c) *Ricerca per proprietà* 'Search per property';

d) *Statistiche occorrenza* 'Statistics per occurrence';

e) *Statistiche per proprietà* 'Statistics per property';

f) *Statistiche testo* 'Statistics per text'.

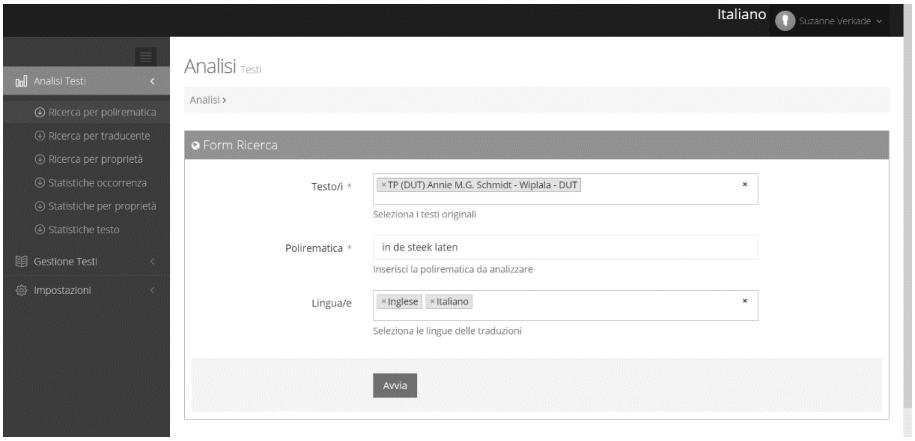

Figure 7 Search per phraseological unit

Search option (a) (Figure 7) allows the user to single out a specific phraseological unit in selected texts, accompanied by its translatants in selected languages.

After selecting the right text(s), and inserting the queried PU and target languages, all occurrences of that specific PU in the selected text(s) will appear in the bottom part of the screen, including all annotated information. Below those, a section per language shows how all those occurrences of the PU have been translated, with all annotated information.

In search option (b) you can look for specific translatants. If, for example, it is relevant to know which PUs (in general, that is to say in all annotated texts present on the platform up until that moment) are translated with "in de steek laten", CREAMY gives the result as shown in Figure 8. This makes it possible to do an inverted search, and to analyse how a specific target language (or rather, 'arrival language', when doing bidirectional analyses in which the

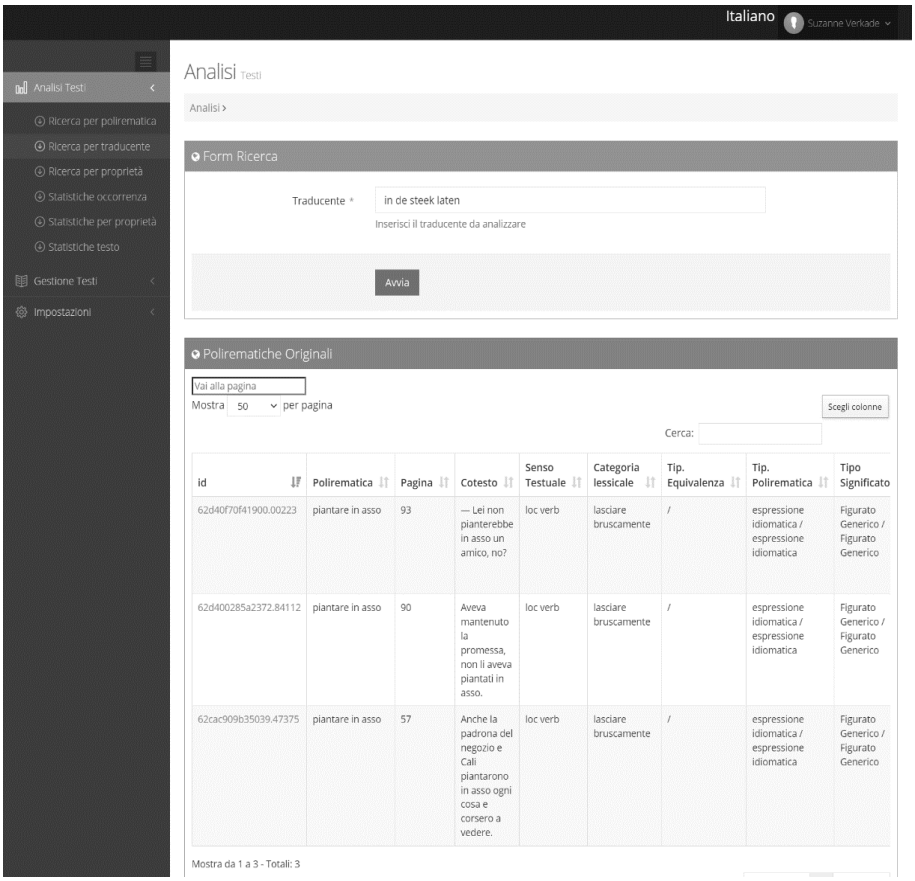

Figure 8 Search per translatant

original text becomes the 'arrival text' of the translation) conveys multiple PUs – and thus multiple denotative and connotative meanings – with the same translatant, and if so, with which acceptations of that translatant.

