Ca' Foscari
University
of Venice

Master's Degree programme

in Language Sciences

DM 270/2004

Final Thesis

Metonymy Resolution with BERT and RoBERTa: what language models can infer about the interpretation of metonymy

Supervisor
Prof. Lebani Gianluca

Assistant supervisor
Prof. Giusti Giuliana
Dr. Dall'Igna Francesca

Graduand
Eleonora Ganio Mego
Matriculation Number 887189

Academic Year
2021 / 2022

To my grandmother Marcella

Abstract

Metonymy is a figure of speech that allows the use of a concept to refer to another concept closely related to the previous. It is used by the speakers to facilitate but at the same time better express the meaning as they intend it. The aim of this thesis is to test the performance of some pre-trained language models, such as BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019), on metonymy resolution. Metonymy resolution is a task that aim at finding the correct "hidden" referent behind a metonymic expression and consists of two parts, namely the recognition and the interpretation of such type of expression (Markert & Nissim, 2007). Since previous research dealt with metonymy resolution as a classification task, the main contribution of this thesis is investigating what these models can infer about metonymic expressions, i.e. whether they can understand the plausible referents of said expressions. The task has been performed on a dataset of 509 metonymic sentences (Pedinotti & Lenci, 2020) and each sentence returned five alternative solutions for each metonymic occurrence. The accuracy of the answers was then judged based on hypernym-hyponym relations as encoded on WordNet (Fellbaum, 2010). Moreover, the contextualized embeddings were considered to understand at what level the models manage to better understand metonymy. The results of this research contribute to advance the understanding of the mechanisms of models like BERT and what kind of semantic information related to the pragmatic use of the human language they are able to process.

Index

I. Introduction

Metonymy is a rhetoric device, which has several advantages: it usually requires little effort both in production and in comprehension (if the right circumstances are granted), and, therefore, is quite often exploited. Examples of such figure of speech can be found in everyday conversation, but they are often employed in several other contexts as well. The following quote can be considered as an example of a metonymic expression used in literature:

"*Friends, Romans, countrymen, lend me your ears*."

- William Shakespeare, Julius Caesar

In this famous citation, the expression "lend me your ears" is indeed to be interpreted figuratively as a metonymy. The reason is that, if we consider the literal meaning, the sentence would loose meaning since a body part can hardly be lent, but if we interpret the expression "lend me your ears" as "give me your attention" given the intuitive connection between "ears" and "listening"/"paying attention" , then the quote is meaningful. As can be seen in this example, it could have been possible to substitute the metonymic expression with its literal meaning since their meanings are equivalent. Nonetheless, using the metonymic version has a whole other effect since it accentuates the action of giving the complete attention to what someone is saying and, therefore, in this occurrence it creates a more theatrical effect. This feature is one of the purposes of metonymy, but, as it will be discussed, there are other reason why said rhetoric figure is so often employed even in normal conversations.

However, even if metonymic expressions can be found in most linguistic exchanges, this phenomenon has received little attention in the computational linguistics, or Natural Language Processing (NLP), field. Hence, this thesis aims to further explore this topic, in order to better understand if machines can deal with the comprehension of metonymy with an accuracy comparable to the

performance of the human mind. Specifically, the aim of this thesis is to further explore a type of task that transformers, a type of deep learning model, are not usually asked to perform: so far metonymy resolution has often been dealt with as a classification task, while the project for this thesis is to employ said transformers to infer some plausible interpretations of metonymic expressions. The transformers included in this research are four: BERT (Devlin et al., 2018), in its base and large versions, and RoBERTa (Liu et al., 2019), also in its base and large version. All four models were asked to process a dataset (Pedinotti & Lenci, 2020) of 509 metonymies, subdivided into six categories: CONTAINER-FOR-CONTENT, PRODUCER-FOR-PRODUCT, PRODUCT-FOR-PRODUCER, LOCATION-FOR-LOCATED, CAUSER-FOR-RESULT, and POSSESSED-FOR-POSSESSOR. In the following table an example of a sentence containing a metonymic expression, taken from the dataset, is provided for each of these categories to better clarify what these types correspond to.

Metonymic type	Example
CONTAINER-FOR-CONTENT	*The gentlemen had somehow spilled his glass*.
PRODUCER-FOR-PRODUCT	*The children memorized the poet.*
PRODUCT-FOR-PRODUCER	*The actress thanked the magazine.*
LOCATION-FOR-LOCATED	*The church was singing the hymns.*
CAUSER-FOR-RESULT	*The woman heard the smoke detector.*
POSSESSED-FOR-POSSESSOR	*The regime assassinated dissenting voices.*

Through the resolution of a masked element contained in a prompt sentence, each of these models returned five alternative referents for each metonymic entry in the dataset. The obtained results were firstly judged on the basis of the hypernyms-hyponyms relations, as encoded on WordNet (Fellbaum, 1998), using three different strategies: the first strategy consists in checking the presence of the lemma produced by the transformer in the lemma lists generated by the

hypernyms; the second strategy created a semantic space for each metonymic type consisting of the synsets contained in the hypernyms and all the synsets of the answer returned by the transformer were given a 0 or 1 score based on their absence or presence in the semantic space and then the average score was computed; lastly, the third strategy checked for the presence of the most frequent synset, i.e. the first synset, of the target lemma in the same semantic space of the previous strategy. By applying the three different strategies, an overview of the overall performances and the performances according to metonymic type was created in order to judge which of the models dealt best with metonymy resolution.

For the second experiment RoBERTa large was chosen since it was the model that overall performed slightly better compared to the other and was used to investigate the process this model implements in order to solve the task of interpreting metonymic occurrences. More specifically, the output of each layer of RoBERTa was analysed to establish at which of the 24 layers the transformer seems to better understand and process metonymy. To do so, for each entry in the dataset three instances were taken into consideration, namely a sentence where the target word was used metonymically, a sentence where the target word was used literally, and lastly a sentence that contained a plausible paraphrase of the metonymic target word. Then, the contextual embeddings of the target words of these three instances were extracted and compared at each hidden state of RoBERTa. For each entry in the dataset, a measure calculated on the basis of the difference between the cosine similarity of the metonymic target word and its paraphrase and the cosine similarity of the metonymic and literal expressions normalised for the cosine similarity of the literal target word and the metonymic paraphrase was generated at each layer of RoBERTa. The 24 measures thus obtained represent the trend of the level of understanding of RoBERTa when dealing with metonymy.

On the basis of the findings from the two experiments, the conclusions were that transformer language models do not seem to perform on metonymy as well as

they do on literal language given their generally low accuracy in the first experiment. Nonetheless, based on the analysis of the second experiment, it is argued that, despite the overall unsatisfactory answers in the first experiment, at least RoBERTa large seems to notice in the processing at the last layers that in the case of metonymic expression there is more than the apparent literal interpretation.

II. Linguistic theories: from formal to neurolinguistics

In this chapter, a definition of metonymy will be presented. Subsequently, the main theories concerning a formal and cognitive linguistic explanation of metonymy will be introduced, as well as the psycholinguistic and neurolinguistic perspectives on such figure of speech.

2.1 A definition of metonymy

Rhetorical devices are devices to add nuances of meaning to utterances. Figures of speech are used with little additional effort in everyday communication to better convey specific ideas and express how we perceive the world. As defined by Kienpointer (2011), figures of speech (also called "rhetorical figures", "rhetorical devices", or "figures of rhetoric") are "the output of discourse strategies for creating communicatively adequate text". Given their versatility, there is a great variety of figures of speech to fulfil different purposes. The downside of using figurative language is that the majority of rhetorical figures requires additional effort in order to be produced. Therefore, speakers, as well as the other participants in the conversation, are more prone to be consciously aware of having employed figurative language in their speech. For example, although very useful, metaphor production and comprehension involve a thoughtful and creative process (Steen, 2009) and, therefore, both the speaker and the hearer are required to increase the effort to successfully deliver the concept expressed in a metaphorical expression. However, this is not always the case: other rhetorical figures are produced and interpreted almost subconsciously given the fact they may involve more spontaneity and automatic processing (Lakoff, 1987; Sperber & Wilson, 2002). Among the several rhetorical devices of this latter category, this thesis has the aim to analyse a case which is less taken into consideration, while it occurs repeatedly in everyday conversation: metonymy. Metonymy can be

described as “a process which allows us to use one well-understood aspect of something to stand for the thing as a whole, or for some other aspect of it, or for something to which it is very closely related” (Gibbs, 1994). For example, in a conversation, “the Crown” might be uttered not to refer to the ornamental headdress worn by a monarch, but rather to designate the person who wears the head ornament. As humans, the participants in said conversation do not usually have much trouble understanding this mechanism and can correctly infer the meaning as intended by the speaker. Said process is first and foremost a cognitive process since it is involved in production, comprehension, and in gaining knowledge. As a matter of fact, Kövecses (2006) defines metonymy as “a cognitive process in which one conceptual element or entity (thing, event, property), the vehicle, provides mental access to another conceptual entity (thing, event, property), the target, within the same frame, domain or idealized cognitive model”. The reason why Kövecses does not define metonymy as a strictly linguistic phenomenon but rather as a broader process, involving abilities other than language is that metonymy can be found even in other aspects of everyday life. Metonymy can be delivered both verbally and visually, and thus metonymic representations are exploited in the most diverse fields, which space from literature to art and cinema. Therefore, metonymy could give us some insights into how we perceive reality and which communication strategies we employ to linguistically or visually convey the connections created in the brain.

2.2 Advantages and disadvantages of metonymy within a conversation

As it has been analysed by Littlemore (2015), there are several reasons why metonymy occurs quite often in ordinary conversations, while one could simply refer to the “original” object. Firstly, as argued by Lakoff and Johnson (2008), metonymy allows us to better express how we perceive the world; for instance, in the case of part-for-whole metonymies, the part we pick to describe the whole referent is not random, but it is determined by which part we focus on and,

therefore, communicates our perception of said object. For instance, in the sentence "the grey beard over there ordered a sandwich" the expression of "grey beard" is probably used to indicate a man whose main facial feature was his grey beard, and in the sentence said feature is used to indicate the whole person because it communicates the first impression that man left to the speaker. Secondly, the use of metonymy helps to facilitate and to speed up a conversation. As Grice (1975) argued by formulating the four maxims[1] of communication, speakers tend to formulate their contributions as informative as required for the current purpose of the conversation, but they aim at being as brief as possible by avoiding unnecessary prolixity as well. In this regard, it could be said that metonymy suits said intents. Moreover, metonymy has a third effect which could be considered ambivalent, since it supports the creation of a unified identity within a community, but at the same time, as it will be discussed later, it creates distance between different communities. This is due to the fact that metonymy often employs specific references or ways to observe the world which might be shared only among people belonging to the same cultural backgrounds, thus creating a sense of oneness. For instance, the slang expressions generally used by younger generations can be considered as metonymic instances to create a sense of belonging: to give an example, "tea" is often used by the youth not to refer to the hot beverage made with boiling water and leaves, but rather to indicate some kind of gossip. Said expression could be difficult to understand for older generation, but a shared slang creates a sense of community among the youth. Lastly, since metonymy involves indirectness, it is frequently used both in politics and comedy. For instance, a politician may rely on such figure of speech to define a position without resulting in harsh statements; on the other hand, a comedian could exploit metonymy to create humour and irony given its ability to offer the occasion for language play. Generally speaking, it could be said that metonymy is

[1] The four maxims suggested by Grice (1975) are as follows: the maxim of quantity, the maxim of quality, the maxim of relation, and lastly the maxim of manner. His theory implies that the speakers involved in a conversation aim at making their statement exactly as informative as necessary, they don't say anything they believe to be false, they intend to be relevant, and they try to be as clear as possible.

a useful and powerful tool in language because of its property to enrich a single word with additional layers of meaning.

However, using a figure of speech comes with some risks. The first disadvantage is an intrinsic property of metonymy: since its intended meaning is not literal, there is the chance that it might be misinterpreted, such as in the case when a metonymic expression is interpreted as literal. Furthermore, the interpretation of metonymy can be considered successful only if there is sufficient shared common ground built between the participants to the conversation so that the hearer can correctly infer whether the intended meaning is literal or figurative. Lastly, as previously mentioned, metonymies can have as references some cultural aspects that might not be shared among different speech communities. Therefore, the risk is that the hearer may not be able to understand the underlying reference if the required knowledge is not granted. For instance, evidence of the above mentioned risk of misinterpretation can be found in the study by Littlemore et al. (2018). In their research they investigated the ability of Japanese speakers to interpret metonymic expression in English. Lacking the knowledge of some common English idioms, the Japanese tried to interpret metonymies, such as “Is there anywhere where I can freshen up?”, in a literal manner; however, the literal interpretations results in confusing ideas since that was not the intended understanding of the expressions.

2.3 Theoretical linguistics: the starting point

The human language has the property of creativity, meaning that it is always possible to formulate new combinations of words according to the expressive needs in a conversation. This is also the case with metonymy: the human brain is able to form new connections and produce novel cases of metonymic expressions. However, throughout the decades researchers have been able to develop several schemas in order to categorise and regulate the majority of the metonymic examples. In her book “Metonymy”, Littlemore (2015) recapitulates the main

models for metonymy, that have so far been proposed. The first taxonomy she goes through is the model suggested by Radden and Kövecses (1999). According to Littlemore, this taxonomy represents the first hierarchical model which provided researchers with a common language to describe the phenomenon of metonymy. Radden and Kövecses show that said figure of speech can be divided into two types. The first category stands for whole-and-part metonymies, indicating the cases where a part of the object we want to mention is being used to designate the whole object and vice versa, such as when the term “America” is used to define just the United States. The second category, instead, is dedicated to the so-called part-and-part metonymies, which relate to the instances where a concept is used to refer to another which is related, such as when it is said that someone married “money”. In this case, “money” is used to indicate a person; however, “money” is not part of a person, but rather it belongs to said person and it is considered as a defining feature. These two types of metonymies just mentioned represent the major division, but for each of the two, there are multiple sub-categories to describe more specifically the various examples of metonymy. In the following schema, this further division is best represented.

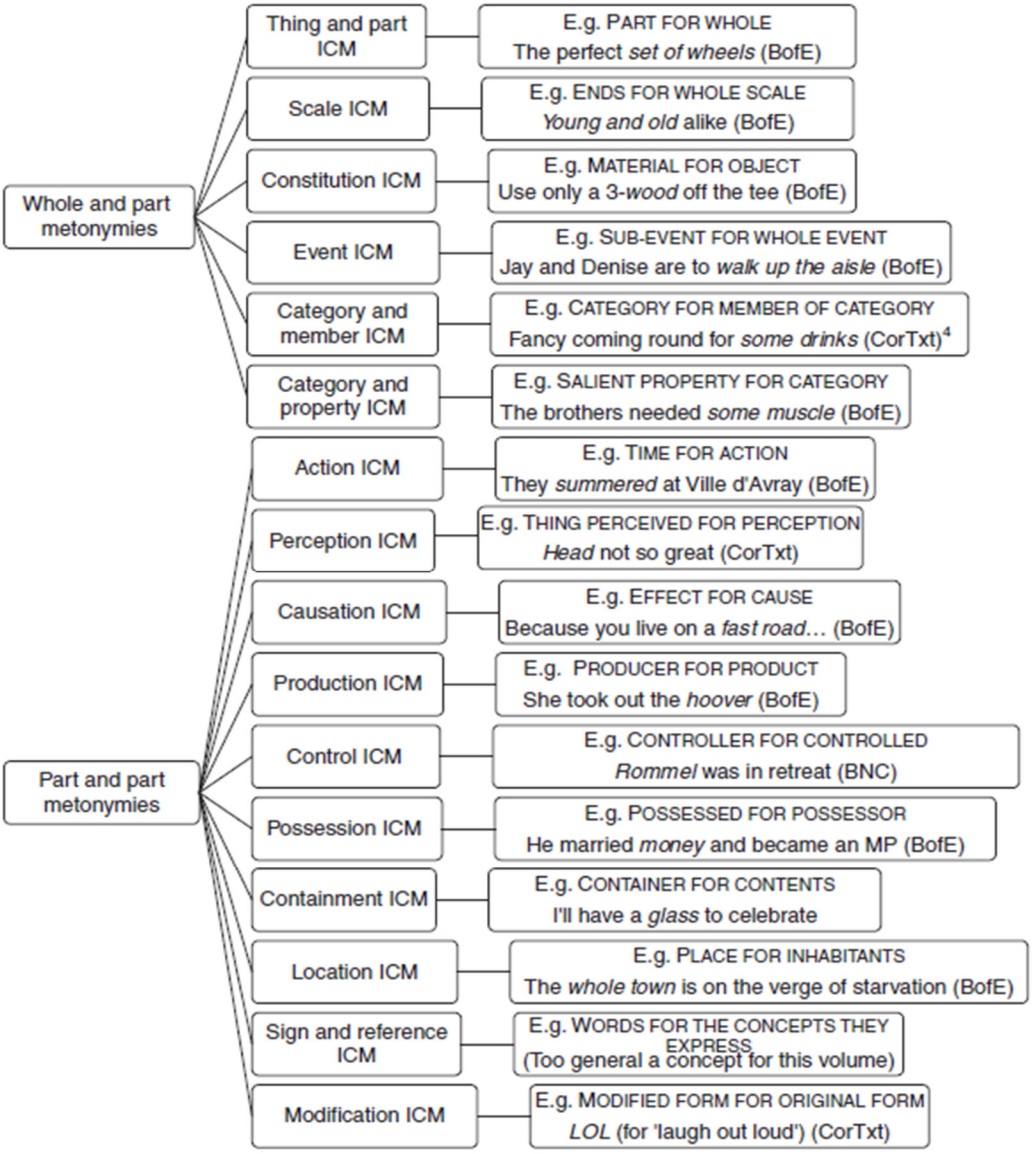


Figure 1 - Radden and Kövecses' taxonomy [source: Littlemore, 2015, pp.22]

The taxonomy presented in Figure 1 shows us that, despite the division in only two categories, each of the two types of metonymy present a further division into several sub-categories. It is interesting to note that even though all the subcategories share a common feature determined by the bigger grouping, they represent the most different instances of metonymies. For example, in the part-and-part metonymies, we can find instances concerned with abstract concept

such as effect-for-cause metonymies and other instances involving human subjects such as possessed-for-possessor metonymies. What distinguishes the different instances of metonymy is the ICM they relate to. An ICM, which stands for "idealized cognitive model", has been defined as "a series of embodied, encyclopaedic, abstract, loosely connected and somewhat idiosyncratic knowledge networks that we have in our minds" (Lakoff, 1987; Radden and Kövecses, 1999). This kind of knowledge is acquired over time without particular effort since it deals with the human experience of the world. Moreover, in the case of metonymy it is rare that the knowledge pertaining to one ICM constitutes enough information to correctly infer the intended meaning, but rather accessing that knowledge requires the activation of several ICMs. To give a general idea of how an ICM can be defined, Lakoff (1987) differentiates five types of models: propositional, image schema, metaphoric, metonymic and symbolic. For example, image schemas represent the building blocks of how human cognition works. Image schemas consist of a kind of knowledge which is the first we acquire as children, and it represents how we perceive the world and how the objects are related to one another. An instance of an image schema could be the container schema: through the experience, the realization that some objects can accommodate something else inside them is acquired. For example, we understand in the very early years of our lives that a glass usually contains some kind of liquid that we may drink. Metonymic thinking works in a very similar way: the two concepts, the one we utter and the referent, are usually linked in a quite evident way. Therefore, taking the same example given for the container image schema, if we utter "glass" as a metonymic expression in a sentence, such as "pour a glass", it should be evident to the hearer that the referent of said expression is not the glass itself, but rather what it contains. Given this evident connection between the referent and the metonymic term, in their paper Radden and Kövecses (1999) argue as well that the connection between said concepts is determined by some non-casual principles. These principles regulate the choice of the vehicle, the metonymic expression, we pick to mention the original referent.

All the principles they list fall into five main categories, namely human experience, perceptual selectivity, cultural preferences, communicative principles, and overriding factors. More specifically, while selecting a the vehicle of a metonymic expression, respecting the human experience principles indicates that human and concrete objects are preferred to non-human or abstract subjects; according to the perceptual selectivity principles, we may prefer immediate over non-immediate and dominant over less-dominant; cultural preferences favours the selection of ideal and stereotypical ideas; communicative principles consist in striving for clarity and relevance in communication; finally, overriding factors can be represented by the use of rhetoric figures and social-communicative effects. To give some more understandable examples, one principle is the property of concreteness: concrete objects are more likely to be picked rather than abstract ideas. An example could be the sentence "the clock is ticking" used to mention that the time is passing: here the abstract concept of time is replaced with a concrete object, the clock, in order to visualise the idea. In addition to this example, another principle guiding the choice of metonymic expression is the so-called "initial or final over middle" principle. According to said assumption, the extremes of a concept are more likely to be selected to form a metonymic expression than things that fall in the middle. For instance, it is most common to hear "from January to December" to designate the whole year rather than "from March to February", even though they are both true and refer to a whole calendar year. Regardless of how the connection between the referent and the word used to designate it is formed, the fundamental aspect is that the two can be found within the same domain. The concept of "domain" has been formally introduced by Langacker (1987) and borrowed from Lakoff (1987), and it is similar to the concept of ICM in that they both constitute a knowledge structure. However, the idea of a "domain" is more concrete and the concepts inside a single domain are somehow related. As a matter of fact, ICM contains ideas that are more loosely connected and abstract, while a domain is more strictly defined and the range of ideas belonging to a domain is not as wide. Some examples could be the domain

of TIME, LOVE, SPACE, JOURNEY, and so on. As Lakoff (1987) suggests, metonymy involves mapping from one concept to another within a single domain. This definition is crucial to distinguish metonymy from metaphor. To explain metaphor the concept of domain is again employed, however in metaphor, the mapping is across different domains. To clarify the matter of mapping, a pair of sentences can be compared to observe such difference. If we consider the expression "that car is a dinosaur" to state that the car is indeed very old, we can define the two nouns "car" and "dinosaur" as belonging to two different domains, namely the domain of VEHICLE and the domain of ANIMAL; in this sentence, however, they are unified through the use of a metaphor to compare an old car to an extinct animal and, therefore, we map "car" from the domain of VEHICLE to an item, "dinosaur", in a different domain. On this evidence, the mapping is defined as across domains. On the other hand, in the case of metonymy the mapping happens within the domain: for instance, to refer to the car we might utter "set of wheels" and, since the wheels are part of the car, the mapping for the creation of this metonymy is conducted within the domain of VEHICLE.

Despite this seemingly clear-cut distinction between metonymy and metaphor, there are borderline cases: Hilper (2006) noticed some instances such as 'kind-of' relations, which sometimes should be interpreted through metonymy and sometimes through metaphor. For instance, two expressions that both belong to said category are "to keep an eye on" and "hand in glove" in sentence such as:

a. *Marcus Judge had kept an eye on her finances from the beginning.*
b. *The drug barons work hand in glove with the pharmaceutical industry.*

In both sentence, the previously mentioned expression have to be interpreted figuratively; however, in sentence a, since "keep an eye on" means "be attentive" and said meaning is a hypernym of "to watch", the knowledge can still be mapped within the same domain, resulting in the expression in a metonymy. The expression "hand in glove" in the context of sentence b means "accordant", which

is an hypernym of “physically fitting”; since “accordant” and “physically fitting” do not belong to the same domain, this expression should be considered a metaphor.

2.4 Further theoretical linguistic theories

As mentioned before, the model by Radden and Kövecses (1999) is not the first and only model created to schematise metonymy. Not only the other models mentioned by Littlemore are relevant to the field, but they are also interesting to be analysed and compared given the different perspectives they offer on the topic of metonymy. For instance, as it will be analysed more in-depth afterwards, in order to classify the different phenomena, some models deal with metonymy considering the aim of the metonymic expressions (Warren, 1999; Panther and Thomburg, 1998,), others centre the observation on the relation between the source and the target domain (Ruiz de Mendoza Ibáñez and Diez Velasco, 2002), further other models focus on the kind of contiguity present in each metonymic utterance (Peirsman and Geeraerts, 2006).

Despite having a similar perspective, Warren (1999)’s and Panther and Thornburg (1998)’s models differ in the categories they distinguish to schematize the different occurrences of metonymy. Warren(1999) argues that there are two main types of metonymy, namely referential and propositional metonymy. The referential type refers to those occurrences where an entity is related to another, such as “Shakespeare” in the sentence “People are hungry for Shakespeare in America” while the propositional type relates to those cases where it’s a proposition that relates to another, for instance “raise the eyebrow” with the meaning of “being surprised” in the sentence “Rosalind raised her eyebrows and held out her hand”. On the other hand, Panther and Thornburg (1998) suggest two different kinds of metonymy. The first category is represented by prepositional metonymies, further broken down into referential, which are the same as those suggested by Warren (1999), and predicational, which usually involve a link between events. The first type is represented by sentences such as “The growing

list of countries where the buses are on strike" where "buses" stands for "drivers"; while the second type contains phrases like "He was able to tell me that it had merely gone into spasm", where "able to tell" does not really entail an ability but rather the fact that the person did communicate the information. The second category, instead, contains the illocutionary metonymies, which are generated based on pragmatic inferences. An example of illocutionary metonymies can be found in a sentence such as "Have you got a fiver? I want to pay the boy for his petrol" where the information required by the speaker is not whether the hearer is in possession of a fiver but whether the money can be lent.