Besides searching for a specific phraseological unit or translatant, CREAMY allows users to search for specific properties or characteristics (i.e. per parameter). For example, it is possible to filter out all idioms. But it is also possible to add search restrictions to multiple parameters, to single out e.g. all idioms that have a metaphoric meaning, and convey an ironic use value, and belong to an informal register, and are allocated within the semantic field

"feelings and emotions". Unfortunately, it is not possible at the moment to do a cross-search of both phraseological units and their respective translatants, which could result, for example, in an overview of all figurative collocations in the starting text(s) that also have a figurative collocation as translatant in the arrival text(s).

The other three options provide a statistic overview. When looking for statistics per occurrence (d), CREAMY allows you to insert one phraseological unit (in any language) leading to a general, numeric outlook on how many occurrences that PU has in any text annotated on the platform that has at least one occurrence. Option (e) (Figure 9) provides a statistical analysis per property in all annotated texts: by selecting one parameter among type of phraseological unit, type of meaning, structural composition, lexical category, language variety, use value, semantic field and translational equivalence, and then a specific subcategory in the chosen parameter, CREAMY users are provided with an

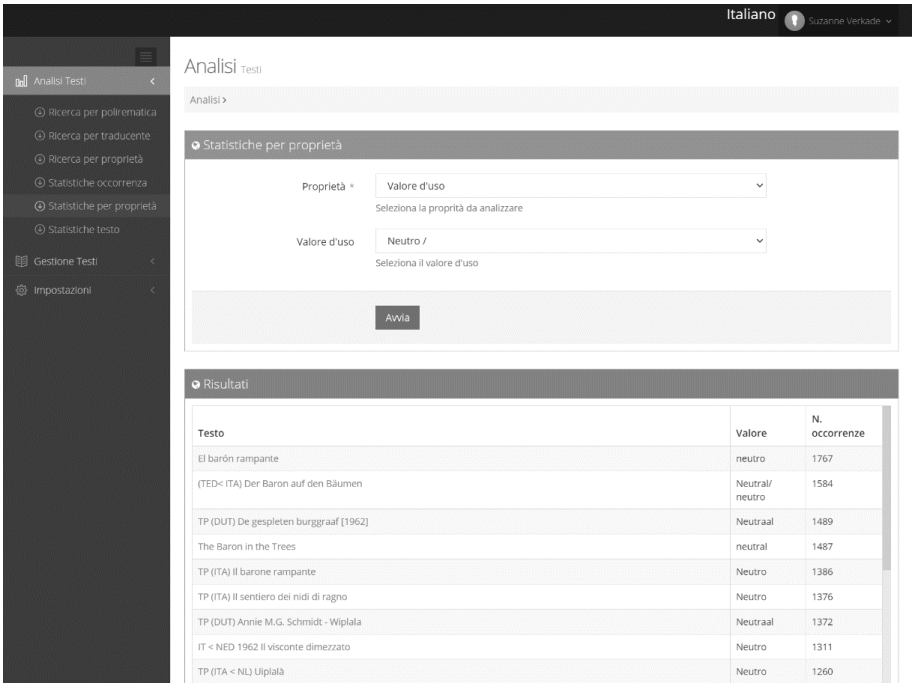

Figure 9 Statistics per property

overview of all annotated texts with at least one occurrence of the selected subcategory, ranked from the highest to the lowest number of occurrences.

The last option, statistics per text (f), provides a graphic overview of a specific starting text and all its annotated arrival texts, focusing on the following parameters: type of phraseological unit, lexical category, language variety, use value, and semantic field.