A few years later, Ruiz de Mendoza Ibáñez and Diez Velasco (2002) propose another hypothesis on how to deal with the phenomenon of metonymy. They argue that it is not relevant to distinguish the aims of a metonymic expression, but rather to consider the relationship between said expression and the intended referent. Consistent with this view, they report two instances of metonymy, namely the "target in source" and the "source in target" metonymies. To the first class belong those examples where the metonymic word is part of its referent and, therefore, the interpretation involves a domain reduction of the metonymic expression, while to the second the opposite case, that is the instances where the referent is part of the metonymic expression and, therefore, the interpretation involves a domain expansion of the metonymic expression. The previous class consists of sentences such as "The great contribution that the Pill has made to personal choice" where the term "Pill", which is normally used to indicate any type of medicament, refers specifically to the contraceptive pill, making "Pill" a sub-domain of the metonymic vehicle, and requires a restriction of the domain PILL; while the latter class is represented by sentences like "All hands on deck" where "hands" refers to the sailors who are doing hard physical work, making "hands" a sub-domain of the referent, and requires an expansion of the domain BODY PART.

The last theoretical linguistics model reported by Littlemore is Peirsman and Geeraerts' (2006) model. This approach takes a radically different perspective on the matter. The distinction between different occurrences of metonymy they

suggest is not based on clear-cut, parallel categories. Instead they suggest the idea of “radial categories”, which indicate a type of classification that separates the prototypical elements from the non-prototypical elements putting them in a continuum. For instance, taking the radial category of “pet”, we may find at the centre as prototypical elements words such as “cat” and “dog”, while at the periphery words such as “snake” or “tiger” could be more likely found. The idea that Peirsman and Geeraerts suggest is that metonymy instances should be analysed according to the radial category principle as well, defined by the type of contiguity between the vehicle, i.e. the metonymic expression, and its referent. For example, taking into consideration the sentence “I’ll be able to eat every day and have a roof over my head”, we can establish that the metonymic expression “a roof over my head” indicates and is part of a house, making this kind of metonymy prototypical. An example of a peripheral metonymy, instead, can be found in the sentence “Clinton plans a round table discussion”: here “round table” refers to the assembly which might be gather around a round table, but not necessarily; the assembly cannot be put in contiguity with the round table, thus collocating this instance of metonymy at the periphery.

2.5 Cognitive linguistic perspective

In addition to the prior mentioned models, which have a more theoretical linguistic perspective, other studies offer a different view on the matter of metonymy, namely the cognitivist approach. Cognitivists argue against Chomsky’s theory of Generative Grammar (1988) and suggest instead that language operates in the brain according to three cognitive principles. The first of these principles is opposed to generative grammar and argues instead that language is not an innate cognitive faculty; the second principle is opposed to truth-conditional semantics and argues instead that the semantic metalanguage should not be evaluated in term of true or false according to a model of the world; lastly, the third principle is opposed to reductionist tendencies that support the hypothesis that

grammatical form and meaning should be represented in an abstract and generic manner and supports instead that “language knowledge emerges from language use” (Croft & Cruse, 2004). Moreover, cognitive linguists emphasised the role of metaphor and metonymy, recognizing that how language is structured is determined by conceptual and pragmatic factors. Concerning metonymic expressions, Langacker (1993), for example, suggests a cognitive linguistic theory, which states that, for a metonymic expression to be produced, a shift in profile has to occur. In other words, the sense of a word changes according to the context in which the same word is uttered: the “Parliament” could refer to the building in which the assembly is held, but also to the group of people which constitute the government. The sense of the expression can only be selected on the base of the context and a shift is required in order to be able to adapt the specific meaning to the whole sentence. Langacker (1993)’s theory can be included in a cognitivist frame since he introduces the concept of “active zone”, the area which gets activated from the shift of focus on the different meanings of an expression. This shift is allowed by the connections that can be formed in the brain. It is always possible to form new connections and, therefore, to create new metonymic links. However, Langacker(1993)’s thesis presents a criticism, namely the fact that from this perspective, any expression could be interpreted as a metonymy since a different focus has to be selected each time to correctly adapt the general idea to the specific context.

To solve said criticism, Barcelona (2003, 2011) proposes the idea that metonymies have to be dealt with in a radial category manner. The type of metonymies that lie at the core of the diagram are those based on contiguity and are usually represented by referential metonymies; they are the so-called prototypical metonymies. Then, there is a bigger group that contains the previous case with the addition of other examples, namely the typical metonymies, in which the target is distinct from the source. Finally, the outer circle, containing the two subdivisions, consists of the schematic metonymies: this is the type of metonymy as intended by Langacker (1993) and requires domain highlighting.

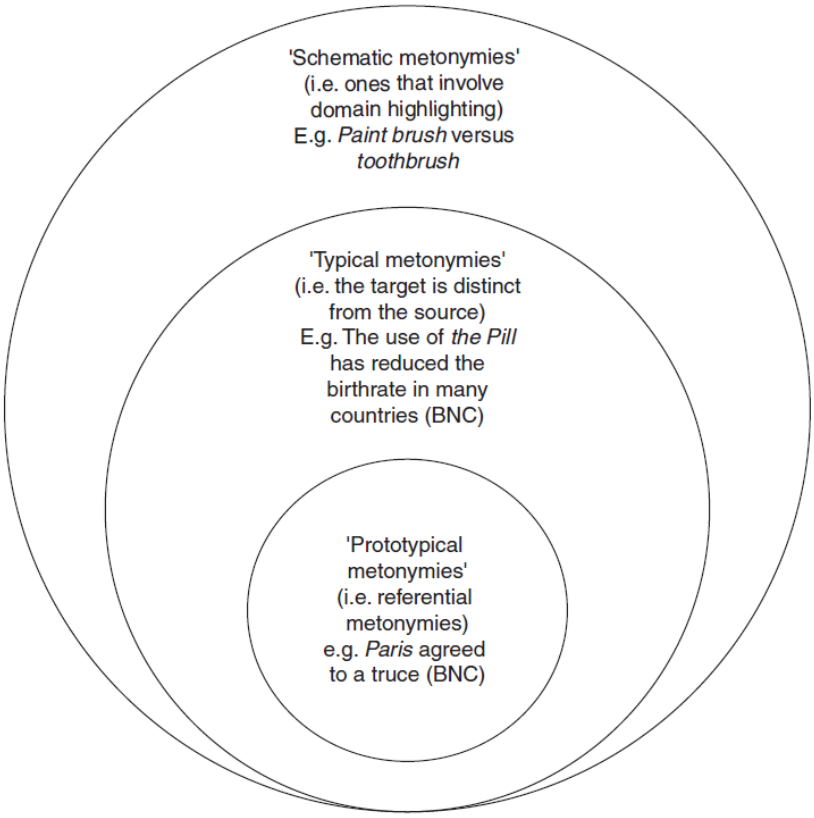


Figure 2 - Barcelona's model. Source: Littlemore, 2015, pg.57

Further development of this theory is provided by Handl (2011). Her model has a very similar schema to Barcelona(2003,2011)'s. However, Handl (2011) introduced the concept of "underspecification of meaning" and subsequently "underspecified metonymies", which accounts for the intermediate category. The reason why this type of metonymy has an underspecified meaning is that both the basic sense of the vehicle term and the contextual meaning contribute to the interpretation of the metonymic expression. In practice, the brain does not distinguish between the referent and the vehicle term, both meanings remain underspecified and are joined to form one "functional unit". Therefore, it is argued that the metonymic expression stands for both concepts and cannot be simply considered as a replacement for the intended referent. The main strength of Handl (2011)'s theory is that is based on the observation of real-world data from the British National

Corpus: the examples are not formulated by the author and, therefore, possibly biased, but they represent how metonymies are produced in everyday conversations.

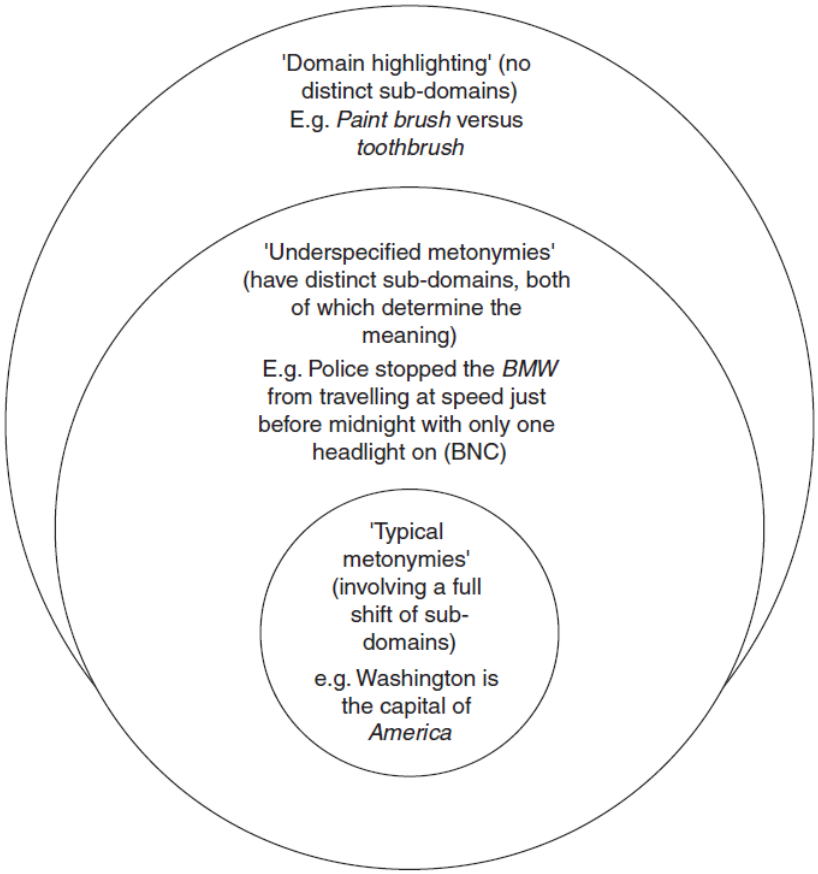


Figure 3 - Handl's model. Source: Littlemore, 2015, pg.58

2.6 Recent theories

More recently, other theories with different approaches have been suggested to deal with metonymy. Three main examples of such theories can be identified, namely the blending theory (Coulson and Oakley, 2003), the relevance theory (Sperber and Wilson, 1987, 2004), and the complex system theory (Biernacka, 2013). What these theses have in common is a less strict approach to outlining metonymy. Moreover, the focus of such hypothesis is less on the different types of domains previously suggested, but rather on the role this figure of speech has

on the influence that context has in the interpretation and production of metonymic expressions.

Coulson and Oakley (2003) are the main proponents of the blending theory, which draws inspiration from Fauconnier and Turner (1999) and states that the way the meaning of an expression is determined has to do with the previously cited concept of idealised cognitive model (ICM) and often requires the combination of more than one of such spaces. In the blending theory, metonymy has the role to allow the comprehension of the different concepts that happen to be compressed into just a few words; from said perspective, distinguishing and understanding the elements that compose the concept being uttered grants a better comprehension of the different facets of meaning and metonymy proves to be a useful tool to do so.

On the other hand, in the relevance theory metonymy is not considered a tool to understand meaning, but a device to foreground the importance of context in the comprehension process. In the relevance theory, as formulated by Sperber and Wilson (1987, 2004), it is argued that the person involved in the conversation as the hearer will take for granted that any information communicated is in some way useful to the purpose of the conversation and such knowledge can, therefore, be used as a clue to understand what is later said. According to more recent theorists, metonymy is interpreted through a mechanism similar to the same principle of relevance: world knowledge should be sufficient to correctly deliver the intended meaning and determines why some metonymic expressions are considered more acceptable than other. For instance, in the case a person calls a friend saying “I am parked out back”, there are a two conditions the hearer must be aware of in order to correctly infer that the “I” is actually referred to a vehicle, probably a car: firstly, there should be a “salient correspondence” between the person and the car, namely the person has to be considered the owner of the car and that same person has parked out back; secondly, it should be granted that communicating this information is useful for the friend because it could be the case that the person is giving a lift to the friend. If these presented conditions are

respected, then the interpretation of the metonymic sentence “I am parked out back” should cause no trouble.

Lastly, the complex systems theory distances itself from linguistics since it is more of a sociological matter. For instance, Larsen-Freeman and Cameron (2008) argue that to explain human behaviours it is necessary to consider and comprehend all the determining factors that took place before the situation in which such behaviours present themselves. The complex systems theory has been applied several times to metaphor, but more recently Biernacka (2013) hypothesized that it could be applied to metonymies as well. The reason is that, similarly to metaphors, by highlighting different aspects metonymies influence the way participants understand the conversation. For instance, Van Dijk (1998) analysed speeches given by politicians and he found a recurring pattern of using positive connotated characteristics to refer a “in-group”, probably the country the politician was speaking to, as opposed to negative characteristics to designate the “out-groups”. Using such opposite delineations for different countries has the aim of influencing the perspective of the citizens, creating a “us-versus-them” rhetoric On the basis of this consideration, the linguistic and the social perspective cannot be considered as different matters but rather contribute to further explain why we say what we say in such specific way.

2.7 Theories on how metonymy is processed in the brain

Given the fact that metonymies do not represent a purely linguistic phenomenon, but they are highly involved in cognitive processes, the study of such rhetoric device has been carried out even in the branches of linguistics which are more involved in how language is processed in the brain, namely psycholinguistics and neurolinguistics. The main interests have been to analyse whether literal language and figurative speech are dealt with in the same way or differently (Frisson and Pickering, 1999; Annaz et al., 2008), and the role of both context and syntactic information in the comprehension and interpretation of metonymy (Lowder and

Gordon, 2013). In order to proceed with said analyses, the most common research methods involve eye tracking, brain scanning, reaction-time studies, and metonymy comprehension tasks.

As reported by Littlemore (2015), the main studies in this field of research are somewhat more recent than in the theoretical linguistic branch. Part of the pioneer work is presented by Frisson and Pickering (1999), who conducted two eye-tracking experiment. In the first experiment, by using an eye-tracking software they registered how long it took the participants to read literal sentences and then compared the results with the time require to read sentences with conventional metonymies. The scholars aimed at analysing how metonymy is processed in the brain compared to the literal meaning. In the second experiment, they added to the experiment material less common metonymies and compared the performance of the participants when dealing with often recurring metonymies or with novel metonymies. Based on their findings, Frisson and Pickering (1999) argue against the idea that neither the literal language nor the figurative meaning is processed first, while they suggest that initially, the meaning remains underspecified and, therefore, both senses can be accessed immediately and simultaneously. Only after sufficient context is provided in the conversation, the hearer can commit to either meaning.

One problematic aspect of Frisson and Pickering (1999)'s research is that the syntax of the metonymic expression is not considered a variable, which could influence the rate at which the meaning is processed. In their study, Lowder and Gordon (2013) suggest investigating if and how phraseology plays a role in the interpretation of metonymy. On the basis of previous studies that demonstrated that the human brain processes at a different depth different parts of a sentence, they hypothesised that the position in which the metonymy appears influences how much attention we pay to said expression. In order to prove their thesis, they repeated Frisson and Pickering (1999)'s experiment, but they divided the dataset according to where the metonymic expression would appear, namely either in the direct object or in an adjunct phrase. Their results showed that when the

metonymy appeared as the direct object the participants required more time to process it, probably because they found it more complicated to make sense of the expression, while the process was quicker when the metonymy was found in the adjunct phrase because the focus was on another part of the sentence.

Other than the previously mentioned studies which exploit traditional techniques, further research has employed more advanced instruments to analyse the process of metonymic interpretation. To give an example, in Rapp et al. (2011) functional Magnetic Resonance Imaging (fMRI) is utilized to investigate which parts of the brain receive activation when dealing with a metonymic expression. To do so, they analysed the scans of the brain of fourteen healthy participants and they look for which parts of the brain were activated respectively when asked to read literal sentences (e.g. “Africa is arid”) , metonymic sentences (e.g. “Africa is hungry”), and non-sense sentences (e.g. “Africa is wollen”). They found out that the most involved side of the brain is the left middle temporal gyrus, as well as the inferior frontal gyrus of both hemispheres. The left middle temporal gyrus is also the part responsible for syntactic processing and the interpretation of novel metaphors, while the inferior frontal gyrus is involved in the unification of discourse information and previously stored knowledge. These results seem to support the hypothesis that ICMs play a fundamental role in metonymy comprehension.

Additional supporting evidence to these previously presented findings comes from a meta-analysis conducted by Rapp et al. (2012) on prior MRI experiments involving figurative language. All the studies that they took into consideration led them to the conclusion that figurative language caused the activation of both sides of the inferior frontal gyrus and, more specifically, the right hemisphere is mainly responsible for processing novel figurative language. Although this study does not focus particularly on metonymy, it is still in line with previous findings and, therefore, provides additional evidence.

2.8 Developmental studies of metonymy

Aside from understanding how and where metonymies are processed in the brain, it is also relevant to understand when metonymy comprehension and production start to develop. The main hypotheses concerning this development are that it either depends on age or vocabulary knowledge. Rundblad and Annaz (2010a)'s study has the aim to understand when children start to develop the ability to comprehend metaphor and metonymy, compare their performance to the adult's and, lastly, determine the role played by vocabulary size. To conduct the experiment, the researchers formulated the experiment as a story/picture comprehension task: children were asked to observe four pictures accompanied by a short capture which contained either a metonymic or a metaphorical expression, and then they were asked a comprehension question at the end. The results seem to show that the ability to process metonymy precedes the ability to process metaphor and it is fully developed by the age of twelve. In addition to that, the researchers observed how vocabulary size influenced more strongly metonymy rather than metaphor comprehension. Some possible explanations for these findings are that the interpretation of metonymy requires fewer cognitive patterns, or that most metonymies are recurrent and very similar versions of the same metonymic phrase can be found in several languages.

Not only children are capable of understanding metonymic expressions, but they showed great ability at producing metonymies, even novel ones. Through a process, which has been defined as "creative metonymical shrinking" (Nerlich et al., 1999), children have proved to be able to connect concepts and form, even unusual, metonymic phrases in order to express their ideas with less effort. For instance, the reasoning behind a child's utterance such as "Mum, I like being a sandwich", which may seem nonsensical at first, is that there may a group, which the child is part of, that usually has sandwich as a snack; the child may particularly enjoy being part of said group and, therefore, what he means to say by shrinking the two concepts together is that he enjoys being part of the group that usually gathers during the break to enjoy their sandwiches together. This result is a highly

creative use of metonymy, whose meaning requires more effort in order to be unpacked.

2.9 Dealing with metonymy in the case of linguistic impairments

Another factor affecting language comprehension and production is represented by linguistic impairments. The impairments can be due to different causes, such as some kind of syndrome or disorder or brain damage. Such cases are often investigated because they can provide an explanation on how the mind and its processes work. This is the reason why cases of children with linguistic impairments are involved also in studies to investigate metonymy: there are a few examples where the conditions gave an insight on how metonymy is dealt with from a cognitive perspective.

In the study conducted by Annaz et al. (2008) children with Williams syndrome, which is a genetic disorder that affects different part of the body, were involved in an experiment to investigate their ability to process metaphoric and metonymic expressions. From a cognitive perspective, people with Williams syndrome show a mild to moderate intellectual disability, however language abilities are normally intact (Burn, 1986). Moreover, people affected by this syndrome usually have shown good performances when dealing with everyday language; however, they lack pragmatic skills. This shortcoming results in stereotyped replies and missing clues on when it is appropriate to intervene in a conversation. Since the interpretation of figurative language is not a purely linguistic task, but it also involves pragmatic knowledge, the question to analyse was whether the lack of pragmatism would result in an inability to comprehend figurative language. Thanks to their experiment, which was built the same way as in Rundblad and Annaz (2010a), Annaz et al.(2008) showed that children with Williams syndrome deal better with metonymy rather than metaphor. Moreover, their ability to comprehend metonymy was directly proportional to the children's vocabulary knowledge. Based on such findings, the authors drew the conclusion that, in

contrast to metaphor, which requires additional cognitive mechanisms, metonymy can be considered comparable to regular language.

However, the previously presented research failed to take into consideration a possibly influential variable, namely novelty versus conformity of the metonymic expressions presented to the participants. Having already heard a specific expression could facilitate the access a second time to the same item, resulting in unreliable experimental evidence. Therefore, in order to solve said criticism, Van Herwegen et al. (2013) repeated the same experiment as the previously presented study to investigate whether such a variable has an impact on the results. The findings show that the development of comprehension skills was delayed in the group affected by the Williams syndrome compared to the control group and it was not affected by the metonymies being conventional or novel. Therefore, the repetition of the experiment seems to prove that the "novelty versus conformity" factor is not as impactful as other variables, such as semantic knowledge, for example.

Another impairment often considered in several linguistic experiments is autism. Autism is better defined as autism spectrum disorder (ASD) and, since it is a spectrum, it is hard to define the general characteristic, which could apply to every individual with said condition. This consideration can be applied to linguistic abilities: some people with ASD are non-speaking, while others perform at the same level as neurotypical people (Lord et al., 2018). Even though language abilities can significantly vary across subjects, a common aspect of people with ASD is their impairment in pragmatic abilities (Tager-Flusberg, 2006). As already mentioned, processing the various rhetorical figures requires not only linguistic knowledge but also a certain degree of pragmatism. Therefore, Rundblad and Annaz (2010b) compared the performance of children with ASD and typically developing children. To do so, they created an experiment which comprehended both metaphor and metonymy comprehension and the structure and the procedure of the experiment was the same story/picture task as in Rundblad and Annaz (2010a). The results showed a remarkably poorer performance regarding

both figures of speech in children with ASD compared to the neurotypical subjects. Moreover, the results of this experiment with children with autism resemble the findings of the experiment with children with Williams syndrome in that the deficit in metaphor comprehension is correlated to chronological and mental age, while the ability to comprehend metonymy was linked to the vocabulary knowledge of the children involved. On the basis of these findings, Rundblad and Annaz (2010b) argue that the performance of children with ASD in metaphor and metonymy comprehension was significantly affected at all ages. In addition to that, based on the different performances when dealing with metaphor or metonymy, they reach a similar conclusion as in their previous study: metonymy comprehension proved to be more easily faced because it functions similarly to literal language and does not require mapping across different domains, like metaphor.

III. Computational linguistics

After having described metonymy from the point of view of theoretical and psycho-linguistics, in this chapter the approach of computational linguistics will be considered. The theories which will be discussed go back to the first experiments with metonymy, but also the most recent studies with advanced technologies will be included.

3.1 Is it or is it not metonymy? First computational approaches to Metonymy Resolution

Since the 1950s when the United States started to experiment with automatic translation. Computational linguistics had to deal with increasingly complex language phenomena, in order to improve performances. Since speaking a language involves creativity, which is a property machines are not primarily designed to deal with, researchers started to investigate how much computers can understand in the case of figurative language. This process involved metonymy as well. Since the 1990s, linguists started to experiment with asking machines to solve metonymic expressions. Since then, even though it has not received as much attention as other linguistic aspects, this area of research has developed. The relevance of computationally solving metonymic expressions is demonstrated by the many use this kind of task could have: for instance, in the NLP field, metonymy resolution tasks are employed in machine translation (Kamei and Wakao, 1992; Brdar & Brdar-Szabó, 2013; Zhi, 2020), question answering (Stallard, 1993; Harabagiu, 2008), and anaphora resolution (Harabagiu, 1998; Markert and Hahn, 2002; Zhao, 2014).

Pioneer studies dealing with metonymy resolution tasks were conducted in the 1990s, but they were based on small-scale corpora; it was only in the early 2000s that scholars started to take a corpus-based approach or to test their models on

larger datasets. A key aspect of the models proposed in this era was that metonymy resolution was treated as a classification task: most times researchers would gather a corpus of one type of metonymies and build models that could classify the target expressions as either literal or metonymic expressions, sometimes further divided into different categories. One of the first papers to go down this newly suggested path was Markert and Nissim (2002). The paper dealt with metonymy and suggested a computational algorithm to solve them. In their study, they considered a very specific type of metonymy to be analysed, namely all the metonymies related to location, subdivided into place-for-people, place-for-event, and place-for-product groups. After creating a corpus where metonymic occurrences were mixed with literal expressions, Nissim and Markert (2002) created a classification task to recognize and divide the metonymic phrases into different classes. The decision to investigate this linguistic phenomenon as a classification was based on the assumption that the interpretation of metonymy is a quite regular task, as previously suggested by Lakoff and Johnson (1980), and therefore can be treated similarly to a word sense disambiguation (WSD) task. The only aspect that distinguishes these two types of tasks is that a WSD task selects a limited number of readings for each entry, while metonymy resolution could potentially require infinite sets of senses related to each word. In order to solve this problem, Nissim and Markert (2002) argued that, even though the senses could be infinite, it is also true that all metonymic expressions can be generally traced back to broader categories. To conduct the experiment, the researchers created a corpus, in which 1000 examples were annotated by hand, and a supervised classification algorithm which could differentiate whether the expressions to be analysed were alternatively literal phrases, place-for-people, place-for-event, place-for-product, mixed or other types of metonymies. In addition to this first consideration, since features, such as co-occurrences, collocations, and grammatical features are usually influential in WSD tasks, Nissim and Markert (2002) included them in the experiment to explore as well the role these feature types play in a metonymy resolution task. The results showed that

co-occurrences do not impact in any way the performance, while the latter features perform well, with the only exception that collocations must be generalised to the semantic class the expression belongs to.

As follow-up research, Nissim and Markert (2003) proposed another more efficient algorithm to perform a similar metonymy resolution task. In comparison to the previous model, the latter innovation was based on employing the similarity among the conventional types of metonymies in order to facilitate their classification. More specifically, two types of similarity were considered in the experiment, namely the similarity among the target words, also called Possible Metonymic Words (PMWs), and the similarity of the context in which said target words occur. So, we can consider the following three sentences as examples:

- *Pakistan had won the World Cup.*
- *England won the World Cup.*
- *Scotland lost the semi-final.*

The similarity was computed through a context reduction step, that simplified the sentences above in “Pakistan-subj-of-win”, “England-subj-of-win”, and “Scotland-subj-of-loose”. Then, it was computed the similarity between the semantic classes of “Pakistan”, “England”, and “Scotland”, the similarity of the roles which was “subj-of” for all three instances, and finally the head similarity between “win” and “loose”. In the following schema, the process is best summarized:

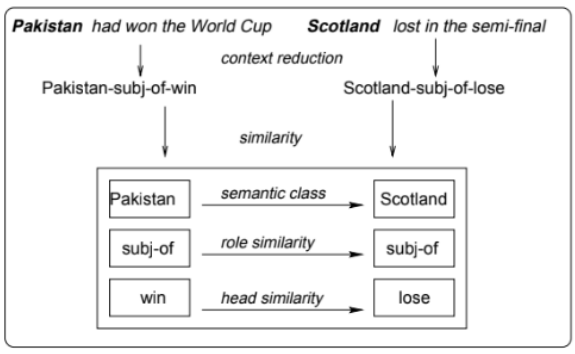


Figure 4 - Computation of the similarity between two examples. Source: Nissim & Markert, 2003, pp.2

Then, through the study of a corpus, Nissim and Markert (2003) observed that the use of a name of a nation to designate its team covered most metonymic instances of this kind and, therefore, they managed to demonstrate the key-role played by metonymic pattern in metonymy resolution. Thus, the same criteria was used to other location metonymies, which did not concern the semantic field of sports but rather events or products. Moreover, Nissim and Markert (2003) demonstrated the usefulness of considering syntactic head-modifier relations as a feature to increase precision, even though they entail data sparseness. In order to solve this problem, the researchers introduced a thesaurus in their study that enabled the model to draw inferences and generalise between the given example and similar lexical heads. This methodology granted a generally better performance of the algorithm on metonymy recognition.

Alternatively to the method suggested by Nissim and Markert (2003), around the same time, another approach was proposed: a probabilistic account was brought forward by Lapata and Lascarides (2003). In their study, they analysed in particular the kind of metonymy called logical. Logical metonymies are those that are involved in the verb phrase, where there is a covered event implicitly included in the verb, such as in the sentence “I began the book”. It could be intuitively argued that the most likely reading is “I began reading the book”, however that is not the only plausible meaning, since it could be possible as well that “I began writing the book”, “I began editing the book”, and so on. On this basis, Lapata and Lascarides (2003) argue that no theory presented until then represented an exhaustive account to deal with the multiple readings each occurrence of said type of metonymy could present. Therefore, they suggest that the likelihood of the possible interpretations should be ranked and a probabilistic model should be built on top of the likelihood analysis. In order to computer the likelihood, Lapata and Lascarides (2003) retrieved from the British National Corpus (BNC)(Burnard, 1995) the co-occurrences of the covered events with the metonymic verbs and with the nouns and ranked them according to their frequencies. Based on the frequencies, they managed to unify them into a formula, which computed the likelihood, or

probability, for the selected covered event. The probabilistic aspect of this kind of model results in an unsupervised approach to metonymy resolution, a new method which was never attempted before. The results from this statistical approach were compared with the judgement of human subjects on the same metonymic sentences and the comparison between the results generated by the model and the answers returned by the participants proved that the performance of the model was overall satisfactory. Thus, the theory suggested by Lapata and Lascarides (2003) showed that a probabilistic model could infer semantic properties from a corpus even if said corpus is not semantically annotated.

3.2 Inclusion of the distributional approach

Towards the late 2000s, a new method was applied to metonymy resolution, namely the distributional approach (Boleda, 2020). This type of approach was based on the Distributional Hypothesis, as formulated by Harris (1954): the similarity in meaning of different words entails a similarity in the linguistic distribution. Therefore, words that occur in similar contexts have probably similar meanings. This hypothesis was the base on which the field of Distributional Semantics was built upon. According to the methods employed in this field, the words should be considered as a whole with the context in which they appear and represented abstractly as vectors. The context influences the direction of such vectors: the same word, in fact, could correspond to vectors with different directions. For instance, a polysemous word such as “wing” indicated both the anatomical part of a bird and a part of a building and to each of these meanings corresponds to at a different vector which would have different inclination according to the sense of the word we are taking into consideration.

Metonymy resolution tasks could benefit from said approach because applying the distributional approach could facilitate the automatic interpretation of metonymic occurrences: if metonymies appear in similar contexts, their referent should be similar at least to a degree.

One of the first studies in which the distributional approach was employed was Brun et al. (2007). In the research, they took into consideration location name and company name metonymies and dealt with them similarly to a Named Entity Recognition (NER) task, which involves the recognition and extraction of name entities with the purpose of information mining. Brun et al.(2007) applied the NER task to metonymy resolution but, in order to do so, they had to adapt it to figurative language. They hybridize the methodology by unifying a syntactic analysis of the sentences and the distributional approach. More specifically, they implemented a deep parsing analysis to the metonymic sentences to extract syntactic information, which roughly corresponded to the agent-experiencer roles. Then, through a corpus study, they identified those instances that presented irregularities according to the agent-experiencer role, i.e. the metonymic expressions, and, on the basis of the observation of those instances, drew inferences such as "if a location name is the subject of a verb referring to an economic action, like import, provide, refund, repay, etc., then it is a place-for-people" (Brun et al., 2007, pp.489). Thus, the parser was adapted with the encoding of said additional information. Also the distributional-approach-based model they employed was implemented with the syntactic information just retrieved: to do so, they instructed the model to take two lexical units at a time and add the syntactic information that unifies the two units in order to form a triple, such the following for the sentence "provide Albania with food aid":

- OBJ-N('VERB:provide','NOUN: Albania').
- PREP_WITH('VERB: provide ','NOUN:aid').
- PREP_WITH('VERB: provide ','NP:food aid').

Then, the context was created for each lexical item, as for instance the following:

Words:	Contexts:
VERB:provide	1.VERB: provide. OBJ-N
NOUN:Albania	1.VERB: provide.PREP_WITH
NOUN:aid	2.NOUN: Albania.OBJ-N

NP:food aid	2.NOUN: aid. PREP_WITH
	2.NP: food aid. PREP_WITH
	1.VERB:provide.OBJ-N+2.NOUN:aid. PREP_WITH
	1.VERB:provide.OBJ-N+2.NP:food aid. PREP_WITH
	1.VERB:provide.PREP_WITH +2.NO:Albania.OBJ-N

For each entry in the dataset, it was then created a sub-context, a list of contexts in which the lexical unit appears, and a sub-dimension, a list of lexical units which co-occur at least once with a given context from the sub-context list. For instance, "Albania" is retrieved in 384 different contexts and 54,183 lexical units are occurring with at least one of the contexts from the sub-contexts list. On the basis of these data, the closeness of the context "VERB:provide.OBJ-N" is computed, thus obtaining the attribution of the nation name "Albania" to a place-for-people metonymy. The performance of the model created on the basis of the above presented method on the train set was adequate, however, it worsened on the test set. Therefore, this study was certainly defective, but it is still worth to be mentioned since it suggested a new combined approach to solve the problem of metonymic interpretation.

The distributional approach was also adopted by Lenci (2011) to analyse logical metonymies. He based his theory on Elman's assumption (2009) which states that the semantically preferred filler of an argument of a verb is determined not only by the verb itself but by the other preferred fillers of the verb as well. Based on this statement, Lenci elaborated the Expectation Composition and Update (ECU) model, which was able to separately compute the semantic expectations determined by the subject and by the verb and afterwards combine them to predict the "updated expectation". To test his theory, he formulated an experiment to assess the performance: the model was evaluated on how well it calculated the thematic fit between an agent-verb pair and a patient-noun

argument of the same verb. The results were satisfactory and therefore the ECU model constituted a potentially interesting method to interpret metonymy.

However, the performance of Lenci (2011)'s model or of any other distributional model had not been compared to the other approaches. Therefore, to assess the validity of this latter approach, Zarcone et al. (2012) proposed a new research paper that would establish which method is better than the other between the two previously mentioned approaches. In contrast to the previous studies that have employed English as the language of research, this study deals with the German language; however, the task resembles the preceding analyses, in that it aims at evaluating the possible readings of logical metonymies. More specifically, Zarcone, Utt and Padó decided on two alternative meanings for each metonymic sentence and asked the model to determine which occurrences were the high thematic fit and which occurrences were the low thematic fit. They chose to compare the probabilistic model suggested by Lapata and Lascarides (2003) and the distributional model proposed by Lenci (2011). Nevertheless, the researchers had to modify the ECU model to be adapted to deal with logical metonymy: they shifted the focus from the object to the covert event. They conducted the experiment and the results showed that the similarity-based model's performance was more convincing because of the better coverage and the higher accuracy compared to the probabilistic account.

After having proven that the distributional approach was a method worth considering, Utt, Lenci, Padò and Zarcone (2013) have taken it as a starting point to further investigate new processes to identify metonymic occurrences. In their study, they investigated whether the measure they called "eventhood" could help to infer whether a verb is a metonymic instance or not. The feature of "eventhood" determines whether a verb selects more likely an event-denoting object or an entity-denoting object. In order to determine the nature of the object-noun relations, the researchers have used WordNet (Fellbaum, 2010), a large database which has encoded semantic relations among English words, to decide which nouns had an event sense. The results showed that, even though there are

exceptions, a significantly higher eventhood score corresponded to metonymic verbs and that the higher the eventhood of a verb is the less likely said verb is to select an entity-denoting object. Therefore, on the basis of these findings, the research team of Utt, Lenci, Padò and Zarcone (2013) concluded that "eventhood" is a feature that can often indicate the occurrence of a metonymic verb and can also help to distinguish different types of metonymies.

3.3 Metonymy Resolution in the Deep Learning era

Neural networks are a fundamental method of machine learning developed to mimick the cognitive processes that humans perform on a daily basis and, therefore, allowing machines to reproduce the same mechanisms that take place in the brain (Guresen & Kayakutlu, 2011). Through machine learning algorithms, computers are able to train on large datasets, identify patterns and then, reproduce those patterns in order to produce results to solve the most diverse tasks (Ray, 2019). The concept of neural networks was created by Warren McCulloch and Walter Pitts (1943) and it represented a major area of research for both computer and neuro-scientists for more than two decades. Since then, neural networks have been going in and out of fashion through the years, but it is not until the 2000s that neural networks entered the world of language modeling. Previously, other strategies to model languages had been used, such as the models cited, but starting to apply neural networks to linguistic tasks was the turning point in the process of making it possible to machine to deal with natural language in a more accurate but at the same time semi-autonomous manner.

The first research published using the neural network technique was by Bengio et al. (2003). In their study, they created the first neural language model, which was a feed-forward model constituted of only three layers: an input layer, a hidden layer, and an output layer. Even though the number of hidden layers has been progressively increased in the following years, the three layers of Bengio et al. (2003)'s network are still the building blocks of any neural model. However, the

layers are not the only element to build a neural network: other than the input data and the corresponding targets, i.e. the output label, there are two major aspects, namely the loss function and the optimizer. The loss function serves to compare the prediction as generated by the network to the targets and then to compute a measure to define how accurately the model is working; the optimizer, instead, takes the loss value generated by the loss function and, based on that measure, updates the weight, thus influencing how the model learns.

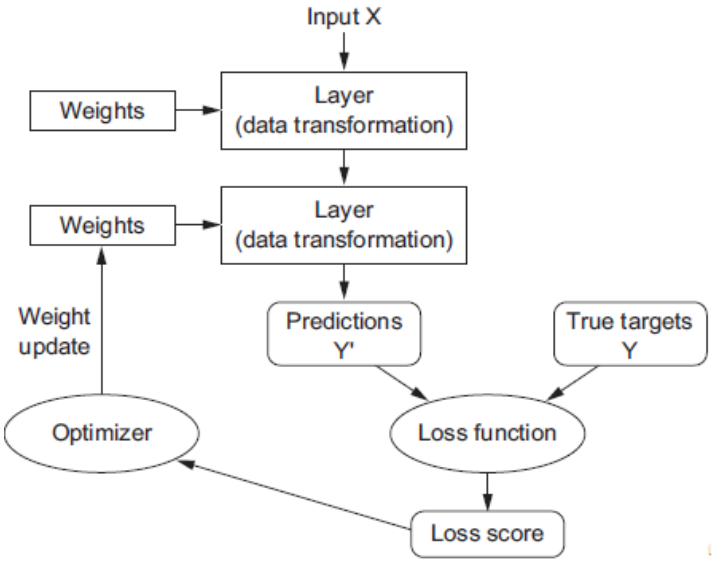


Figure 5 - Representation of a neural network. Source: Chollet, F. (2021)

Taking this neural network as a starting point, other models have been created on top of it: such kind of neural network has been modified and models like recurrent neural networks (RNN) (Mikolov et al., 2010) and long-short term memory networks (LSTM) (Graves, 2013).

In particular, given their ability to store information such as word sequence, LSTMs prove to be a useful tool for analysing language sequences and, therefore, metonymic occurrences. Hence, neural networks and, more specifically, LSTMs started to be employed in order to investigate metonymic sentences. This shift in the approach had one main consequence: after the distributional approach which treated metonymy resolution as an interpretative task where the model was asked

to predict the “missing element” of a metonymic expression, with neural networks metonymy resolution began to be treated again like a classification task. For instance, Gritta et al. (2017) intended the resolution of location metonymies as a classification task, i.e. identifying whether a location name was intended literally as the name of a place or figuratively as, for example, the people who find themselves in that place. The main contributions of this study are a new dataset created purposely for location metonymies (ReLocaR) and a novel predicate window (PreWin) method. This method is particularly interesting because, instead of considering the whole sentence as the informative source to classify metonymy, the PreWin allows to select a smaller and more relevant section of the sentence, granting more accurate results. By combining the PreWin method with dependency parsing, Gritta et al. (2017) produced convincing results, proving that at that state-of-the-art knowledge a minimalistic neural approach worked best for location metonymy resolution.

3.4 Transformer Language Models

A substantial turning point is represented by the appearance of a new type of pre-trained language models, namely transformers such as BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019). What differentiates these models from the rest is the mechanism of self-attention “heads”, which serves to compute a weighted representation of the token, consisting of a key, a value, and query vectors for each input token. Each layer of this type of neural network is composed of several heads and has to combine the outputs of said heads (Rogers et al., 2021). The main advantage of the self-attention heads is that they have a more in-depth understanding of the dependencies and influences between the words, in comparison with the previous models. As a matter of fact, the weighted representation allows distinguishing the most relevant words in long and complex sentences: this was the main problem for LSTMs, which were trained to take into equal consideration all words, but thus they would start forgetting information

rather quickly. The ability of transformers through attention heads to memorise the most salient words allows the model to keep track of relevant information and therefore to refer to words that appear in previous contexts.

As previously mentioned, two examples of such transformer models are BERT and RoBERTa. BERT, which stands for Bidirectional Encoder Representations from Transformers, is a model elaborated by Google AI Language in 2018 and it is able to process unlabelled text, extract information from the data based on the left and right context of the analysed token (Devlin et al., 2018). BERT was developed in two steps: the first step was the semi-supervised training on large amount of texts, such as books and Wikipedia, and it was trained on the basis of an unmasking task and a next sentence prediction task; the second step consist of the fine-tuning of the model for a particular purpose and require a supervised training on labeled data.

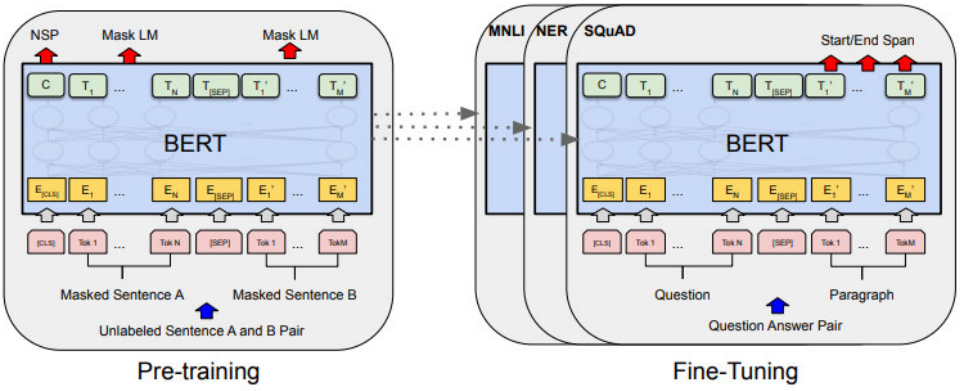


Figure 6 - BERT development: training and fine-tuning. Source: Devlin et al.(2018)

On the other hand, RoBERTa is architecturally the same model as BERT; however, BERT was undertrained according to Liu et al. (2019) and therefore they retrained the model modifying some aspects of the training process, namely the training procedure was made longer over additional data, the next sentence prediction task was removed, the training sentences were significantly longer, and it involved

dynamic masking instead of static. With this process, they created a new better performing model, that they called RoBERTa (Robustly optimized BERT pre-training Approach).

BERT, RoBERTa and the other variants of this type of transformers received immediate attention and recognition from the NLP community because of their impressive performances. The aim of most studies related to transformers is to understand how they manage to deal with language so successfully. To do so, several researchers tried to comprehend what kind of information BERT is able to process and then reproduce. As reported by Rogers et al. (2021), BERT is able to learn and represent information similar to syntactic tree structures, parts of speech, syntactic chunks, and roles. Nonetheless, even though it has some knowledge about semantic roles, BERT seems to struggle with other semantic-related tasks, like for example name entity replacement (Balasubramanian et al., 2020). Furthermore, also gaining and applying world knowledge does not seem an easy undertaking for this type of transformers, making any type of task where any kind of pragmatic knowledge is required rather difficult.

As discussed at the beginning of this thesis, metonymy interpretation is a multi-faceted task: in order to correctly infer the referents of a metonymic expression, knowledge about the syntactic structure of the sentence, the semantic information and, lastly, world knowledge are all required. Given the just discussed difficulties BERT has to apply all these skills at the same time, it is indeed interesting to evaluate whether this transformer model can perform on a metonymy resolution task as well as it does on literal language. The research proposed so far goes in two directions, either comparing the performance among transformers and with the previous theories or dealing with metonymy resolution as a classification task.

A work that goes in the direction of testing the performance of these newly created models is the research by Rambelli et al. (2020). In their study, they compared the performance of probabilistic, distributional, and transformer-based

models on the task of logical metonymy interpretation. Specifically, Rambelli et al. (2020) considered the probabilistic model by Lapata and Lascarides (2003) for the probabilistic account, the Distributional Semantic Models (DSMs) by Zarcone et al. (2012) and the Structured Distributional Model (SDM) by Chersoni et al. (2019) for the distributional approach, and lastly they chose BERT, RoBERTa, XLNet, and GPT-2 as the transformers. The task said models were required to fulfil was covert event recovery, meaning that they had to guess which was the implied verb in the sentence. The alternative verbs returned by the models were then judged on the basis of a plausibility score determined by a <subject, metonymic verb, object> triple and this rating was compared to human judgements in order to calculate the accuracy of the performance. The findings showed that the model that best performed on metonymy resolution was the SDM by Chersoni et al. (2019); however, Rambelli et al. (2020) argued that further research may be required to attest the performance of the transformers.

On the other hand, considering the strategy of testing models on a classification task, the study conducted by Li et al. (2020) is worth to be mentioned. Their aim was to conduct the first classification task experiment on location metonymies without the bias of external information, like the knowledge derived from taggers, parsers, or annotated dictionaries. The main contribution of this study is the inclusion of an interesting feature, namely a technique called target word masking: using six datasets, they substituted the target metonymic word with an X token, the masking, and let the models, BERT-base and BERT-large, guess whether the masked token was a metonymic or literal expression based only on the context surrounding said word. They compared the performances of the regular models, the models with data augmentation, and with the masking technique. The results proved that target word masking significantly improved the accuracy with which the models were able to predict metonymies.

Another research that has employed the technique of target word masking, even though with a different aim, is by Pedinotti and Lenci (2020). Their work is worth mentioning because they manage to combine both previously mentioned

approaches into one study. Their aim was to assess whether BERT has the ability to capture the meaning shift that occurs when a literal expression is replaced with a metonymic expression; moreover, they also wanted to compare BERT-base's performance with a model inspired by the SDM by Chersoni et al. (2019). This second model was based on the Generalized Event Knowledge (GEK) theory, which is a theory formulated by McRae and Matsuki (2009), and states that the way a sentence is processed gets influenced by expectations about the upcoming information determined by conceptual knowledge. In order to test the models, they subdivided the experiment into two parts. In the first part, they investigate the comprehension of the meaning shift, meaning that for each dataset entry, they calculated the contextual embeddings of the target word when used literally and when used metonymically, and then they compared both results with the contextual embedding of the target referent in the metonymic sentence. In this first step of the experiment, Pedinotti and Lenci (2020) also employed the masking strategy to observe whether BERT would predict the metonymic expression or the intended referent of said metonymy as more likely. In the second experiment, instead, the researchers wanted to assess whether the models would be able to associate the target word with the corresponding sense. To do so, they further divided the task into two parts: Metonymic Matching and Literal Matching. Metonymic Matching was correctly performed if the metonymic expression was more similar to the metonymic paraphrase than the literal expression to the same metonymic paraphrase, while Literal Matching was confirmed in the case of the literal paraphrase being more similar to the literal expression than to the metonymic expression. Also in the second experiment, masking was included in a similar way as in the first experiment: they observed whether the model would more likely predict the metonymic interpretation or the literal interpretation. On the basis of their findings, Pedinotti and Lenci (2020) argued that BERT struggled in the interpretation of the meaning shift that occurs in the case of metonymy, however, BERT's performance was superior to SDM when dealing with matching and choosing between two possible interpretations.

IV. The Project

After having introduced what the concept of metonymy and seen how it has been dealt with from different research fields, the aim of this thesis is suggesting a new analysis of how such figure of speech is handled by recent technologies, such as transformer language models.

4.1 The background

Transformers are receiving a lot of attention at the present moment given their impressive performances when dealing with language (Zhang et al., 2023; Illina & Fohr, 2023; Kadan et al., 2023). They are indeed a powerful tool that could be employed in different tasks, from natural language generation (NLG) (Nguyen & Tran, 2023) to information extraction (Li et al., 2023) to many others. Their main advantage is that transformers are able to operate on unlabelled data, but still acquire relevant information. This type of language model has proved to generally deal quite well with literal language; however, they showed to encounter some trouble with figurative language. Furthermore, the main issue with transformers is that researchers still have to figure out how they exactly work. Rogers et al. (2021) discussed in their paper which aspects of human language BERT is able to process. As they have argued on the basis of previous research (Wu et al., 2019; Hewitt and Manning, 2019), BERT seems to have some syntactic understanding, but it lacks as far as semantics and pragmatics are concerned. In Chapter 2.1-2 of this thesis, it was mentioned that, even though metonymy resolution seems an automatic task for humans to perform in everyday communication, in reality semantic skills and world knowledge indeed play a fundamental role in the correct interpretation of such figure of speech, as the psycholinguistic experiments have shown. On the basis of said considerations, the evaluation of what kind of

information transformer models are able to process when dealing with metonymy resolution is still a matter of discussion.

As mentioned in the previous chapter, several studies on metonymy resolution have been conducted in order to try to answer the question of whether a transformer language model such as BERT could distinguish the literal use from the metonymic use of an expression. However, to the best of my knowledge, no study has so far been proposed to further investigate what transformers can infer about the interpretation of referential metonymy, i.e. they have never been asked to return possible referent interpretations of a metonymic occurrence. Therefore, the first experiment of this thesis aims at exploring if and how well some transformers, namely BERT (base and large) and RoBERTa (base and large), can predict the intended referents of a corpus of metonymic sentences. To do so, said transformers were asked to produce words that could replace the metonymic expression, and the accuracy was judged both on the basis of hypernym-hyponym relations as encoded on WordNet.

Furthermore, a second experiment is proposed in order to analyse the performance of the models at the different stages and more specifically to understand why the accuracy of the answers the models return varies. The model that will be analyse is RoBERTa large since it was the transformer which performed slightly better than the others in the first part of the project. To investigate the trend of the performance of RoBERTa large, the contextual embeddings of target words will be considered and compared. The comparison will be between three different instances of sentences: metonymic sentences, literal sentences, and sentences with a metonymic paraphrase. By comparing the contextual embeddings of the target words contained in these sentences it will be possible to compute the cosine similarity at each layer of the chosen model.