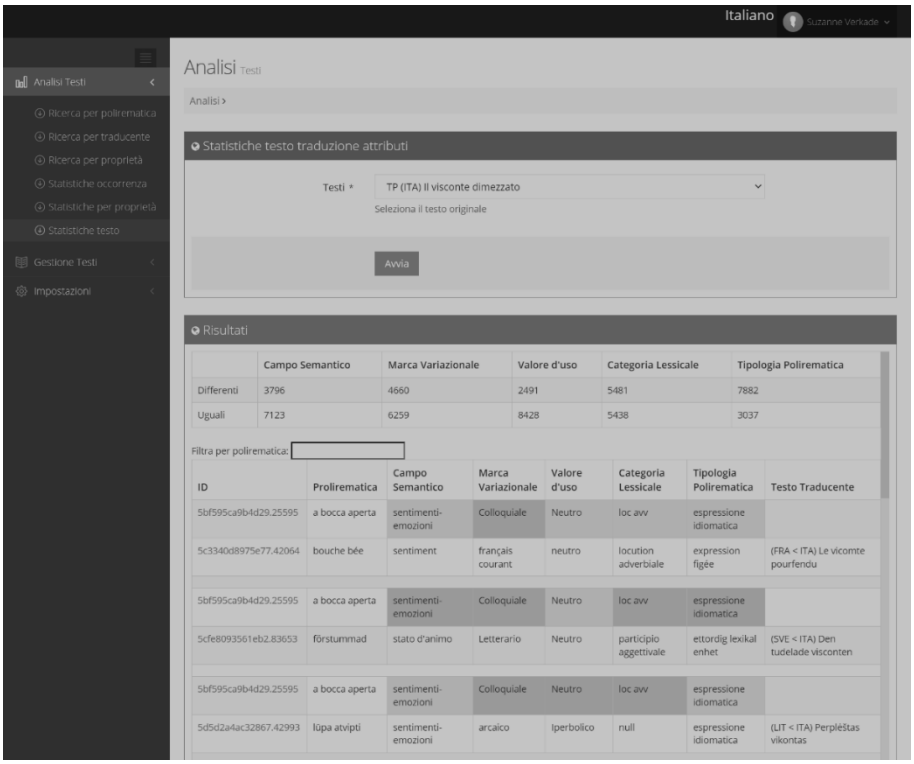

Figure 10 Statistics per text

In the upper part of the screen (Figure 10) a table, divided per parameters, gives an insight on how many translatants share the same subcategory[77] with their source phraseological unit, and how many differ. Below,

[77] Obviously, subcategories can differ between languages. To guarantee mutual understanding, they have been connected through a metalanguage, "linguistichese".

all PUs of the selected starting text are shown, with their respective translatant marked in different colours to show which parameters they share and which differ. The results shown in Figure 10 are those of the Italian starting text *Il visconte dimezzato* that currently has fourteen annotated translations in thirteen different languages. If you are interested in as single, specific PU, it is possible to filter the results.

4.3. Other research instruments

A very useful tool for this research – especially in the annotation phase of this project and directly after, in order to double check if each occurrence of all phraseological units had been inserted – has been the aligned Dutch and Italian texts. To do so, we have first extracted the text of the .epub and .pdf files of the original Dutch text and the Italian translation. Next, we have divided the text in separate files per chapter and cleaned it of the numerous errors caused by the OCR (optical character recognition). We then converted the files to a .txt UTF-8 format and aligned each chapter by using LF Aligner[78]. The alignment is formatted in .tmx (Translation Memory eXchange) files, and has been uploaded in this form to SketchEngine, which provides numerous ways to interrogate the corpus.

CREAMY provides easy linking between starting and arrival texts, and hence phraseological units and their translatants in multiple languages and/or in multiple translations in the same language, but it does not yet provide all the search and analysis options needed for complex analyses such as the present. For that reason, all data has been copied to Excel (one file for each direction: NL→IT and IT→NL), but the linkage between each pair of phraseological unit and

[78] We made a first attempt with MemoQ – which from certain points of view is definitely more user friendly than LF Aligner. However, as this project at an earlier phase also included English, we could not continue to use MemoQ because it does not allow more than two languages to be aligned contemporarily.

translatant needed to be restored. To reconnect PUs and their TLs, the same numerical identifier was added to both rows containing all respective data of the PU and TL. In this way, it is possible to use a PivotTable to cross-search both phraseological units in the starting text and translatants in the arrival text. For example, it is now possible to filter out only those collocations with a figurative meaning in the starting text that have a metaphorical idiom as a translatant.

WIPLALA: DATA ANALYSIS AND INTERPRETATION

The empirical part of this dissertation is divided into three chapters, in which various aspects of the data will be highlighted. Following the detailed annotation of all phraseological units and translatants in the Dutch text and in the Italian text, and using both the search and analysis options on the CREAMY platform, as well as Excel for more complex cross-searches, a quantitative analysis has been carried out.

In Chapter 5 the quantitative analysis of the Dutch phraseological units in *Wiplala* and their Italian translatants will be presented, accompanied by a qualitative discussion of examples[79]. Chapter 6 regards the Italian translation, here assumed as the starting text, and hence the Dutch original text as the arrival text[80]. In Chapter 7 the results of the first two analyses will be confronted in a bidirectional analysis, highlighting the most important differences and what those entail.

[79] The examples are visually separated from the main text. For each example the Dutch and Italian co-texts are given, in which the phraseological unit and translatant are underlined. The description of the examples is divided into two parts by a dash (–); the first part refers to the PUs and the second part to the TLs, unless otherwise stated. In the main text, Dutch and Italian phraseological units are given in cursive. Parts of phraseological units or non-phraseological translatants are placed between double quotation marks. Single quotation marks are used for the meaning of Dutch and Italian expressions.

[80] See the introduction to Chapter 4 on the choice to adopt "starting text" and "arrival text" throughout this dissertation, in stead of "source text" and "target text".