4.2 Experiment 1

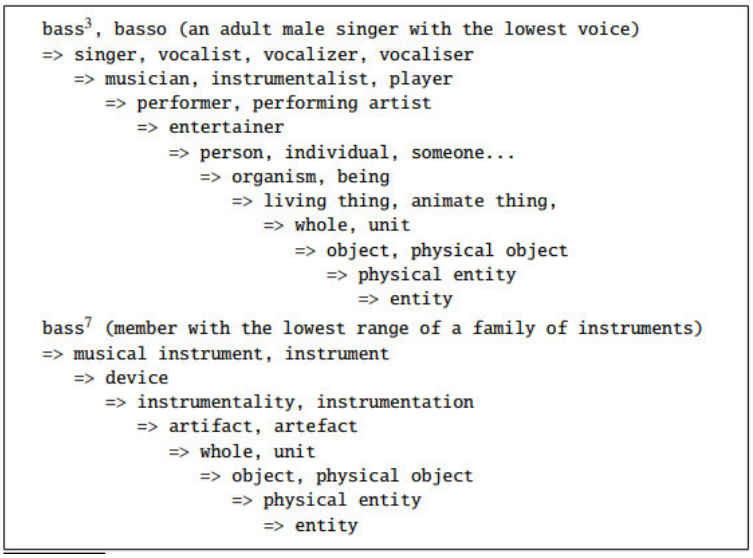
4.2.1 Models, tools, and dataset used in the study

In my work I considered a total of four models: BERT base, BERT large, RoBERTa base, and RoBERTa large. Both BERT and RoBERTa are models built on the base of stacked layers of encoders and the difference between the two versions of each model is mainly in the different number of layers: the base version of both consists of 12 layers, while the large version of 24 layers[2]. A higher number of layers corresponds to a larger quantity of parameters and, therefore, large versions have more parameters than the base versions. However, as far as parameters are concerned, BERT and RoBERTa differ: BERT base has 110 million parameters, while RoBERTa base has 125 million parameters, and BERT large has 340 million parameters, while RoBERTa large has 355 parameters. For the purpose of this thesis, the above-mentioned transformers were retrieved from the open source library Transformers (Wolf et al., 2020) on the Hugging Face website.

Other than the models, another essential tool was used in the experiment, namely WordNet. WordNet is a large electronic lexical database for English (Fellbaum, 1998), which was created to represent concepts are connected to one another in the human brain. WordNet consists of three separate databases: the first database is dedicated to nouns, the second is for verb entries, and in the third adjectives and adverbs are encoded. For the purpose of this thesis, we will be considering and working on only the database dedicated to nouns. Each entry in WordNet represents a sense which might encode a list of lemmas, i.e. a synset, and is defined by a gloss and some examples of sentences in which said lemmas may occur. However, the same lemma could appear in multiple synsets: this is due to the lexical ambiguity that may occur, as for instance in the case of polysemous words. To explain how WordNet was built, two theories must be considered: componential analysis, which states that more generic concepts are “contained”

[2] https://huggingface.co/transformers/v2.5.1/pretrained_models.html

into specific concepts, and relational semantics, which just relates words without highlighting any particular relation between them. As a matter of fact, these theories have determined the basic structure of this lexical database: the main relations in WordNet are of two types: either IS-A-KIND-OF/IS-A-PART-OF or IS-AN-ANTONYM-OF/ENTAILS, and their opposites. An example of how a lemma is encoded is given with the following image.



On the basis of these assumptions, working on the two types of relations is the method to navigate WordNet: the first approach can be used to investigate the hypernym-hyponym relations to understand if a sense is contained in another sense, while the second approach allows us to investigate which words share a similarity or are instead in opposition.

The last relevant element used in the experiment was the dataset, which was created by Pedinotti and Lenci (2020), and was retrieved from the GitHub website of the project[3]. It was chosen because it is a rather large example of a corpus of 509 referential metonymies, divided into six types of metonymies: CONTAINER-

[3] https://github.com/ppedin/MetonymyData

FOR-CONTENT, PRODUCER-FOR-PRODUCT, PRODUCT-FOR-PRODUCER, LOCATION-FOR-LOCATED, CAUSER-FOR-RESULT, POSSESSED-FOR-POSSESSOR. An important aspect to be highlighted is that this corpus is not equally split into the six categories: the following table reports how many entries there are for each type.

Type of metonymy	Number of entries for each type
CONTAINER-FOR-CONTENT	89
PRODUCER-FOR-PRODUCT	110
PRODUCT-FOR-PRODUCER	47
LOCATION-FOR-LOCATED	94
CAUSER-FOR-RESULT	92
POSSESSED-FOR-POSSESSOR	77

Even though some categories contain about the same number of entries, it should be noted that PRODUCER-FOR-PRODUCT is a remarkably larger category, while PRODUCT-FOR-PRODUCER is the type of metonymy with significantly fewer examples. Noting these differences is relevant especially when evaluating the performances of the models because they could be an undesired affecting variable. For example, even though it may not be the case, there could be a category of metonymy which is more “novel” than others, the model could therefore struggle to correctly interpret it and, if this category is larger than the others, it could have a greater effect on the overall accuracy of the performance of the models. However, other than this downside, it is a carefully drafted dataset: it contains for each entry four sentences to illustrate how similar words can have different senses according to the contexts in which they occur. In fact, a column is dedicated to “Metonymic Sentences”, the sentences in which the metonymic expression occurs; a second column contains “Literal Sentences”, the sentences in

which the same word as in the previous case but this time in its literal meaning is included; a third column lists “Sentences with Metonymic Paraphrase”, in which a plausible referent of the metonymic sentence is inserted in a sentence; lastly, a fourth column consists of “Sentence with Literal Paraphrase”, which are very similar to the literal sentence, but the target expressions refer to a possible paraphrase of the intended meaning of the metonymic phrase and they are used literally. An instance of an entry in the dataset could be as follows:

- “Metonymic sentence”: the pot bubbled on the fire
- “Literal Sentence”: the pot reflected the light
- “Sentence with Metonymic Paraphrase”: the water cools to 25.0 degrees
- “Sentence with Literal Paraphrase”: the container showed signs of damage

The association of the metonymic and literal uses of the same word or with a plausible referent is useful in order to investigate how the models are performing while solving the task through the comparison of the desired target words and what kinds of embeddings were obtained from the transformer.

4.2.2 Methodology

As previously explained, the aim of this study was to understand whether transformers such as BERT and RoBERTa are able to trace back a metonymic expression to the intended referent or set of referents. Therefore, the first step of the experiment was to ask these models to provide five alternative words, which should be equal in meaning to the metonymic expression. In order to conduct this part of the experiment, the models were asked to interpret the metonymic phrase through the masking technique, which was used alongside a prompt. Practically, to each entry of the dataset, a second sentence was manually added, which was the same for every occurrence of the dataset. This second sentence is defined as the prompt, and for BERT base and large it was realized as follows:

*Sentence with metonymic expression. + "Therefore, the *metonymic expression of the previous sentence * is a type of [MASK]"*

For RoBERTa base and large, the prompt used was the same, only the masking was changed to:

*Sentence with metonymic expression. + "Therefore, the *metonymic expression of the previous sentence* is a type of <mask>"*

This was just due to a difference of the models in how the mask is signalled to them as a clue that indicates that that element should be substituted with a word. Other than this discrepancy, the mechanism was the same: the models were asked to substitute the mask with five alternative words on the basis of the precedent context, which was automatically processed; after the sentence with the mask had been solved, with a for-loop it was substituted from the next sentence in line until all the dataset was processed. The five results for each sentence were stored in a CSV file for each model and split into five columns, which correspond to the order in which the answers were returned so that the solutions provided as firsts were included in the first column, the solutions provided as second in the second column and so on, until the fifth and last solution and column. The results were then further split according to the type of metonymy the phrase would classify as, so six CSV files were created for each model to facilitate the evaluation of said results.

4.2.3 Evaluation of the interpretations: three strategies for a more comprehensive analysis

The first evaluation process dealt with the accuracy of the interpretations produced by the four models. The chosen tool to establish which solutions were correct and which were false was WordNet and, more specifically, the first step to conduct the evaluation was analysing the hypernym-hyponym relations as encoded in WordNet. In the dataset from Pedinotti and Lenci, a column was

dedicated to the list of the possible desired referents, one for each metonymic sentence. However, only one possible referent was given for each sentence. Thus, in order to generalize this list and create some sort of semantic space that would determine the correctness of the answers, the referent words from the dataset were taken and, with the aim of finding synsets that could contain as many words from the target referent list of the dataset as possible, two or three hypernyms for each category of metonymy have been searched on WordNet. The idea was that the accuracy of BERT's and RoBERTa's answers would be judged positively if the interpretations the models provided were found in the categories, i.e. the synsets, selected.

It is important to mention that, even though most of the target words were included in the selected hypernyms, there were some exceptions which were not included in the hypernym selection. To name some, in the CONTAINER-FOR-CONTENT type of metonymy the target word "book" and "merchandise" were excluded, and no hypernym was chosen for them. The reason was that both these target referents occur respectively only once in this category of metonymy and, given the fact that most metonymies belonging to this category refers to either food or drinks, selecting a hypernym that could have included both such as "artefact" impacted in an undesired way the count for accuracy. In fact, it was observed that some answers produced by the models for CONTAINER-FOR-CONTENT metonymies could have been found as hyponyms of "artefact", but it would not have been ideal because the answer would have been judged as accurate even though it was not. On this ground, I chose to exclude some less-often-occurring words if finding a hypernym that included them would have meant distorting the computation of accuracy.

Not only it was necessary to select two or three hypernyms for each category, but also the correct synset for each hypernym. All the synsets selected are noun synsets because all the metonymies in the dataset are referential metonymies. The number of the synset was chosen with the same criteria as the selection of the hypernyms, namely the synset that could contain the largest number of target

referents was preferred. The hypernym synsets selection for each category of metonymy ended up as follows:

METONYMY TYPE	SYNSET	SYNSET ID	LEMMAS OF THE SYNSET
CONTAINER-FOR-CONTENT	substance.n.07	{04941723}	{consistency, consistence, substance, body}
	food.n.02	{07571428}	{food, solid food}
PRODUCER-FOR-PRODUCT	artefact.n.01	{00022119}	{artifact, artefact}
	communication.n.02	{00033319}	{communication}
	event.n.01	{00029677}	{event}
PRODUCT-FOR-PRODUCER	person.n.01	{00007846}	{person, individual, someone, somebody, mortal, soul}
	group.n.01	{00031563}	{group, grouping}
LOCATION-FOR-LOCATED	person.n.01	{00007846}	{person, individual, someone, somebody, mortal, soul}
	group.n.01	{00031563}	{group, grouping}
CAUSER-FOR-RESULT	sound.n.04	{07385893}	{sound}
	communication.n.02’	{00033319}	{communication}

	sensation.n.01'	{05720023}	{sensation, esthesis, aesthesis, sense experience, sense impression, sense datum
POSSESSED-FOR-POSSESSOR	person.n.01	{00007846}	{person, individual, someone, somebody, mortal, soul}

As it might be noticed from the list, the synset selected is specified through the a main lemma determined by WordNet of a synset plus a "n" that indicates that the synset refers to a noun and a number which states to which synset we are referring to. This latter feature is especially useful in the case of multiple senses associated with a single lemma: by defining the synset number, we are selecting the specific sense we want to take into consideration.

Other than the synsets, it was necessary to decide on which method to adopt in order to check the presence of the BERT's and RoBERTa's answers in the hyponyms of the chosen hypernym synsets. Firstly, it was decided to use NLTK (Natural Language Toolkit)(Bird et al., 2009), which is a suite of open-source Python modules and it allowed to automatically retrieve the synsets used in the experiment. Secondly, three different strategies ended up being preferred in order to try to make the evaluation of the answers of the models as accurate as possible.

The first strategy suggested is perhaps the simplest one out of the three: a list for each category of metonymy containing all the hyponyms of the selected

hypernyms for said category was generated. Afterward, the interpretations provided by the four models were iterated in order to check whether they were included in the previously created list. Thus, to each word it was attributed either a probability of 0 if it was not found in the list and a probability of 1 in case it was found. To sum up the results, the accuracy (Manning, 2008) was calculated according to the following formula:

$$accuracy = \frac{number\ of\ correct\ predictions}{total\ number\ of\ predictions}$$

It is relevant to mention that the average accuracy of the solutions was calculated at k. This type of measurement of accuracy is often employed in information retrieval (Manning, 2008) and fits the purpose of this experiment, given the relevance of considering all the generated solutions up to a point (the k-number). So, firstly, the average accuracy of the first answers was computed, secondly the average accuracy of the second answers was calculated on the basis of the results of the first and second answers, and so on until the fifth solution, whose accuracy comprehended all the scores.

The overall accuracy of the models was calculated, as well as the accuracy for each category of metonymy in order to take into consideration any possible variation according to the type of metonymy. The following plots will show the performance of the models when dealing with metonymy resolution judges on the basis of the computation of the accuracy. Each plot represents the performance of all four models. On the y-axis the interval of the score is reported: the interval for all the plots was set to be between 0 and 1 because these are the values given to each solution. Ideally, if the models were performing perfectly or at least nearly perfectly the results would tend to be shifted towards the 1-score; contrarily, the worse the models perform the scores would tend more towards 0. On the x-axis,

instead, it is mentioned to which number of solution the averages refers, and therefore the numbers 1 to 5 are reported.

Figure 7 shows the behaviours of the four models according to the first strategy. As can be seen, the performances are far from excellent: on average, all four models tend to correctly guess the intended referent of a metonymic expression only about half of the times. Moreover, BERT base is the model which performs slightly worse, while the answers generated by RoBERTa large seems to be the most convincing out of the four models. However, given the proximity of all four lines at any point, it can be observed that the performance of each model is not drastically different from the performances of the other models. Only at the first solution, the divergence in the performance seems to be slightly increased, while at the last solutions, the scores almost overlap.

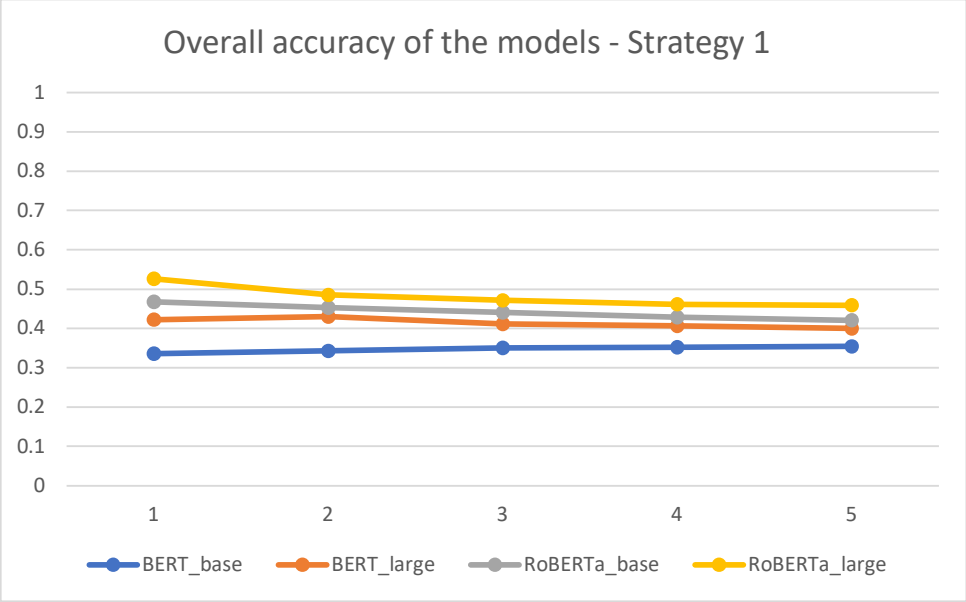


Figure 7 – Plot of the trend of the overall accuracy of the models. Each line represents the trend of the performance of a model according to average score on the y-axis and solution number on the x-axis.

As previously mentioned, it is relevant to apply this first evaluation strategy to the results divided into metonymic categories to observe whether the metonymic

type is an affecting variable in the performance of the models, i.e. whether there is any remarkable difference in how the models manage to interpret metonymies according to the metonymic type they belong to. To do so, the following plots were drawn to show the performances of BERT base, BERT large, RoBERTa base, and RoBERTa large according to the different categories of metonymic expressions.

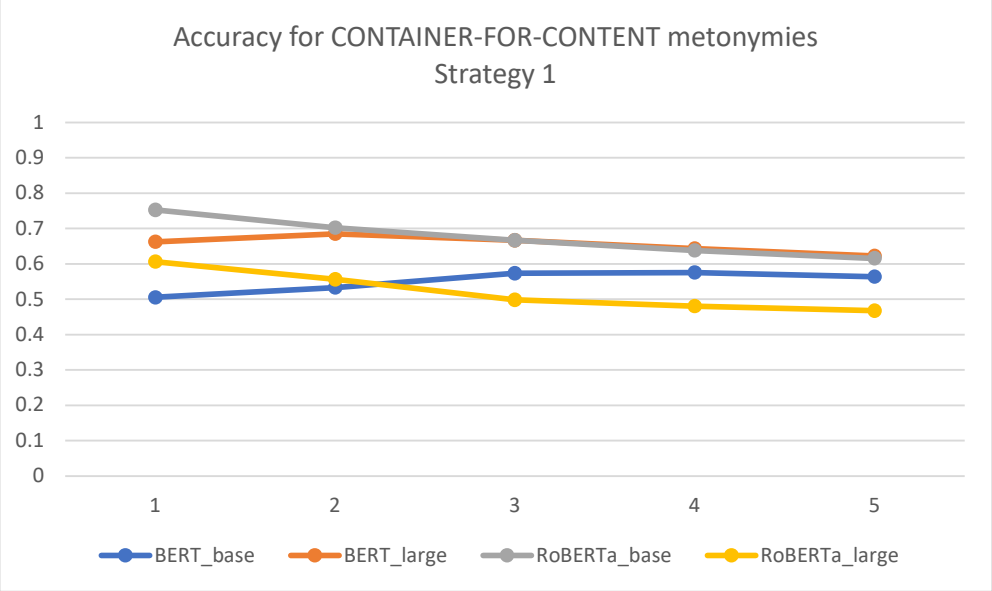


Figure 8 - Plot of the trend of accuracy for CONTAINER-FOR-CONTENT metonymies according to the first strategy of evaluation. Each line represents the trend of the performance of a model according to average score on the y-axis and solution number on the x-axis.

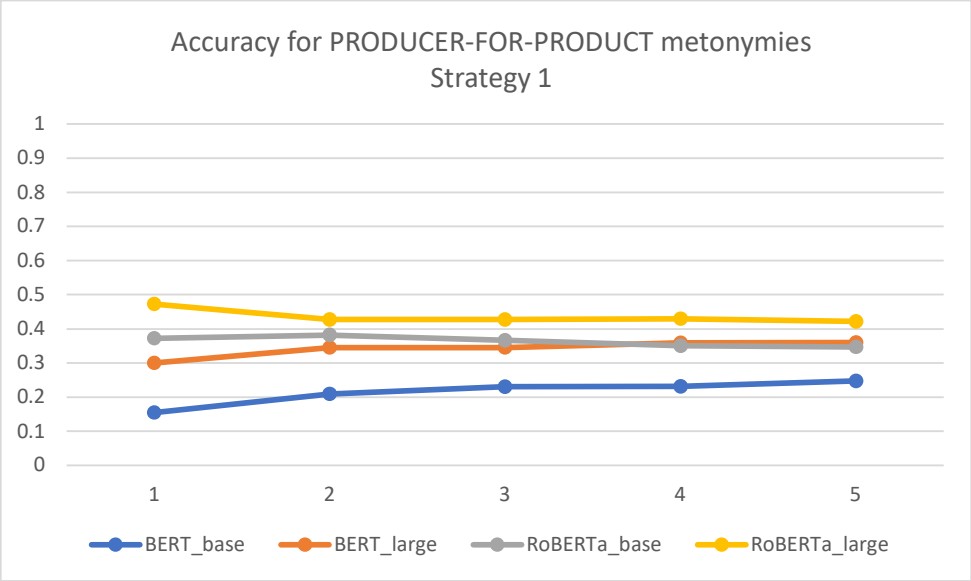


Figure 9 - Plot of the trend of accuracy for PRODUCT-FOR-PRODUCER metonymies according to the first strategy for evaluation. Each line represents the trend of the performance of a model according to average score on the y-axis and solution number on the x-axis.

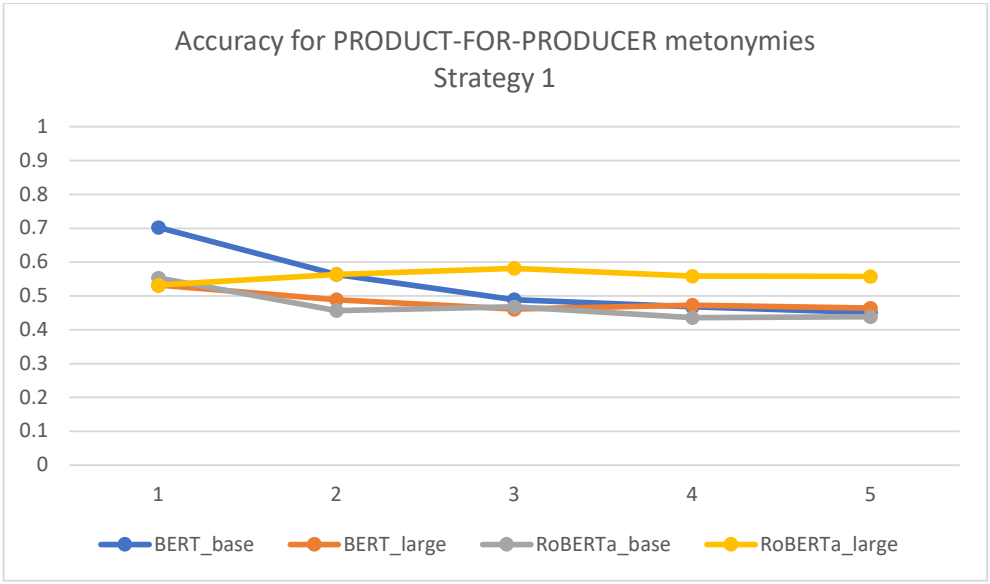


Figure 10 - Plot of the trend of accuracy for PRODUCER-FOR-PRODUCT metonymies according to the first strategy for evaluation. Each line represents the trend of the performance of a model according to average score on the y-axis and solution number on the x-axis.

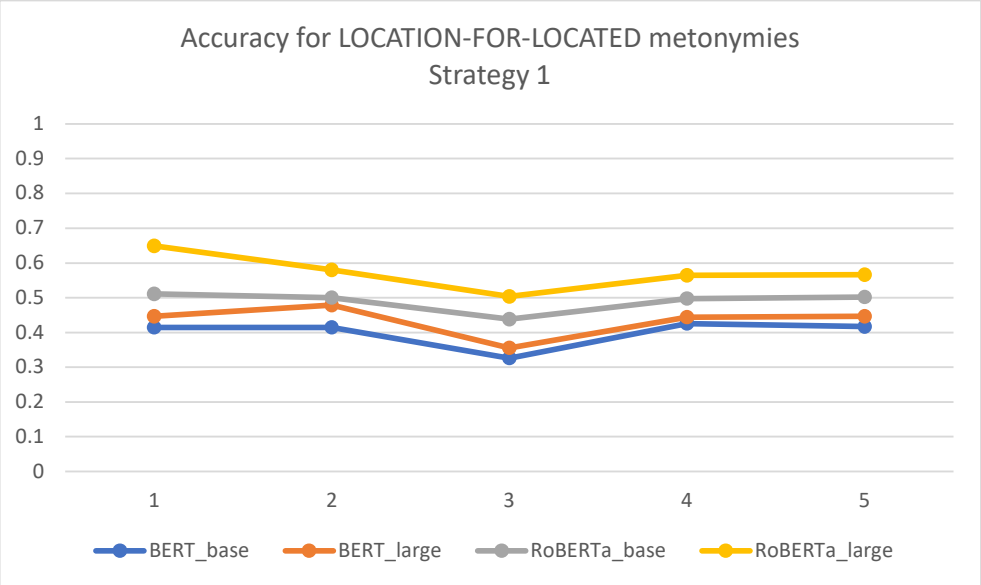


Figure 11 - Plot of the trend of accuracy for LOCATION-FOR-LOCATED metonymies according to the first strategy for evaluation. Each line represents the trend of the performance of a model according to average score on the y-axis and solution number on the x-axis.

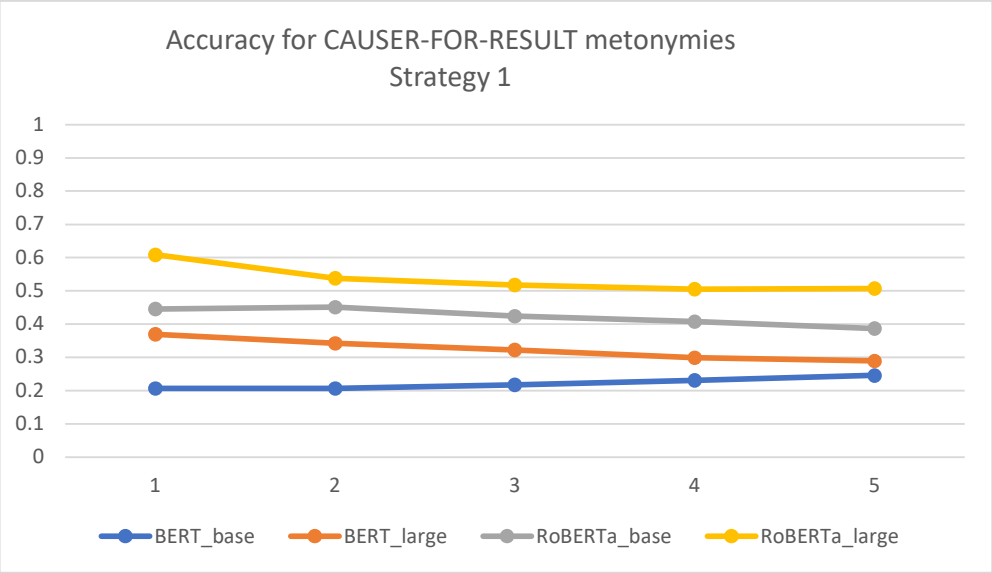


Figure 12 - Plot of the trend of accuracy for CAUSER-FOR-RESULT metonymies according to the first strategy for evaluation. Each line represents the trend of the performance of a model according to average score on the y-axis and solution number on the x-axis.

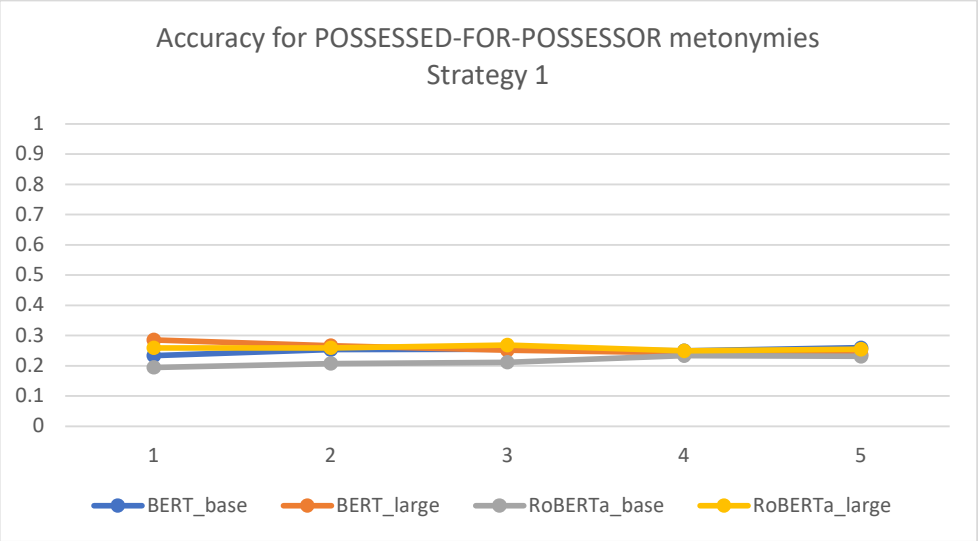


Figure 13 - Plot of the trend of accuracy for POSSESSED-FOR-POSSESSOR metonymies according to the first strategy for evaluation. Each line represents the trend of the performance of a model according to average score on the y-axis and solution number on the x-axis.

As can be seen from the plots above, there are indeed some categories that seem to be better processed and other categories that are dealt with way worse. An example of the previous is the category of CONTAINER-FOR-CONTENT metonymies: here the accuracy reaches almost the score of 0.8, which is not perfect but, since the models are dealing with figurative language, it is remarkable. Moreover, it is interesting to note that at the first solution the performances of the models in this same category do not reflect the accuracy of the overall performances: as a matter of fact, the model that seems to better resolve metonymy is RoBERTa base, followed by BERT large. RoBERTa large, which in the general evaluation was the best model for this task, comes only in third place at the first solution and even ends up last at the last solution.

On the other hand, the category in which the models seem to struggle more is the last one, POSSESSED-FOR-POSSESSOR metonymies, as shown in Figure 13. In this case, all models do not go beyond the 0.3 scores, leaning strongly toward the side which indicates a lack of understanding. A possible interpretation of this phenomenon could be that this type of metonymy is arguably the category that leaves the most room for a creative use of language: compared to the CONTAINER-

FOR-CONTENT type of metonymy which is more frequent and the relations between metonymic expression and intended referent are more fixed and probably more sedimented in language use, POSSESSED-FOR-POSSESSOR allows for more novel combination of concepts.

Another remarkable point to highlight is that in most categories of metonymy, the compared performances of the models do not show great variability. As already mentioned in the analysis of the overall accuracy, on average RoBERTa large seems to deal more accurately with metonymy resolution than the other transformers, but the divergence is hardly noticeable. However, if we take the plot for the accuracy of the solution for CAUSER-FOR-RESULT metonymies shown in Figure 12, the difference in the scores is greater than in all other instances of metonymy and, even though the difference is more substantial at the first solution, it is maintained at each number of solutions.

With some less relevant exceptions, the other categories are dealt with pretty much as expected from the overall evaluation of the models. However, there is one aspect that emerges from some of the lines of the previous plots and is somewhat unexpected: generally speaking, transformers are trained to return the answer that they find to be the most accurate as the first, and then their results start to get increasingly worse, but this is not always the case according to the analysis of the accuracy conducted in this thesis. If the performances of the models had followed this principle, the lines of all plots should have had a progressive downward trend with the progression of the number of solutions. As can be seen in the majority of the plots, this is hardly the case: the accuracy scores stay either more or less the same or it even happens that they increase over time. For instance, this aspect is most noticeable in the trend of BERT base performance when dealing with PRODUCER-FOR-PRODUCT metonymies: as it can be seen from Figure 9, even though the accuracy does not drastically improve, the enhancement is still remarkable.

Nonetheless, the first strategy to evaluate the performance of the models must be considered too forgiving at times. This is due to the fact it is enough for a word to be included in the hyponyms list for it to be judged as a completely correct match. However, this evaluation does not take into consideration the fact that a word may have multiple referents, or senses, and it may be the case that not all senses should be evaluated as correct answers from the models. Therefore, a second strategy should be formulated in order to correctly judged such cases. More specifically, we need to create a list of synsets associated to each lemma returned as a solution of the metonymic expression and a list, called semantic space, of the synsets included in the previously selected hypernyms, and then check which of the synsets of the lemmas returned by the model are included in the semantic space generate by the hypernyms. By doing so we take into account cases such as “glass”, which was the interpretation of the metonymic sentence “the man sips the glass”. This solution is judged with a match by the first strategy, therefore completely correct. This is due to the fact that, if we look for the word “glass” on WordNet, we would find that there is a synset encoded under the lemma “glass”, which represents the sense of “glass” as “an amphetamine derivative (trade name Methedrine) used in the form of a crystalline hydrochloride; used as a stimulant to the nervous system and as an appetite suppressant”. This particular sense, and therefore the lemma “glass”, is indeed considered an hyponym of “substance”. Thus, when applying the first strategy to check the inclusion of “glass” in the lemma list of the hypernym “substance”, “glass” would be found and judged as a correct solution. Even though “glass” intended as a drug could technically be a referent for the metonymic expression “glass”, intuitively it is an unlikely referent for the metonymic sentence “the man sips the glass”, since it is usually some kind of beverage that is sipped. Thus, the second strategy aims at highlighting this evidence: a solution should be considered only partially correct if not all the synset connected to the lemma returned by the model can be found in the semantic space of hypernyms. On this ground, the second strategy measures the accuracy of each solution by attributing a 0 or a 1

score to each synset of the lemma on the basis of whether it was included in the semantic space or not, and then dividing the sum of all the scores by the number of synsets connected to said lemma. Lastly, the average score for each solution number was computed like in the first strategy.

Figure 14 shows the overall accuracy of the four models, computed accordingly to this presented method.

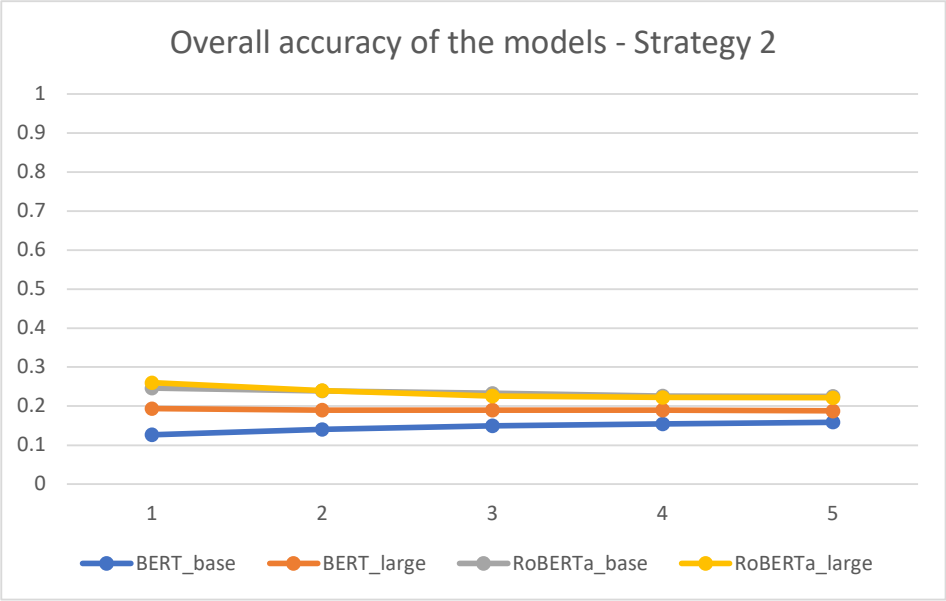


Figure 14 - Plot of the trend of the overall accuracy of the models according to the second strategy for evaluation. Each line represents the trend of the performance of a model according to average score on the y-axis and solution number on the x-axis.

As can be seen in Figure 14, applying this strategy meant drastically reducing the accuracy of the model because not all the senses of the previously positively evaluated answers were included in the semantic space. The strategy of investigating the synsets rather than the lemmas resulted in lower scores, which determined an overall lower accuracy for all models. It is relevant that the ranking of the performances of the models is unchanged in relation to the ranking defined by the first strategy: RoBERTa large is still the model that among the four performs best, followed by a practically equally performing RoBERTa base, then BERT large,

and, closing the ranking, BERT base. However, even RoBERTa large can hardly be defined as a great model for metonymy resolution: particularly with this strategy, it emerges that the accuracy is quite low, since it does not reach the 0.3 scores even with the most solid solutions.

As with the previous strategy, also in this second stage the performances of the models were considered according to one type of metonymic expression at a time. Figures 15 to 20 show the results of the split analysis.

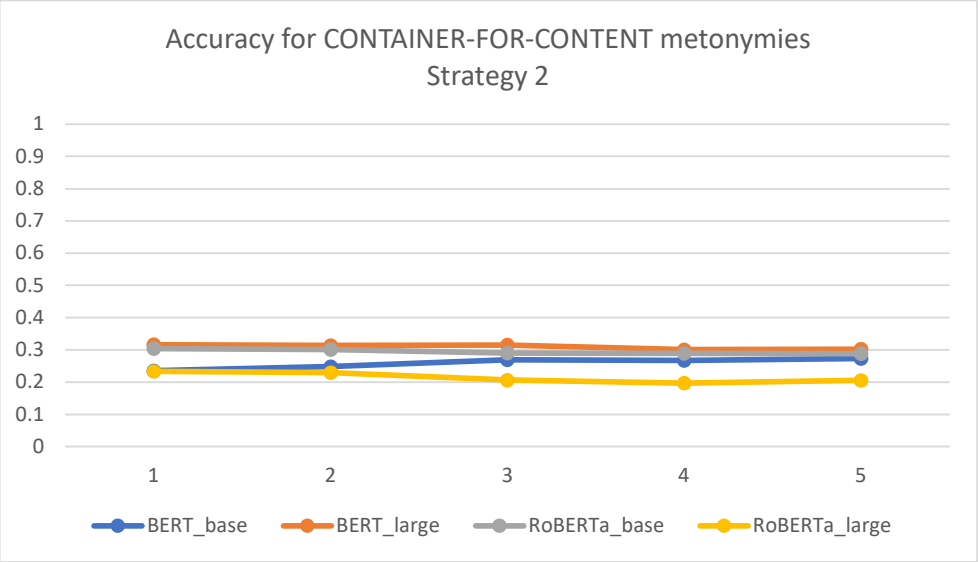


Figure 15 - Plot of the trend of accuracy for CONTAINER-FOR-CONTENT metonymies according to the second strategy for evaluation. Each line represents the trend of the performance of a model according to average score on the y-axis and solution number on the x-axis.

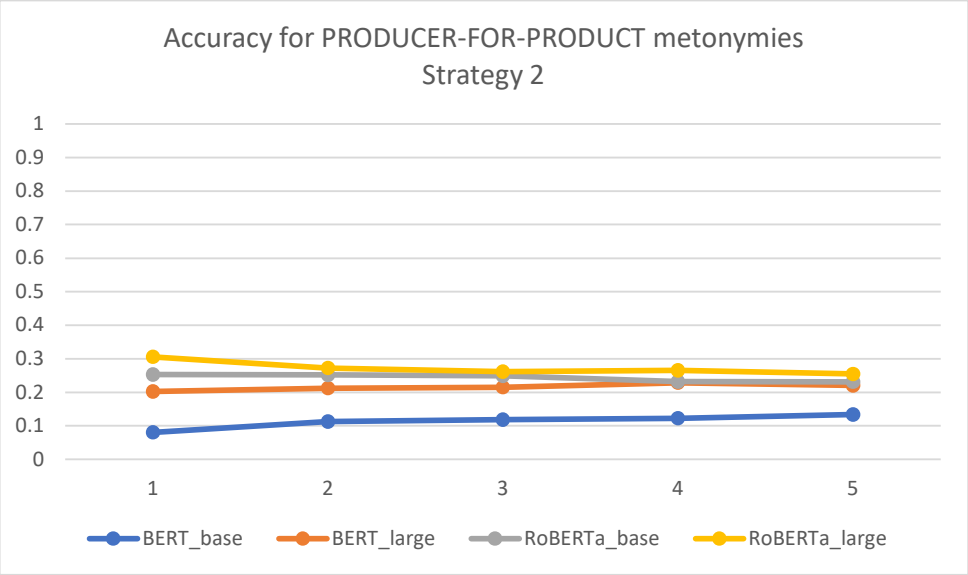


Figure 16 - Plot of the trend of accuracy for PRODUCER-FOR-PRODUCT metonymies according to the second strategy for evaluation. Each line represents the trend of the performance of a model according to average score on the y-axis and solution number on the x-axis.

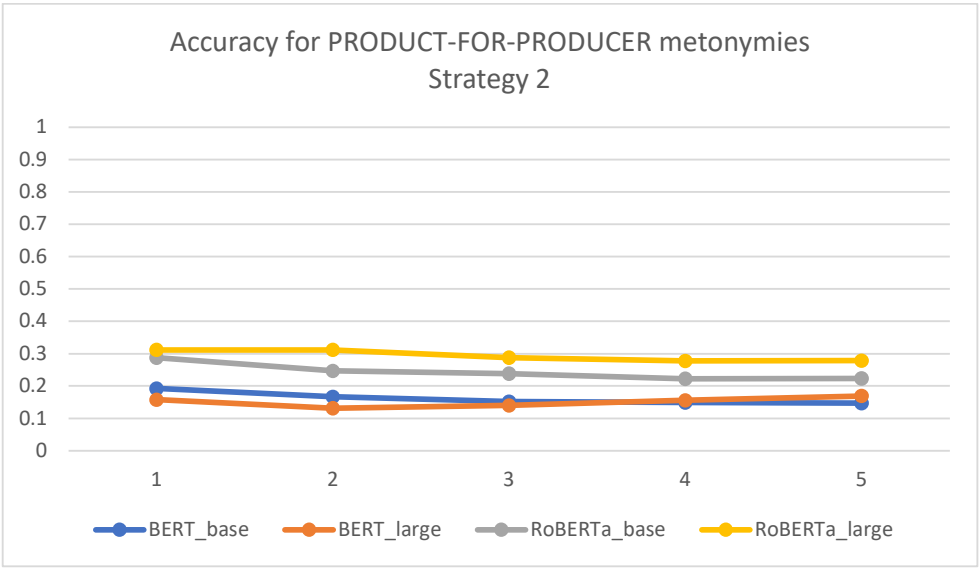


Figure 17 - Plot of the trend of accuracy for PRODUCT-FOR-PRODUCER metonymies according to the second strategy for evaluation. Each line represents the trend of the performance of a model according to average score on the y-axis and solution number on the x-axis.

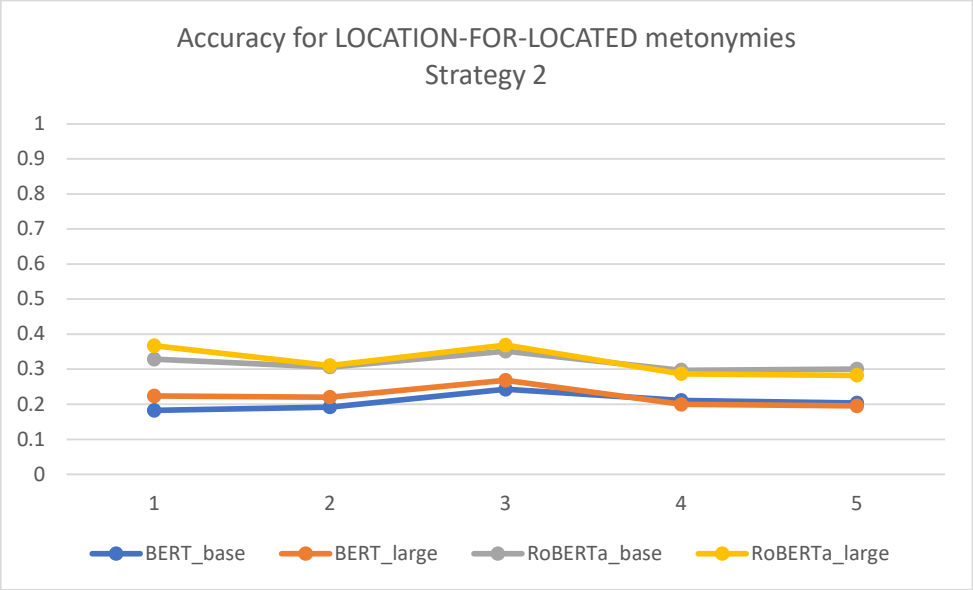


Figure 18 - Plot of the trend of accuracy for LOCATION-FOR-LOCATED metonymies according to the second strategy for evaluation. Each line represents the trend of the performance of a model according to average score on the y-axis and solution number on the x-axis.

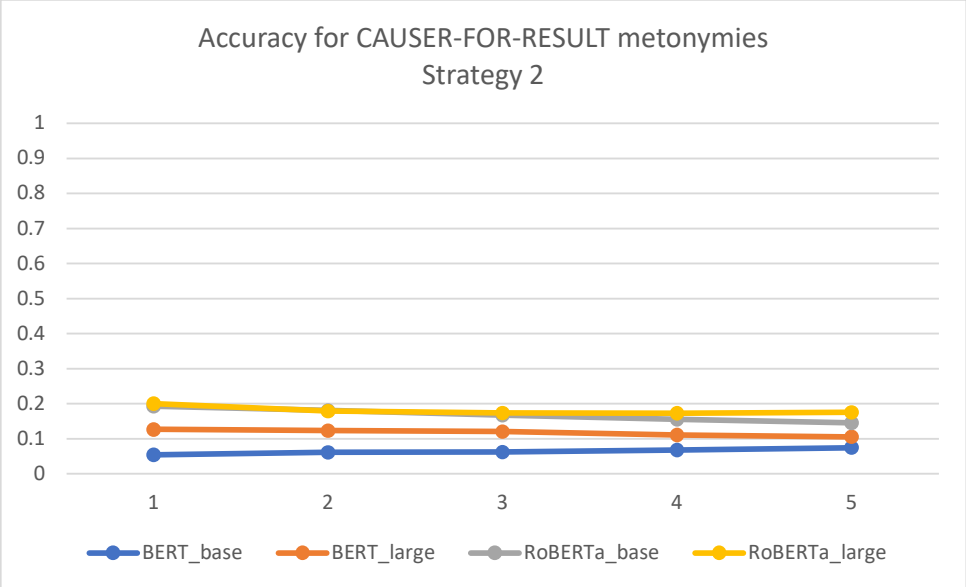


Figure 19 - Plot of the trend of accuracy for CAUSER-FOR-RESULT metonymies according to the second strategy for evaluation. Each line represents the trend of the performance of a model according to average score on the y-axis and solution number on the x-axis.

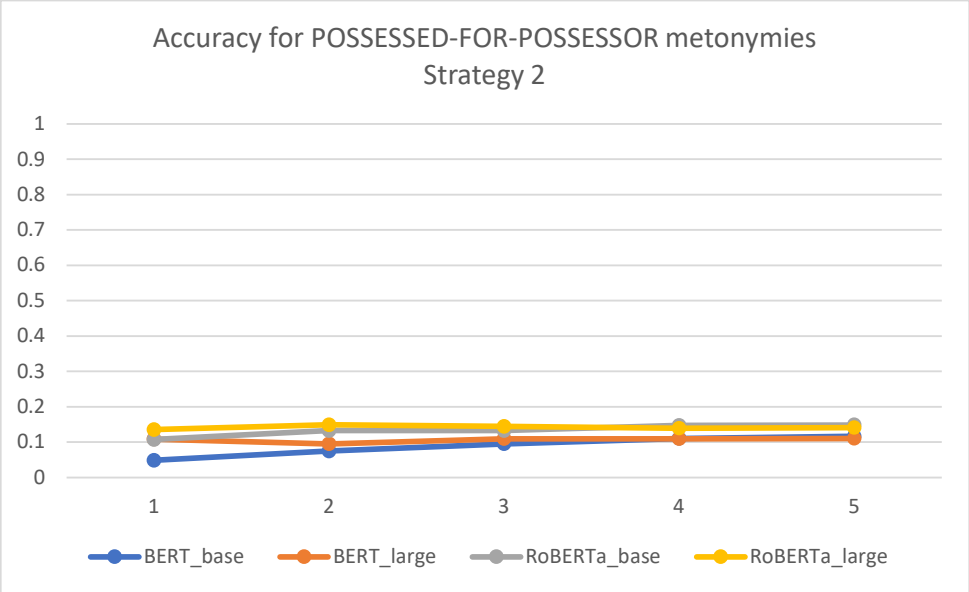


Figure 20 - Plot of the trend of accuracy for POSSESSED-FOR-POSSESSOR metonymies according to the second strategy for evaluation. Each line represents the trend of the performance of a model according to average score on the y-axis and solution number on the x-axis.

As shown by Figures 15 to 20, the variability is not as great as in the first strategy for the accuracy analysis. There are some types of metonymies that are better dealt with by the transformers, such as CONTAINER-FOR-CONTENT and LOCATION-FOR-LOCATED, but even in these instances the scores even at their "peak" are far from an ideal score, staying below 0.4.

Furthermore, even in this second analysis, the category of metonymic expression that seems to be the hardest for all models to process remains the POSSESSED-FOR-POSSESSOR type. In this case, as shown in Figure 20, the trend of the performances of the four models stays below the 0.2 scores at every solution number.

Lastly, it can be noticed that, contrary to the previous strategy, judging the answers with the second method results in quite stable performances, meaning that the anomaly of the increasing improvements highlighted before is not detected here. With few exceptions, the accuracy tends to remain the same or even deteriorate iterating through the solutions. This consideration is more in line

with how such transformers are built to function. Therefore, it might be argued that the first method chosen to evaluate the accuracy is not precise to describe the behaviour of the transformers when executing metonymy resolution and the second strategy might be preferred.

Even though the second strategy is the most comprehensive to evaluate the behaviour of transformers when dealing with metonymy resolution because it takes into consideration all the alternative senses of the answers returned by BERT and RoBERTa, it is very strict in the sense that the scores are generally very low. Such low scores are due to the fact that, except for few lemmas to which correspond just one sense, most lemmas have at least two senses or synsets. Therefore, the scores determined with the previous strategy must be analysed bearing in mind that also less frequent senses are still encoded in WordNet. For instance, when deal with PRODUCER-FOR-PRODUCT, “art” should be intuitively judged in most cases as a completely plausible answer; however, with the second strategy, its score is only 0.75 because of the four synsets there is one that is not included in the synsets list of the hypernym “artifact” because that synset encodes the less frequent sense of “art” as a skill. Therefore, in such cases the scores are falsely diminished because of less common synsets.

Based on this consideration, a third strategy is suggested. It is suggested that the most frequent sense should be the synset investigated; in fact, since it is the most likely out of the list of synsets, the assumption is that determining whether it is included in the semantic space should probably give us the most probable accuracy score, or at least the accuracy of the most probable meaning of that word. As stated by Jurafsky and Martin (2023), “senses in WordNet are generally ordered from most frequent to least frequent based on their counts in the SemCor sense-tagged corpus”. On the basis of this consideration, the third strategy was designed: since the assumption that the first synset is usually the most frequent, the aim of this method is to observe whether the sense encoded as the first synset and therefore considered as the most frequent is the one comprehended in the list of synsets determined by the hypernym. This way the accuracy analysis is not

affected by the less frequent synsets and is based solely on the most frequent and therefore probable sense of a specific word.

First, the plots for the analysis of the overall performance according to this third strategy will be presented below.

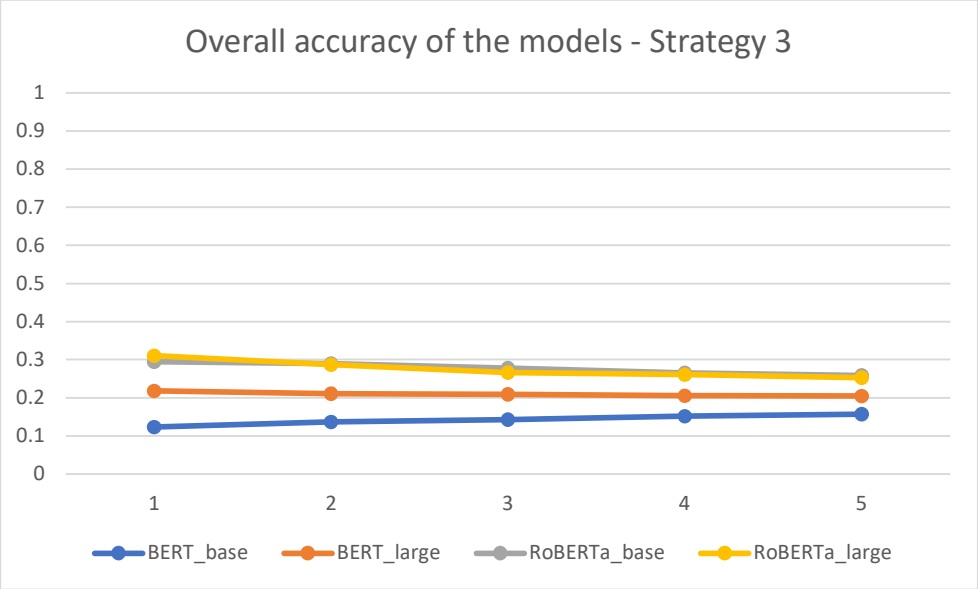


Figure 21 - Plot of the trend of the overall accuracy of the models according to the third strategy for evaluation. Each line represents the trend of the performance of a model according to average score on the y-axis and solution number on the x-axis.

As can be observed in Figure 21, despite slight variations, the third strategy is more in line with the second than the first. First of all, the performance of all the models is again quite poor when judged with this method, since the accuracy estimations do not manage to get past the 0.3 scores. Secondly, the ranking is confirmed: RoBERTa large proves once more to be the model that best deals with a metonymy resolution task. This time, however, the difference in the performance with RoBERTa base is practically invisible and at solution number five it seems that RoBERTa base outperforms, even just slightly, RoBERTa large. Also, the lines that represent the performances of the other two models, BERT base and BERT large, are not quite far apart from the others.

Finally, also for the third strategy the results were analysed separately for each category of metonymy. The plots showing the performance according to metonymy type can be found below.

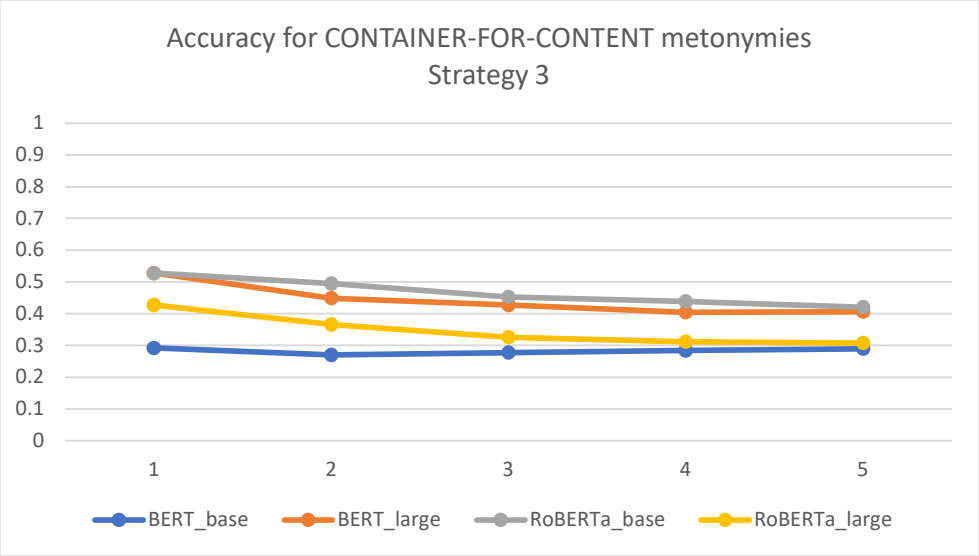


Figure 22 - Plot of the trend of accuracy for CONTAINER-FOR-CONTENT metonymies according to the third strategy for evaluation. Each line represents the trend of the performance of a model according to average score on the y-axis and solution number on the x-axis.

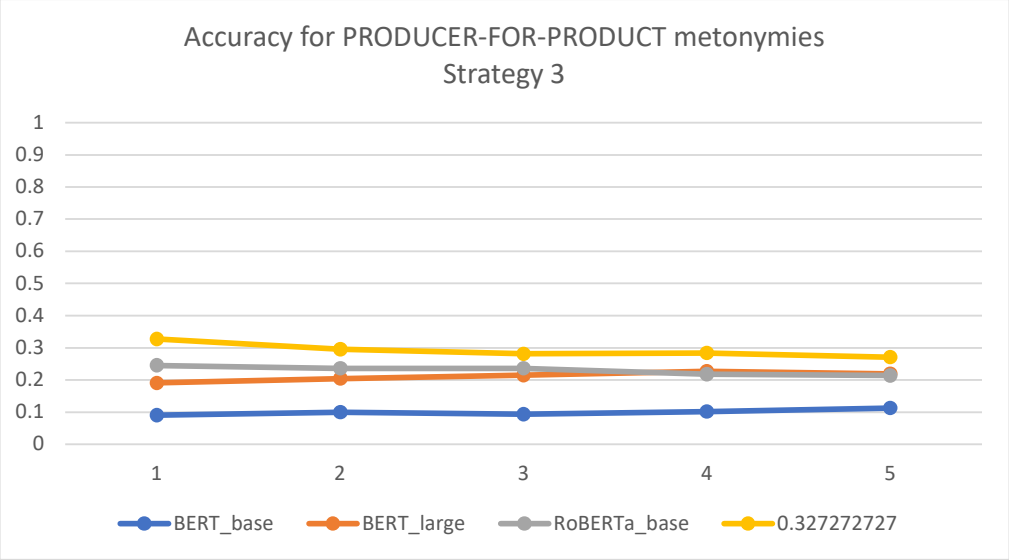


Figure 23 - Plot of the trend of accuracy for PRODUCER-FOR-PRODUCT metonymies according to the third strategy for evaluation. Each line represents the trend of the performance of a model according to average score on the y-axis and solution number on the x-axis.

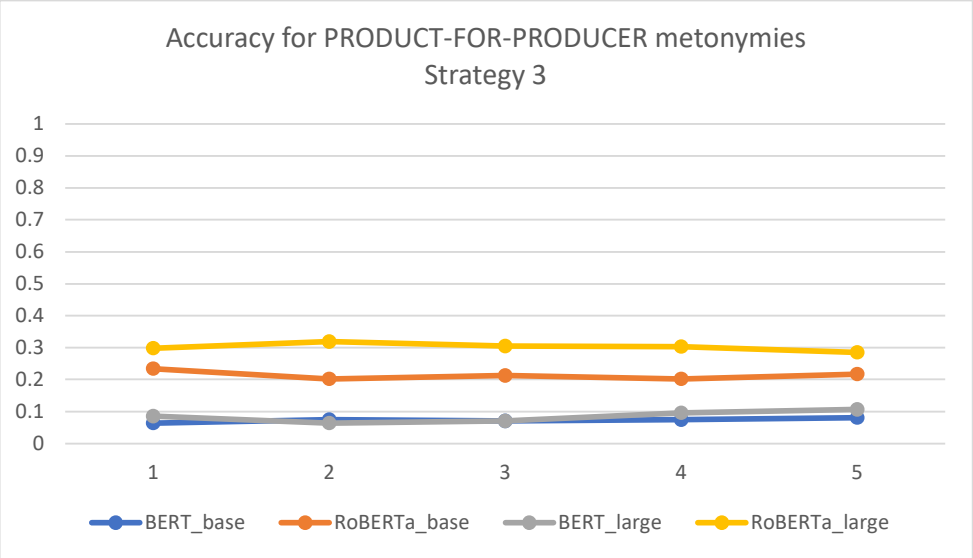


Figure 24 - Plot of the trend of accuracy for PRODUCT-FOR-PRODUCER metonymies according to the third strategy for evaluation. Each line represents the trend of the performance of a model according to average score on the y-axis and solution number on the x-axis.

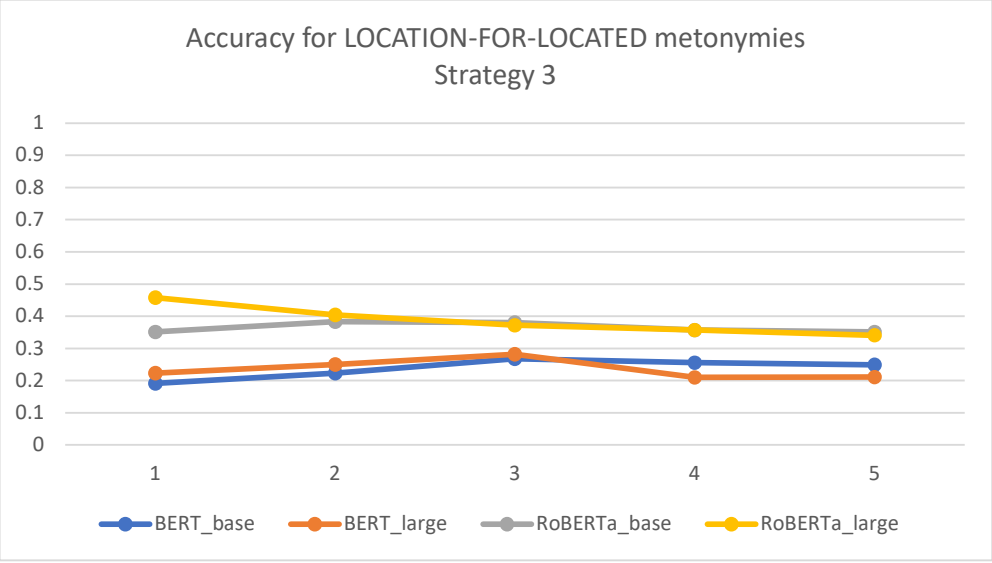


Figure 25 - Plot of the trend of accuracy for LOCATION-FOR-LOCATED metonymies according to the third strategy for evaluation Each line represents the trend of the performance of a model according to average score on the y-axis and solution number on the x-axis.

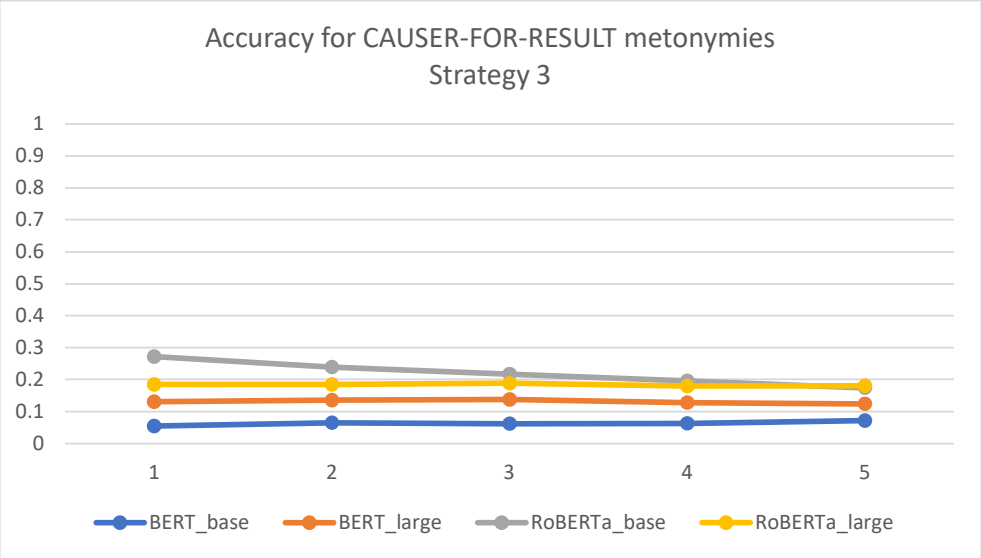


Figure 26 - Plot of the trend of accuracy for CAUSER-FOR-RESULT metonymies according to the third strategy for evaluation. Each line represents the trend of the performance of a model according to average score on the y-axis and solution number on the x-axis.

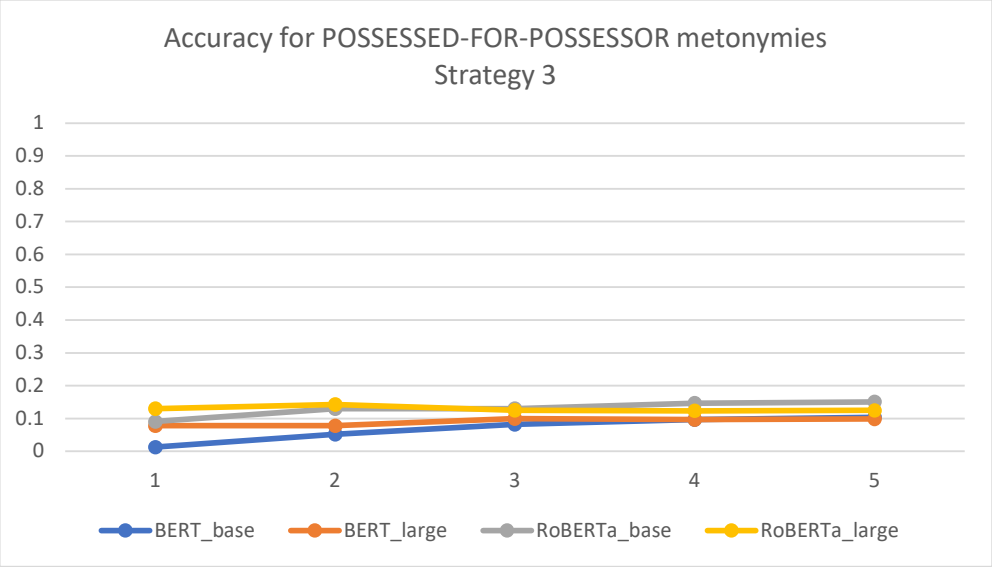


Figure 27 - Plot of the trend of accuracy for POSSESSED-FOR-POSSESSOR metonymies according to the third strategy for evaluation. Each line represents the trend of the performance of a model according to average score on the y-axis and solution number on the x-axis.

As can be observed in Figures 22 to 27, applying the third strategy to the analysis of the performances of the models according to the type of metonymy results in very similar evidence as while applying the second strategy. As a matter of fact,

there are again some categories that have proven to be generally better dealt with by transformers, such as CONTAINER-FOR-CONTENT as shown in Figure 22, and categories with which transformers struggle to identify the correct referent, such as POSSESSED-FOR-POSSESSOR as shown in Figure 27. Moreover, with few exception, the third strategy confirm what the second methodology had shown: the improvement over the iteration of the solutions that the first strategy highlighted was probably due to a too forgiving evaluation procedure. Both the second and third analysis showed that an improvement in the solutions is quite unlikely since the accuracy tends to either be maintained or decrease.

4.2.4 Discussion

The aim of this first experiment was to investigate what semantic information transformers, such as BERT and RoBERTa, can understand about an instance of figurative language, namely metonymy. More specifically, the experiment was built with the goal of gathering information about whether language models are able to trace back a metonymic instance to the intended referents. After having conducted the experiment and analysed the behaviour of the models through the three strategies, it could be observed that the performances of the models are far from satisfactory. The three strategies returned slightly different evaluations: in particular, the first method judged the model less severely and the accuracy scores according to the first strategy are generally higher than the other two procedures. Nonetheless, even according to the first strategy, the average overall accuracy scores of all four models do not go much beyond 0.5, meaning that their guesses for the intended referents of the metonymic expressions are precise just about half the times. Said accuracy scores dropped even lower according to the second and third strategies.

Secondly, the comparison among the performances of the four transformers is worth to be mentioned. Considering the overall accuracy, on average RoBERTa large was the model that performed the best on metonymy resolution; however,

its performance according to strategy 2 and strategy 3 almost overlaps with the performance of RoBERTa base, meaning that both these models deal with metonymy basically in the same way. Following in the ranking RoBERTa base, BERT large is the third model, while BERT base is the model that performs the worst among the four transformers selected. Nonetheless, it has already been mentioned that the difference in the performances according to transformer type was not remarkable in most instances.

The last aspect to be mentioned is the difference in the performance according to metonymic type. The second step for each strategy was to consider the evaluation according to each category of metonymy as well in order to understand if and how the type of metonymic expression affected the performance, i.e. whether some kinds of metonymy are more easily processed than others. From the observation of the scores split for metonymic type, a quite remarkable difference emerged in how precisely different metonymic expressions are dealt with by the transformers. In fact, given the higher scores, it could be argued that this kind of language model seems to find categories such as CONTAINER-FOR-CONTENT and LOCATION-FOR-LOCATED metonymies easier to interpret, while the metonymic type that appears to be the most ambiguous for transformers is POSSESSED-FOR-POSSESSOR. This aspect would require further investigation, maybe involving a psycholinguistic experiment in order to study whether this divergence in the performance is replicated in the human brain; however, a possible hypothesis could be that the first types of metonymic expressions are more easily dealt with because of two possible aspects: firstly, both CONTAINER-FOR-CONTENT and LOCATION-FOR-LOCATED metonymies are more frequent in everyday language and secondly, they are somewhat more “fixed” and the image schema, suggested by Lakoff (1987), proves to be a useful tool to interpret this kind of metonymic occurrences. POSSESSED-FOR-POSSESSOR metonymies, on the other hand, highlight the way people perceive the world and, for that, there is no fixed rule; thus, this type of metonymy allows for a higher degree of creativity and consequently unpredictability. This property may result in a problematic aspect for

transformers: even though it may seem that artificial intelligence reached a level of linguistic ability comparable to the human mind in most tasks, language models are still trained on corpora, that despite their sizes could never cover all the instances of natural language, and, moreover, transformers like BERT and RoBERTa still struggle to inference meaning in contexts where what is said is not to be interpreted literally, but rather requires the intuition to go beyond the literal information and interpret it figuratively.

4.3 Experiment 2: contextual word embeddings comparison

Experiment 1 showed that most times BERT and RoBERTa struggle to trace back a metonymic expression to its intended referent and, therefore, their performance produce in inaccurate results. However, there are also instances where these transformers managed to retrieve a possibly correct referent, and, although the successful attempts are just a minority, it cannot be stated that it is just a coincidence. If that would be the case, the accuracy score on such a relatively large corpus as that considered would have been way lower. On the basis of this argument and the fact that the hidden mechanisms of BERT and RoBERTa are still quite mysterious, what happens "behind the scenes" could be worth the research. For this reason, a second experiment is proposed in this thesis with the aim of investigating what happens at each of the layer of these transformers when said transformers are asked to deal with metonymy resolution. Specifically, RoBERTa large was selected since on average it was the model that proved to be the best performing on the interpretation of metonymic expressions out of the four transformers.

4.3.1 Background

The inspiration for this second experiment was drawn from the same paper from which the dataset of metonymies was retrieved, namely the research by Pedinotti and Lenci (2020). As previously mentioned, in their study the researchers investigated how well BERT base was able to interpret metonymic expressions. In order to do so, in their first experiment they computed the cosine similarity between two instances for each entry in the dataset: the first cosine was measured between the embeddings of the target word in a metonymic sentence and the same target word in a sentence where it was used with its literal meanings ($sim(\overrightarrow{met}, \overrightarrow{lit})$), while the second cosine was measured between the embedding of the target word in the metonymic sentence again and the embedding of a

possible paraphrase of the metonymic expression ($sim(\overrightarrow{met}, \overrightarrow{metpar})$). Then, the results were compared: if the first cosine similarity $sim(\overrightarrow{met}, \overrightarrow{lit})$ was greater than the second cosine similarity $sim(\overrightarrow{met}, \overrightarrow{metpar})$, then they argued that it would imply that the meaning of the metonymic expression was more similar to the literal meaning than to the meaning of the paraphrase and that, therefore, the metonymic sentence was not correctly processed by BERT. On the other hand, in the case the first cosine was found to be smaller than the second, then it could be proved that the transformer was indeed able to correctly interpret metonymic instances. Pedinotti and Lenci observed that the cosine similarity between the metonymic sentences and the literal sentences was bigger that the cosine similarity between metonymic sentences and the corresponding metonymic paraphrases and therefore, according to their findings, the researchers argued that BERT base was not in fact able to carry out metonymy resolution.

The reason why a similar experiment is inserted in this thesis is that Pedinotti and Lenci's experiment failed to take into consideration some aspect that could have influenced their conclusions. First of all, they took into consideration only BERT base as example of a transformer, which is a critical choice given the evidence from the first experiment of this thesis, which highlighted that BERT base was the worst model of the four taken into consideration in solving metonymic expressions. This was the ground on which the choice of RoBERTa large was made: observing the behaviour of a model that seems to perform better could potentially impact the judgement on the interpretation process. Other than the type of transformer chosen for the experiment, the comparison of the cosine similarities could be improved. In fact, Pedinotti and Lenci compared the two measures of similarity only by defining the success or failure of BERT only on the base of which of the two was the larger cosine without considering how great the shift from the literal to the metonymic expression was. On this ground, this second experiment has the aim of understanding how RoBERTa large works while dealing with metonymy by computing a measure that unifies the two cosine as calculated by Pedinotti and Lenci but with the addition of a normalisation process, which

consisted of calculating the difference between the cosine similarity between the embedding of the literal sentence and the embedding of the sentence with the metonymic paraphrase and dividing the result of the subtraction by the cosine similarity between the literal sentences and the sentences with the metonymic paraphrase. Thus, the formula obtained was as follows:

$$metonymy\ comprehension\ score\ = \frac{sim(\overrightarrow{met}, \overrightarrow{metpar}) - sim(\overrightarrow{met}, \overrightarrow{lit})}{sim(\overrightarrow{lit}, \overrightarrow{metpar})}$$

By normalizing the difference between the cosine $sim(\overrightarrow{met}, \overrightarrow{metpar})$ and the cosine $sim(\overrightarrow{met}, \overrightarrow{lit})$ by the cosine $sim(\overrightarrow{lit}, \overrightarrow{metpar})$, the distance between the literal meaning of the target word and the paraphrasis of the metonymic sentence is taken into consideration as an affecting variable. The normalization steps takes into account that the distance between the target word used literally and the paraphrase of the target word used metonymically could affect how much the model is able to infer about the connection between the target word used metonymically and its paraphrase. Too distant vectors can cause trouble in understanding the link between the two, as well as confusion in the distinction between metonymic and literal used may emerge from too near vectors. Therefore, the cosine similarity $sim(\overrightarrow{lit}, \overrightarrow{metpar})$ should be considered as an affecting variable.

4.3.2 Tools used in the study

While for the first experiment it was not required much knowledge about the architecture and the functioning of transformers in order to understand the concepts behind the study, for this second experiment it is essential to spend a few words to give a more in-depth explanation of how this kind of language model

are able to process texts and what steps must be performed in order to return valid results.

Firstly, it is relevant to mention that neural networks are not able to process raw texts because they do not have the linguistic knowledge to understand the semantics behind the sentence structure. Therefore, in order to render natural language available to machines, the information given in a raw text has to be transformed into data that the neural network is able to process and the only "language" machines can deal with is numbers. The way we can transform words into numbers is by using vectors. Said vectors consist of arrays of numbers and provide a way to represent meaning in a high-dimensional space. The number of dimensions of space is determined by the number of contexts in which the word meaning is considered. Ideally, to get an accurate representation of word meaning, the number of contexts should correspond to the size of a vocabulary. This, however, comes with a downfall: if all the contexts in which a word could appear are considered, a long and sparse[4] vector is generated. This kind of vector is indeed problematic for machine learning purposes because they would require more parameters and the model is at a higher risk of overfitting and, thus, it is used only in count models. An ideal vector for machine learning, instead, should be short and dense, and this type of vector is also called word embedding.

Comparing the vectors, given their different angles, is possible by computing the cosine similarity. This value is computed based on the dot product between two vectors normalised by the product of the lengths of each vector and it ranges from -1 if the vectors are in completely opposite direction to 1 if the direction is exactly the same for both of them.

$$sim_{NDP}(\vec{v}, \vec{w}) = \frac{\vec{v} \cdot \vec{w}}{|\vec{v}| \cdot |\vec{w}|} = \frac{\sum_i^N (v_i \cdot w_i)}{\sqrt{\sum_i^N (v_i^2)} \cdot \sqrt{\sum_i^N (w_i^2)}}$$

[4] A sparse vector is a vector whose dimensions are mostly zeros.

As previously mentioned, transformers such as BERT and RoBERTa cannot deal with raw text. Thus, the text fed to the models needs to be pre-processed for these models to interpret it: this means that the sentences must undergo a tokenization step, in which from a complex string the text is deconstructed into its building blocks, namely the tokens. In a second step, to each token a vocabulary index is attributed. In order to render this value contextualized, the model should be able to trace back the token embedding to the sentence it was taken from; to do so, a sentence embedding is generated as well. The last element a transformer needs to process a contextual embedding is the positional embedding, which indicates the position at which the target word occurs in the sentence and correspond to a contextualized index. With these three elements, the vocabulary indexes of the tokens, the sentence embedding, and the positional embedding, the models are able to process the contextual embedding of a target word in a given sentence.

As just stated, if a transformer is fed with this information, it returns one value that corresponds to the output of the model. However, transformers have a variable number of hidden layers, which, as the name indicates, cannot be seen, but each of them receives in input the output of the previous layer. This means that each layer actually produces a result, even though just the output of the last layer, which is supposed to be the most accurate result the model was able to achieve, is shown. By default, the outputs for layer are not access because they are considered irrelevant for the purpose of solving a task since, except for the last output, what has been produced in the previous layers is not judged to be an accurate solution. Nonetheless, the model can be manually asked to separately return the output of each layer: this aspect is of remarkable importance to investigate what kind of contextual embedding the transformer was able to elaborate at each level.

4.3.3 Methodology

In order to proceed to process the sentences through RoBERTa large, the first was to manipulate the dataset because it was necessary to insert a column with the target words corresponding to the referred sentence to facilitate the access to the positional embedding. So, to the column of “Metonymic Sentence” corresponded the column of “Target Word for Metonymic Sentence”, to the “Literal Sentence” column corresponded the “Target Word for Literal Sentence” column, and finally, to the “Sentence with Metonymic Paraphrase” corresponded the “Metonymic Paraphrase”. Moreover, the target words were corrected or reported to correspond in number to the term in the referred sentence.

Afterward, the second step in the pre-processing phase was rendering the semantic information contained in the sentences available to the model so that it could be processed to return the embeddings. Therefore, through a code which was written in Python, the sentences and the target words were imported by using a Pandas dataframe (McKinney, 2010). Then, the sentences were tokenised using the tokenizer for the English language named Stanza[5] (Qi et al., 2020). Based on the observation of the result of the tokenization process performed by Stanza, it could be said that this tokenizer is a whitespace tokenizer, meaning that the tokenizer infers the divisions between words on the basis of the white spaces. The reason why it was important to match the target words and the words in the context of the sentences in number was thus due to the tokenizer, since the tokens are not trace back to their lemma.

In the second step of pre-processing, the vocabulary indexes of the tokens, the sentence embedding, and the positional embedding for each entry in the dataset were initiated by working recursively through the sentences. It is relevant to mention that the Autotokenizer[6] for RoBERTa large was used to retrieve the word embeddings, since Stanza does not seem to have this option and it is only able to

[5] https://stanfordnlp.github.io/stanza/tokenize.html
[6] https://huggingface.co/docs/transformers/v4.26.0/en/autoclass_tutorial#autotokenizer

create an index for the positions of the words in the sentences, i.e. the positional embeddings. Moreover, since the sentences were processed recursively through a for-loop, the sentence embeddings for all sentences were set to 1.

With the vocabulary indexes of the tokens, the sentence embedding, and the positional embedding of the target token for each sentence, it was possible to proceed to the processing of said sentences through RoBERTa large in order to generate the contextual embeddings for the target words. The model was fed with this information, and it was asked to return all the hidden states, except for the first hidden state because it was the input layer, i.e. the original embeddings. The remaining 24 hidden states were stored in a separate list for each sentence.

The last step before computing the cosine similarity was isolating the embeddings from each layer of the target word of each sentence. To do so, the embeddings of all layers for each sentence were once again worked though recursively, saving in a separate list just the 24 embeddings of each target word. Thus, 509 lists of 24 lists were obtained for each of the three types of sentences, the metonymic sentences, the literal sentences, and the sentences with the metonymic paraphrase.

For each triple three cosine similarities were computed: the cosine similarity between the target word in a metonymic sentence and the same word in a literal context, the cosine similarity between the target word in the metonymic sentence and the paraphrase of the corresponding metonymic word, and lastly the cosine similarity between the target word used in its literal sense and the metonymic paraphrase. All the values were stored in three different Excel files, each containing 509 rows corresponding to the entries in the dataset for 24 columns corresponding to the layers of RoBERTa large.

Then, the values of all three files were split according to metonymic type, namely CONTAINER-FOR-CONTENT, PRODUCER-FOR-PRODUCT, PRODUCT-FOR-PRODUCER, LOCATION-FOR-LOCATED, CAUSER-FOR-RESULT, POSSESSED-FOR-

POSSESSOR, and the average cosine similarity was calculated for each of these types.

The metonymy comprehension score created to compute a measure that could sum the meaning of the three cosine similarity was used to establish a single method to judge the performance of the model at each hidden state.

4.3.4 Evaluation of the performance of RoBERTa large

Applying the formula that subtracts from the cosine $sim(\overrightarrow{met}, \overrightarrow{metpar})$ the cosine $sim(\overrightarrow{met}, \overrightarrow{lit})$ and normalize the subtraction by the cosine $sim(\overrightarrow{lit}, \overrightarrow{metpar})$ produced values that can be used for the evaluation of the performance of RoBERTa large. As already mentioned, each of said values correspond to the average measurement of each layer or hidden state of the transformer, according to a specific type of metonymy.

Before, taking into consideration the measurements that put together the three values, it is worth to consider the trend of the average cosine similarities for each layer for each type of metonymy. From the data obtained, a few observations can be made. Ideally, the cosine similarity between metonymic sentences and sentences with metonymic paraphrase should tend to values closer to 1 if the model managed to understand metonymy. As matter of fact, higher values would mean that the embeddings of the metonymic expression and the embeddings of its paraphrase are very near in the semantic space and therefore it could be possible to argue that the model understood that the metonymic expression and its paraphrase have, if not the same, at least very similar referents. As can be observed from the data, the values of the cosine similarity between the metonymic sentences and the sentences with metonymic paraphrase tend to increase and by the last layers of the model the similarities get quite close to higher values around 0.9.

Based on the final results of the cosine $sim(\overrightarrow{met}, \overrightarrow{metpar})$, which represent the final output of the transformer model, it seems that the model is indeed able to understand that a metonymic instance and its referent are the same, given the quite impressive performance it achieves by the end of the processing. However, if so, the question would be why the results were not as satisfactory in the previous experiment, where, even at the best of it the ability, the scores of accuracy did not manage to get past 0.6. The answer to this question can be found when analysing the values of the cosine similarity between the metonymic sentences and the literal sentences. In the case of the similarities $sim(\overrightarrow{met}, \overrightarrow{lit})$, the opposite situation should be desired if we would expect the model to perform well on metonymy resolution: the lower the cosine similarity is, the better the model would seem to understand that the referent of the metonymic expression is not the same referent of that same word but used with its literal meaning. Looking at the computed cosine similarities, it can be immediately noticed that unfortunately this is not what happens according to the evidence reported in this experiment. In fact, the values of the cosine similarities tend to be quite high already in the first layers and they are maintained at a high values until the last layers. Nonetheless, an interesting feature of these values to pay attention to is that the trend is not constant towards higher or lower scores, but it fluctuates instead: sometimes the similarity decreases, other times it increases without following a specific pattern. This aspect could be meaningless since the fluctuation is generally not wide, but it could also signal an uncertainty of the model in the prediction of the target word embeddings.

The high values in both types of cosine similarity are the reason why the performances in the first experiment were on average quite poor and the models managed to guess the correct referents only about half the time. However, a last value should be considered in order to make the judgement as objective as possible, namely, as stated at the beginning of this chapter, the cosine similarity between the literal sentences and the sentences with metonymic paraphrase. The reason why it is relevant to include this variable in the study is that the distance

between the target word of the literal sentence and the target word of the metonymic paraphrase could possibly affect the performance of the model and change in the different layers. If the distance between these two target words was greater than expected, it would be harder for the model to understand that the intended referents of the metonymic sentences correspond to the metonymic paraphrases because the meanings of the two target words would be too further apart, but it will cause trouble as well if the distance was too narrow since it would cause uncertainty and ambiguity, since the model would not be able to distinguish the metonymic from the literal use.

Observing the cosine similarities obtained, we can see how the trend fluctuates also in this case as in the previous one, but similarly to the values of the first cosines computed they started a bit lower but by the last layer end up very close to 1. Thus, it could be argued that the model is actually more accurate at the beginning than at the end of the process because the values at the beginning are lower but above zero, so the transformer understands that there is a similarity in meaning but not to the point where the meaning is exactly the same, which is the conclusion it seems to draw at the end of the cycle, given the fact that a score so close to one means that the referent is almost exactly the same.

After these considerations, the average cosine similarities for metonymic type should be combined in a single value in order to generate a representation of how the model is performing. To compute said value, the previously mentioned formula was employed, and a plot was created in order to represent the trend of the performance of the transformer.

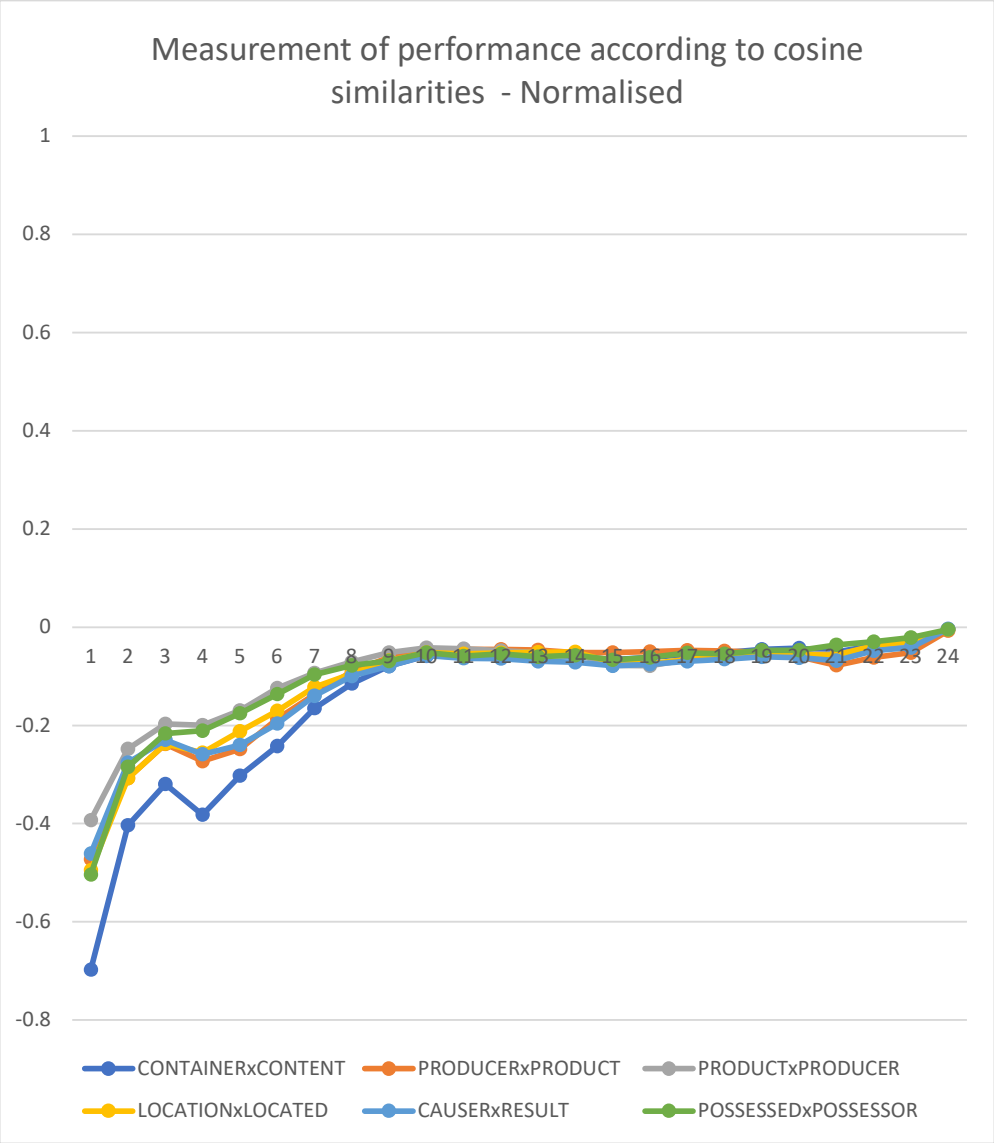


Figure 28 - Plot of the trend of performance for metonymic type according to the metonymy comprehension score (with normalization) on the y-axis and to the number of layer on the x-axis

In Figure 28, each line represents how well the model was performing on a particular type of metonymy. As we can observe from this representation, on average the model performed worse in the first layers but, even with some fluctuations, it improved over the last layers, and by the 24th layer RoBERTa large reached the zero score. This result is still far from ideal, and it cannot be stated that this language model correctly infers that a metonymic sentence and its

paraphrase have the same referent while the target word of a metonymic sentence and that same target word used in a literal context have separate meanings.

An ideal result that would prove that the transformer detected the meaning shift would have been an at least positive value, keeping in mind that it would not be possible to ask RoBERTa to produce results very close to one. The reason why that is not realistic is that the value of the cosine $sim(\overrightarrow{met}, \overrightarrow{metpar})$ cannot possibly go above one, since the cosine similarity ranges from -1 to 1, and the only way to get a result close to 1 from the subtraction between $sim(\overrightarrow{met}, \overrightarrow{metpar})$ and $sim(\overrightarrow{met}, \overrightarrow{lit})$ would be if the values of the cosine $sim(\overrightarrow{met}, \overrightarrow{lit})$ were very close to zero. However, this is again unlikely because the contextual embeddings of the same word but used in different contexts, even though they should not be corresponding, must be similar to a degree, especially in the case of metonymic expression, where the concepts are related. Substantially different embeddings of the same word form may be expected only if said word has completely separate meanings: for example, the cosine similarity between the contextual embedding of “bank” in the sense of the building and the contextual embedding of “bank” in the sense, instead, of the slope beside a body of water can be expected to be quite low. In the case of metonymy, this phenomenon should not be repeated because a fundamental property of such figure of speech is that the term used in the metonymic expression and its referent should belong to the same domain, and therefore be similar in some regards. This connection implies a similarity in the embedding and therefore the cosine similarity between the metonymic sentences and the literal sentences cannot be expected as a low value. On the basis of this consideration, the expectation of a performance measurement close to one is not considered realistic because even if the cosine similarities between the metonymic sentences and their paraphrases were close to 1, they would be balanced out by the cosine similarity between metonymic and literal sentences. On the basis of this reflection on the realistic expectations for a transformer, it could be stated that a language model would be evaluated positively if the final

values went at least a bit above zero. Said positive values would mean that model understood that the similarity between the metonymic sentences and their paraphrases was greater that the similarity between the metonymic and literal sentences, and therefore predict as more likely the figurative use of a metonymic sentence than the literal use. However, as previously argued, this does not seem the case with RoBERTa: the values obtained from the difference of the cosine similarities showed that not only the model does not predict a greater similarity between metonymic sentences and their paraphrase, but, given the negative scores in the first layers, at least at the beginning of the process the transformer seems to prefer the literal interpretation in the case of metonymic expression, given the similarity of the embeddings of the metonymic and the literal target word in the sentences.

Nonetheless, it is remarkable that the performance over the iteration of the twenty-four layers noticeably improves. In fact, the performance measurements at the first layers are negative, but slowly but steadily they improve over the iteration through the layers. Therefore, even though it cannot be stated that the performance of RoBERTa large on the task of interpreting metonymy is the most effective, this improvement showed that over the process the transformer seems to start to better infer the similarity between the metonymic sentences and their paraphrases. The problem is that the improvement concerns only the cosine similarity between metonymic sentences and their corresponding paraphrases. Through the observation of the data, it can be seen that for all metonymic types the cosine $sim(\overrightarrow{met}, \overrightarrow{metpar})$ started at values around 0.5-0.6 and then it reaches scores around 0.9 by the end of the process; this improvement is significant for a positive trend in the performance of the model. The reason why this enhancement in the inference of the interpretation of metonymy is not enough for an overall positive judgement of the model is because this measurement is balanced out by the other two trends: the trend of the cosine similarity between metonymic and literal sentences remains almost unvaried during the iterations, while the trend of the cosine between literal sentences and metonymic paraphrases shows an

undesired increment in the progression through the layers despite some uncertainties. The data shows us that by the end of the process, the average values of the three cosine similarities is almost the same and for that the performance measurements obtained at the 24th layer are close to zero. If the cosine similarity between metonymic sentences and their paraphrases had prevailed on the other two similarities, as desired, then the measurements would have gone above zero.

Other than the overall judgement on the trend of the performance of the model, it is relevant to analyse how differently RoBERTa large behaves according to metonymic type. It has already been reported that by the end of the processing the measurements of all metonymic type reach values around the 0-score. However, the divergence is remarkable in the first layers: the category that is best dealt with is PRODUCT-FOR-PRODUCER, while the metonymic type that seems to cause more struggle for the transformer is quite surprisingly CONTAINER-FOR-CONTENT, whose score at the first layer is placed around -0.7. This evidence is unexpected because in the first experiment of this thesis this latter type of metonymic expressions was the one processed with most ease by all models. On the other hand, the class of POSSESSED-FOR-POSSESSOR, which was connected to the worst performance in the previous experiment, is just the second category worst interpreted at the first layer.

The last observation worth to mention is a peculiar characteristic of the trend that is repeated in each metonymic category. Even though the trends tend to increase over the iteration over the layers, there is a common point where all the performances seem to step back, namely between the third and the forth layer the measurements decrease to increase again after the forth layer. This step back is more marked for some metonymic type, such as CONTAINER-FOR-CONTENT metonymies that from -0,31 returns to -0.38, but it is still evident in case where it is less steep the decrease curve. The only exception to this phenomenon is the category of POSSESSED-FOR-POSSESSOR, whose trend line does not show any drop.

A last note should be mentioned on the matter of normalization and how it impacts the measurement of performance. A second plot was drawn in order to show such difference: the measures were this time computed on the sole basis of the differences between the cosine $sim(\overrightarrow{met}, \overrightarrow{metpar})$ and the cosine $sim(\overrightarrow{met}, \overrightarrow{lit})$. The results thus obtained are as shown in Figure 29.

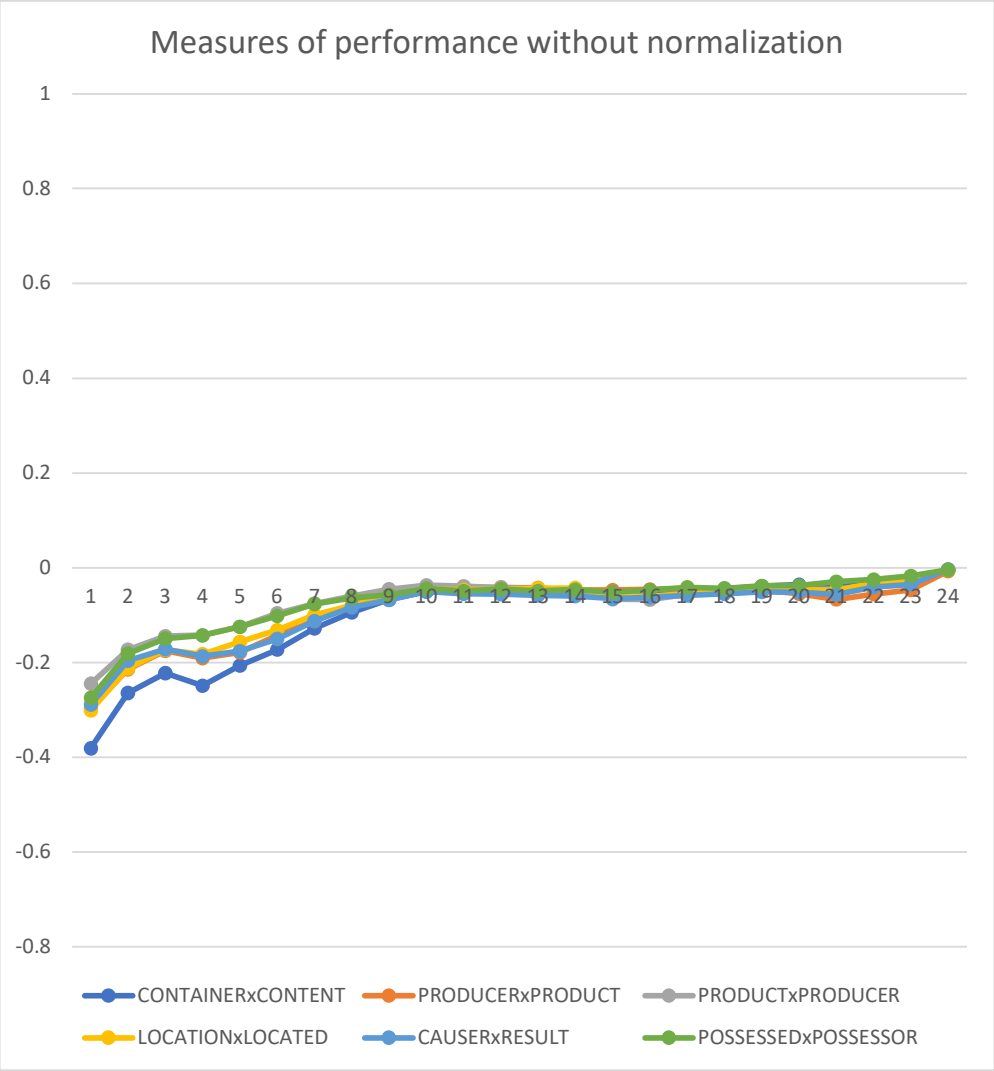


Figure 29 - Plot of the trend of performance for metonymic type according to the metonymy comprehension score (without normalization) on the y-axis and to the number of layer on the x-axis

As it can be observed from Figure 29, the measurements starting from the middle until the last layers are not really affected by the normalization since they remain almost equals to the previous computation. This is due to the fact that the difference between the cosine $sim(\overrightarrow{met}, \overrightarrow{metpar})$ and the cosine $sim(\overrightarrow{met}, \overrightarrow{lit})$ is approximately close to zero and, therefore, the normalization division does not affect the end results. On the other hand, the difference between the cosine $sim(\overrightarrow{met}, \overrightarrow{metpar})$ and the cosine $sim(\overrightarrow{met}, \overrightarrow{lit})$ is greater given the lower similarity between the metonymic sentences and their paraphrase, thus the results in the first layers are indeed altered compared to Figure 28, since the values are halved. This proves that the distance between the embeddings of the literal sense of the metonymy and the paraphrase of the metonymy affects the ease with which RoBERTa interprets metonymic expressions. The reason is that the meaning shift is not judged as great as expected, since RoBERTa produced contextual embeddings for the literal sentences which are quite close to the contextual embeddings for the metonymic paraphrases. The normalization takes into consideration that the meaning shift is probably not sufficiently wide for the mode to correctly distinguish between metonymic and literal instances, and this results in confusion while returning the interpretation as process by the model for the metonymic expression in the first experiment.

4.3.5 Discussion

The second experiment confirmed that, even the model that was performing a bit better in the first experiment, RoBERTa still has a lot to learn about how to deal with the task of metonymy resolution. However, this analysis showed that, even though there is still room for improvement, a model, such as RoBERTa large, notices that there is something happening to the meaning in the case of metonymic expressions which does not correspond to literal language. In fact, as the results have demonstrated, the behaviour of the model is changed over the course of the iteration of the layers, namely the cosine similarity between the

metonymic sentence and the possible paraphrases increases for each metonymic type. This enhancement in the interpretation means that the transformer began to understand that even though the two target words apparently differ, they have instead a common referent. The problem that makes the performance of RoBERTa still not satisfactory is that the model does not seem to be able to distinguish when the same common target word is embedded in different contexts and this determines a difference in the meaning, such as in the case of metonymic sentences and literal sentences. In this case, despite the fluctuation in the cosine similarities which could potentially signal that the model has some doubts, RoBERTa still identifies the two terms with a single referent. This mistake has a huge impact on the performance of the language model since it prevents said model from detaching the two referents and therefore it does not seem to be able to go beyond the literal meaning of the target words.

A second consideration concerning the trend of the performance is worth to be mentioned, namely the comparison of the findings according to the experiment in this thesis and the findings according to previous research. As a matter of fact, according to Rogers et al. (2020) it was argued that "the final layers of BERT are the most task specific" (pp.848). Moreover, Tenney et al. (2019) suggested that semantics, contrary to syntax, is processed at all layers and this is evident when some examples are incorrectly processed at first but they are better dealt with over the iteration of layers. Even though RoBERTa was employed instead of BERT, this observation is indeed reflected in the results produced by the second experiment since the models are architecturally the same: the metonymy comprehension scores were on average higher in the last layers compared to the score in the first layers. Thus, it could be hypothesised that the model in the last layers better understood the task of metonymy resolution, thanks to a better comprehension of the semantic information as well, and therefore started to produce, even if not ideal, more accurate results.

V. Conclusions

The project of this thesis dealt with the task of metonymy resolution as processed by transformer language model. The aim was investigating whether such models were able to understand the hidden meaning of a figure of speech such as metonymy. In order to analyse if transformers have this ability, four models were taken into consideration: BERT base, BERT large, RoBERTa base, and RoBERTa large. Moreover, two experiment were built to generate an inclusive evaluation of the performance of the previously cited models.

In the first experiment the performance of these transformers were studied to determine whether they were able to generate their own alternative plausible referents for a dataset of six different types of metonymic expressions, namely CONTENT-FOR-CONTAINER, PRODUCER-FOR-PRODUCT, PRODUCT-FOR-PRODUCER, LOCATION-FOR-LOCATED, CAUSER-FOR-RESULT, and POSSESSED-FOR-POSSESSOR. The results were judged against the relations between the hypernyms and the hyponyms as encoded in WordNet. The findings of the first experiment showed that the ability of these transformers when dealing with metonymy resolution is not in fact satisfactory, given the average scores returned from the three different analyses. Moreover, it was observed that, even though the majority of metonymic expressions was not interpreted correctly by the models, the scores were not sufficiently low as well to just determined that the correct answers returned were purely a matter of chance.

On this ground, the second experiment was formulated in order to investigate the hidden mechanisms put into action when the task of metonymy resolution was asked to be performed. To conduct this study, RoBERTa large, the model that on average performed best in the first experiment, was chosen to be analysed. For each metonymic type the contextual embeddings of three different instances of sentences were extracted from each layer and the cosine $sim(\overrightarrow{met}, \overrightarrow{metpar})$, the cosine $sim(\overrightarrow{met}, \overrightarrow{lit})$, and the cosine $sim(\overrightarrow{lit}, \overrightarrow{metpar})$ were calculated. A formula

that subtracted the cosine $sim(\overrightarrow{met}, \overrightarrow{lit})$ from the cosine $sim(\overrightarrow{met}, \overrightarrow{metpar})$ normalized by the cosine $sim(\overrightarrow{lit}, \overrightarrow{metpar})$ was generated to create a single measure to evaluate the performance at each hidden state. On the basis of the results, it was observed that the understanding of metonymy by RoBERTa was especially poor in the first layers, while it got better over the iteration of the final layers. Nonetheless, the performance was found to be insufficient to be judged positively since even in the final layers, the performance measure did not manage to go beyond the zero score. Thus, it is understandable why the answers in the first experiment were not as accurate as it was desirable. Moreover, it emerged from the results that the confusion in indicating the correct referents of said metonymic expressions probably arose not because of a lack of understanding the semantic similarity between the metonymic sentences and the corresponding sentences with metonymic paraphrase, but rather it was due to a too high similarity between the metonymic sentences and the literal sentences. It may be argued that it could have been possible to obtain better interpretations from RoBERTa if said model would have been able to better distinguish the use of figurative language from the use of literal language.

5.1 Limitations

Although the results from both experiments show that these models do not seem to be able to deal accurately with metonymy resolution and seem to provide enough evidence for the evaluation of the selected models, there are a few limitations to the present study which should be mentioned.

First of all, this thesis took into consideration four models, which is a relatively large sample given the fact that most research compares the analyses of a pair of models at time. However, BERT and RoBERTa belong to the same generation of transformers created a few years ago. Even though a few years does not seem such a long time ago, in a fast-developing field such as NLP it is still a considerable

amount of time. Since the release of BERT and RoBERTa, several new models have been shared with the research community that seem to perform way better than the previous ones. Therefore, the evidence carried out from this experiment cannot be comprehensive evaluation of all language model, but rather should be limited to the models taken into consideration.

Moreover, another problematic aspect was the evaluation of the results produced by the transformers in the first experiment. As already mentioned, there is a possibly infinite numbers of interpretations of a metonymic expression and this aspect turns out to be an obstacle for the judgement of the solutions because of the inherent problem of finding a method to determine which answers can be considered correct and which wrong that includes the highest number possible of correct referents. The methodology used in this project was checking the semantic spaces generated by the hypernyms, but it is possible that, since the hypernyms were manually selected, the choice was biased and the hypernyms did not include possibly correct solutions, the same way they possibly included wrong answers.

5.2 Future work

This project aimed at beginning the investigation of what some language models infer about the interpretation of metonymic instances. However, this study could offer a cause of reflections on what future work concerning transformers and metonymy could look like.

Firstly, as mentioned in the limitations of this project, it could be worth to extend the investigation on metonymy resolution to other, more advanced models in order to observe whether an overall improvement of their ability to process and produce natural language includes an improvement as well on how well they are able to deal with instances of figurative language, like metonymy.

Secondly, this project was focused on referential metonymy, but it would be interesting to extend the analysis to logical metonymy to evaluate whether

transformer struggle more or find easier to process metonymic verbs rather than metonymic nouns.

Thirdly , an aspect worth further exploring is, for instance, the matter of contextual embeddings of the solutions returned by the models. In this thesis it has been investigated the similarity of the embeddings of pre-formulated sentences, meaning that the sentences, and in particular the paraphrases of the metonymic expressions, processed by RoBERTa were taken from a human-generated corpus. However, since said metonymic sentences were also analysed and interpreted in the first experiment of this project, it would make sense to employ the paraphrases generated by the model and feed them again to the model in order to observe its behaviour. Namely, the embeddings of each of the five returned solutions could be firstly compared to the original metonymic sentence and, secondly, compared the embeddings among the solutions.

Other than the computational approach, another feature could be comprehended in a similar investigation: the inclusion of a psycholinguistic study in order to understand how the human brain interprets metonymic instances could offer some insights on which connections between the form in which the metonymic expression appears and the intended plausible referents are establish. Then, the human performance could be compared to the behaviour of transformer language model.

Appendix 1 – Accuracy

Overall accuracy of the models:

- strategy 1

solution	BERT_base	BERT_large	RoBERTa_base	RoBERTa_large
1	0.335952849	0.422396857	0.467583497	0.526522593
2	0.342829077	0.430255403	0.452848723	0.485265226
3	0.350360183	0.411918795	0.440733464	0.47151277
4	0.3521611	0.406679764	0.428290766	0.461689587
5	0.354420432	0.4	0.420825147	0.458939096

- strategy 2

solution	BERT_base	BERT_large	RoBERTa_base	RoBERTa_large
1	0.127092338	0.194302554	0.246208251	0.260176817
2	0.140903733	0.190265226	0.239096267	0.239950884
3	0.149882122	0.1902685	0.233346431	0.22632613
4	0.15490668	0.189842829	0.226640472	0.222765226
5	0.158954813	0.187972495	0.225312377	0.221897839

- strategy 3

solution	BERT_base	BERT_large	RoBERTa_base	RoBERTa_large
1	0.123772102	0.218074656	0.294695481	0.310412574
2	0.13654224	0.211198428	0.28978389	0.286836935
3	0.143418468	0.209561231	0.27832351	0.266535691
4	0.151768173	0.205304519	0.265717092	0.260805501
5	0.157170923	0.204715128	0.258939096	0.253045187

Accuracy for CONTAINER-FOR-CONTENT metonymies:

- strategy 1

solution	BERT_base	BERT_large	RoBERTa_base	RoBERTa_large
1	0.505617978	0.662921348	0.752808989	0.606741573
2	0.533707865	0.685393258	0.702247191	0.556179775
3	0.573033708	0.666666667	0.666666667	0.498127341
4	0.575842697	0.643258427	0.637640449	0.480337079
5	0.564044944	0.62247191	0.615730337	0.46741573

- strategy 2

solution	BERT_base	BERT_large	RoBERTa_base	RoBERTa_large
1	0.235393258	0.316067416	0.303932584	0.233707865
2	0.248089888	0.313651685	0.301235955	0.229438202
3	0.269775281	0.315393258	0.29	0.205842697
4	0.267022472	0.300730337	0.289044944	0.196938202
5	0.273191011	0.302337079	0.288561798	0.204921348

- strategy 3

solution	BERT_base	BERT_large	RoBERTa_base	RoBERTa_large
1	0.292134831	0.528089888	0.528089888	0.426966292
2	0.269662921	0.449438202	0.494382022	0.365168539
3	0.277153558	0.426966292	0.453183521	0.325842697
4	0.283707865	0.404494382	0.438202247	0.311797753
5	0.28988764	0.406741573	0.420224719	0.307865169

Accuracy for PRODUCER-FOR-PRODUCT metonymies:

- strategy 1

solution	BERT_base	BERT_large	RoBERTa_base	RoBERTa_large
1	0.154545455	0.3	0.372727273	0.472727273
2	0.209090909	0.345454545	0.381818182	0.427272727
3	0.23030303	0.345454545	0.366666667	0.427272727
4	0.231818182	0.359090909	0.35	0.429545455
5	0.247272727	0.36	0.347272727	0.421818182

- strategy 2

solution	BERT_base	BERT_large	RoBERTa_base	RoBERTa_large
1	0.080181818	0.202727273	0.253	0.305818182
2	0.112909091	0.212363636	0.251772727	0.272590909
3	0.118090909	0.215090909	0.24930303	0.261393939
4	0.122772727	0.228431818	0.232409091	0.265545455
5	0.133745455	0.220436364	0.231109091	0.254927273

- strategy 3

solution	BERT_base	BERT_large	RoBERTa_base	RoBERTa_large

1	0.090909091	0.190909091	0.245454545	0.327272727
2	0.1	0.204545455	0.236363636	0.295454545
3	0.093939394	0.215151515	0.236363636	0.281818182
4	0.102272727	0.227272727	0.218181818	0.284090909
5	0.112727273	0.22	0.214545455	0.270909091

Accuracy for PRODUCT-FOR-PRODUCER metonymies:

- strategy 1

solution	BERT_base	BERT_large	RoBERTa_base	RoBERTa_large
1	0.70212766	0.531914894	0.553191489	0.531914894
2	0.563829787	0.489361702	0.457446809	0.563829787
3	0.489361702	0.460992908	0.468085106	0.581560284
4	0.468085106	0.473404255	0.436170213	0.558510638
5	0.45106383	0.463829787	0.438297872	0.557446809

- strategy 2

solution	BERT_base	BERT_large	RoBERTa_base	RoBERTa_large
1	0.192340426	0.15787234	0.28787234	0.311489362
2	0.166489362	0.131276596	0.246382979	0.311489362
3	0.151631206	0.139716312	0.238156028	0.288014184
4	0.149308511	0.156010638	0.222606383	0.277340426
5	0.14693617	0.169234043	0.223787234	0.278765957

- strategy 3

solution	BERT_base	BERT_large	RoBERTa_base	RoBERTa_large
1	0.063829787	0.085106383	0.234042553	0.29787234
2	0.074468085	0.063829787	0.20212766	0.319148936
3	0.070921986	0.070921986	0.212765957	0.304964539
4	0.074468085	0.095744681	0.20212766	0.303191489
5	0.080851064	0.106382979	0.217021277	0.285106383

Accuracy for LOCATION-FOR-LOCATED metonymies:

- strategy 1

solution	BERT_base	BERT_large	RoBERTa_base	RoBERTa_large
1	0.414893617	0.446808511	0.510638298	0.64893617

2	0.414893617	0.478723404	0.5	0.579787234
3	0.32587234	0.355361702	0.43806383	0.503212766
4	0.425531915	0.444148936	0.497340426	0.563829787
5	0.417021277	0.446808511	0.50212766	0.565957447

- strategy 2

solution	BERT_base	BERT_large	RoBERTa_base	RoBERTa_large
1	0.182553191	0.223617021	0.328617021	0.36712766
2	0.191968085	0.220531915	0.306329787	0.310265957
3	0.24312766	0.268340426	0.351553191	0.369106383
4	0.21087766	0.199680851	0.29731383	0.2875
5	0.20387234	0.195212766	0.300829787	0.282659574

- strategy 3

solution	BERT_base	BERT_large	RoBERTa_base	RoBERTa_large
1	0.191489362	0.223404255	0.35106383	0.457446809
2	0.223404255	0.25	0.382978723	0.404255319
3	0.26793617	0.281787234	0.380212766	0.371957447
4	0.255319149	0.210106383	0.356382979	0.356382979
5	0.24893617	0.210638298	0.35106383	0.340425532

Accuracy for CAUSER-FOR-RESULT metonymies

- strategy 1

solution	BERT_base	BERT_large	RoBERTa_base	RoBERTa_large
1	0.206521739	0.369565217	0.445652174	0.608695652
2	0.206521739	0.342391304	0.451086957	0.538043478
3	0.217391304	0.322463768	0.423913043	0.518115942
4	0.230978261	0.298913043	0.407608696	0.505434783
5	0.245652174	0.289130435	0.386956522	0.506521739

- strategy 2

solution	BERT_base	BERT_large	RoBERTa_base	RoBERTa_large
1	0.054130435	0.126847826	0.192934783	0.200326087
2	0.061086957	0.123369565	0.180380435	0.17875
3	0.062391304	0.120217391	0.167246377	0.173514493

4	0.06798913	0.110706522	0.155896739	0.172690217
5	0.074543478	0.1055	0.145043478	0.175130435

- strategy 3

solution	BERT_base	BERT_large	RoBERTa_base	RoBERTa_large
1	0.054347826	0.130434783	0.27173913	0.184782609
2	0.065217391	0.135869565	0.239130435	0.184782609
3	0.061594203	0.137681159	0.217391304	0.188405797
4	0.0625	0.127717391	0.195652174	0.179347826
5	0.07173913	0.123913043	0.173913043	0.180434783

Accuracy for POSSESSED-FOR-POSSESSOR metonymies:

- strategy 1

solution	BERT_base	BERT_large	RoBERTa_base	RoBERTa_large
1	0.233766234	0.285714286	0.194805195	0.25974026
2	0.253246753	0.266233766	0.207792208	0.25974026
3	0.255411255	0.251082251	0.212121212	0.268398268
4	0.25	0.243506494	0.233766234	0.25
5	0.25974026	0.236363636	0.231168831	0.254545455

- strategy 2

solution	BERT_base	BERT_large	RoBERTa_base	RoBERTa_large
1	0.048571429	0.108571429	0.107402597	0.135194805
2	0.074415584	0.095064935	0.132792208	0.149090909
3	0.095238095	0.10974026	0.132683983	0.144588745
4	0.110162338	0.10974026	0.146980519	0.138993506
5	0.116285714	0.110545455	0.148571429	0.141324675

- strategy 3

solution	BERT_base	BERT_large	RoBERTa_base	RoBERTa_large
1	0.012987013	0.077922078	0.090909091	0.12987013
2	0.051948052	0.077922078	0.12987013	0.142857143
3	0.082251082	0.0995671	0.12987013	0.125541126

4	0.097402597	0.097402597	0.146103896	0.123376623
5	0.103896104	0.098701299	0.150649351	0.124675325

Appendix 2 – Cosine similarity and measure of the performance

CONTAINER-FOR-CONTENT metonymies

	1	2	3	4	5	6	7	8
met_lit	0.925	0.914	0.913	0.896	0.892	0.892	0.905	0.913
met_para	0.544	0.65	0.691	0.647	0.686	0.719	0.777	0.82
lit_para	0.546	0.655	0.696	0.652	0.683	0.714	0.773	0.816
normalised	-0.698	-0.403	-0.32	-0.382	-0.302	-0.242	-0.165	-0.115
not_normal	-0.381	-0.264	-0.223	-0.249	-0.207	-0.173	-0.128	-0.094

9	10	11	12	13	14	15	16
0.920	0.916	0.908	0.905	0.899	0.892	0.88	0.882
0.853	0.865	0.856	0.857	0.850	0.842	0.825	0.83
0.852	0.865	0.859	0.86	0.855	0.849	0.841	0.849
-0.079	-0.058	-0.06	-0.056	-0.057	-0.059	-0.066	-0.061
-0.067	-0.050	-0.052	-0.048	-0.049	-0.05	-0.055	-0.052

17	18	19	20	21	22	23	24
0.886	0.885	0.886	0.864	0.848	0.877	0.881	0.979
0.839	0.841	0.848	0.828	0.807	0.845	0.853	0.975
0.857	0.86	0.863	0.842	0.813	0.852	0.861	0.977
-0.055	-0.052	-0.045	-0.043	-0.05	-0.038	-0.033	-0.005
-0.047	-0.045	-0.039	-0.036	-0.041	-0.032	-0.028	-0.005

PRODUCER-FOR-PRODUCT metonymies

	1	2	3	4	5	6	7	8
met_lit	0.928	0.914	0.915	0.896	0.898	0.906	0.925	0.936
met_para	0.632	0.699	0.739	0.705	0.719	0.763	0.812	0.857
lit_para	0.628	0.698	0.737	0.699	0.719	0.767	0.822	0.868
normalised	-0.473	-0.307	-0.238	-0.273	-0.248	-0.186	-0.137	-0.091
not-normal	-0.297	-0.214	-0.176	-0.191	-0.178	-0.143	-0.112	-0.079

9	10	11	12	13	14	15	16
0.951	0.953	0.952	0.948	0.948	0.947	0.946	0.95
0.896	0.909	0.91	0.907	0.905	0.9	0.899	0.904

0.909	0.923	0.926	0.92	0.921	0.915	0.915	0.919
-0.062	-0.047	-0.045	-0.045	-0.047	-0.052	-0.052	-0.05
-0.056	-0.043	-0.042	-0.042	-0.043	-0.047	-0.047	-0.046

17	18	19	20	21	22	23	24
0.954	0.956	0.956	0.942	0.917	0.945	0.953	0.993
0.91	0.91	0.909	0.888	0.85	0.89	0.906	0.985
0.923	0.923	0.92	0.902	0.866	0.897	0.912	0.986
-0.048	-0.049	-0.051	-0.06	-0.077	-0.062	-0.051	-0.007
-0.044	-0.045	-0.047	-0.054	-0.067	-0.055	-0.047	-0.007

PRODUCT-FOR-PRODUCER metonymies

	1	2	3	4	5	6	7	8
met_lit	0.872	0.878	0.886	0.869	0.867	0.878	0.898	0.914
met_para	0.628	0.706	0.742	0.727	0.743	0.782	0.822	0.854
lit_para	0.622	0.697	0.731	0.713	0.733	0.773	0.812	0.848
normalised	-0.393	-0.248	-0.197	-0.199	-0.169	-0.124	-0.094	-0.07
not-normal	-0.244	-0.173	-0.144	-0.142	-0.124	-0.096	-0.076	-0.06

9	10	11	12	13	14	15	16
0.928	0.931	0.925	0.92	0.917	0.912	0.909	0.908
0.883	0.893	0.886	0.878	0.872	0.865	0.843	0.841
0.875	0.889	0.88	0.87	0.865	0.861	0.845	0.85
-0.052	-0.042	-0.044	-0.048	-0.052	-0.055	-0.078	-0.079
-0.045	-0.037	-0.039	-0.042	-0.045	-0.047	-0.066	-0.067

17	18	19	20	21	22	23	24
0.909	0.909	0.911	0.901	0.89	0.909	0.91	0.986
0.853	0.859	0.868	0.855	0.844	0.873	0.883	0.982
0.855	0.858	0.865	0.852	0.833	0.871	0.879	0.981
-0.066	-0.059	-0.049	-0.0534	-0.055	-0.041	-0.031	-0.004
-0.056	-0.05	-0.043	-0.046	-0.045	-0.036	-0.027	-0.004

LOCATION-FOR-LOCATED metonymies

	1	2	3	4	5	6	7	8
met_lit	0.864	0.873	0.881	0.865	0.868	0.87	0.886	0.897

met_para	0.562796	0.660718	0.707614	0.682745	0.711162	0.738758	0.786516	0.817685
lit_para	0.608	0.688	0.731	0.711	0.738	0.771	0.814	0.84
normalised	-0.495	-0.308	-0.238	-0.256	-0.212	-0.171	-0.122	-0.094
not-norm	-0.301	-0.212	-0.174	-0.182	-0.157	-0.131	-0.099	-0.079

9	10	11	12	13	14	15	16
0.904	0.905	0.889	0.887	0.877	0.87	0.849	0.846
0.845	0.86	0.842	0.842	0.834	0.827	0.79	0.785
0.864	0.877	0.863	0.862	0.851	0.847	0.817	0.813
-0.068	-0.052	-0.054	-0.052	-0.051	-0.05	-0.072	-0.074
-0.059	-0.046	-0.047	-0.045	-0.043	-0.043	-0.059	-0.06

17	18	19	20	21	22	23	24
0.857	0.858	0.864	0.852	0.854	0.865	0.866	0.981
0.803	0.807	0.817	0.806	0.807	0.833	0.842	0.976
0.832	0.836	0.842	0.833	0.832	0.847	0.854	0.979
-0.065	-0.06	-0.056	-0.056	-0.057	-0.038	-0.028	-0.005
-0.054	-0.05	-0.047	-0.046	-0.047	-0.032	-0.024	-0.004

CAUSER-FOR-RESULT metonymies

	1	2	3	4	5	6	7	8
met_lit	0.912	0.908	0.916	0.904	0.903	0.904	0.912	0.917
met_para	0.623	0.711	0.744	0.717	0.727	0.754	0.799	0.834
lit_para	0.626	0.712	0.747	0.722	0.733	0.762	0.805	0.837
normalised	-0.461	-0.276	-0.23	-0.259	-0.24	-0.197	-0.14	-0.1
not-norm	-0.288	-0.196	-0.171	-0.187	-0.176	-0.15	-0.113	-0.084

9	10	11	12	13	14	15	16
0.922	0.919	0.909	0.907	0.897	0.892	0.88	0.877
0.854	0.868	0.855	0.852	0.839	0.833	0.815	0.815
0.858	0.871	0.859	0.857	0.845	0.841	0.825	0.826
-0.079	-0.058	-0.063	-0.064	-0.07	-0.071	-0.079	-0.076
-0.068	-0.05	-0.055	-0.055	-0.059	-0.06	-0.065	-0.062

17	18	19	20	21	22	23	24
0.884	0.887	0.889	0.875	0.869	0.88	0.877	0.977
0.826	0.832	0.838	0.823	0.812	0.84	0.843	0.974
0.836	0.842	0.847	0.833	0.823	0.845	0.845	0.974
-0.07	-0.065	-0.06	-0.062	-0.068	-0.048	-0.041	-0.003
-0.058	-0.055	-0.051	-0.052	-0.056	-0.041	-0.035	-0.003

POSSESSED-FOR-POSSESSOR metonymies

	1	2	3	4	5	6	7	8
met_lit	0.83	0.829	0.845	0.829	0.838	0.84	0.856	0.87
met_para	0.556	0.648	0.696	0.686	0.714	0.738	0.78	0.807
lit_para	0.545	0.636	0.689	0.678	0.712	0.746	0.786	0.813
normalised	-0.504	-0.284	-0.216	-0.211	-0.175	-0.136	-0.097	-0.078
not-norm	-0.274	-0.181	-0.149	-0.143	-0.125	-0.102	-0.076	-0.067

9	10	11	12	13	14	15	16
0.879	0.878	0.865	0.862	0.85	0.84	0.813	0.806
0.822	0.834	0.816	0.817	0.801	0.794	0.761	0.759
0.831	0.844	0.828	0.828	0.815	0.809	0.774	0.7
-0.069	-0.052	-0.059	-0.054	-0.06	-0.057	-0.067	-0.061
-0.057	-0.044	-0.049	-0.045	-0.049	-0.046	-0.052	-0.047

17	18	19	20	21	22	23	24
0.817	0.825	0.834	0.815	0.819	0.835	0.834	0.971
0.776	0.782	0.796	0.778	0.79	0.81	0.816	0.966
0.787	0.794	0.807	0.796	0.805	0.827	0.828	0.968
-0.052	-0.054	-0.047	-0.047	-0.036	-0.03	-0.021	-0.005
-0.041	-0.043	-0.038	-0.037	-0.029	-0.025	-0.017	-0.004

Bibliography

Annaz, D., Van Herwegen, J., Thomas, M. S.C., Fishman, R., Karmiloff-Smith, A., and Rundblad, G. (2008). The comprehension of metaphor and metonymy in children with Williams syndrome. *International Journal of Language and Communication Disorders*, 44 (6): 962–78.

Barcelona, A. (2003). Metonymy in cognitive linguistics: an analysis and a few modest proposals. In H. Cuyckens, Th. Berg, R. Dirven and K.-U. Panther (eds.) *Motivation in Language. Studies in Honour of Gunter Radden*. Amsterdam: Benjamins, 223–55.

Barcelona, A. (2011). Reviewing the properties and prototype structure of metonymy. In R. Benczes, A. Barcelona and F. J. Ruiz de Mendoza Ibáñez, (eds*.) Defining Metonymy in Cognitive Linguistics: Towards a Consensus View.* Amsterdam: John Benjamins, 7–57.

Bengio, Y., Ducharme, R., Vincent, P. & Jauvin, C. (2000). A neural probabilistic language model. *Advances in neural information processing systems*, 13.

Biernacka, E. (2013). 'The role of metonymy in political discourse. Unpublished PhD thesis, Milton Keynes: The Open University.

Bird, S., Loper, E., and Klein, E. (2009). Natural Language Processing with Python. O'Reilly Media Inc.

Brdar, M., & Brdar-Szabó, R. (2013). Translating (by means of) metonymy. Cognitive linguistics and translation: Advances in some theoretical models and applications, 199-226.

Brun, C., Ehrmann, M., & Jacquet, G. (2007). XRCE-M: A hybrid system for named entity metonymy resolution. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)* (pp. 488-491).

Burn, J. (1986). Williams syndrome. *Journal of Medical Genetics*, 23(5), 389.

Burnard, L. (1995). *The Users Reference Guide for the British National Corpus*. British National Corpus Consortium, Oxford University Computing Service, Oxford, U.K.

Chersoni, E., Santus, E., Pannitto, L., Lenci, A., Blache, P., & Huang, C. R. (2019). A structured distributional model of sentence meaning and processing. *Natural Language Engineering*, 25(4), 483-502.

Chollet, F. (2021). *Deep learning with Python*. Simon and Schuster.

Chomsky, N. (1988). Generative grammar. *Studies in English linguistics and literature*.

Coulson, S., and Oakley, T. (2003). Metonymy and conceptual blending. In K.-U. Panther and L. Thornburg (eds.) *Metonymy and Pragmatic Inferencing.* Amsterdam: John Benjamins, 51–79.

Croft, W., & Cruse, D. A. (2004). *Cognitive linguistics*. Cambridge University Press.

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

Elman, J.L. (2009). On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive Science*, 33(4): 547–582.

Evans, V. (2012). Cognitive linguistics. *Wiley Interdisciplinary Reviews: Cognitive Science*, 3(2), 129-141.

Fauconnier, G., and Turner, M. (1999). Metonymy and conceptual integration. In K.-U. Panther and G. Radden (eds.) *Metonymy in Language and Thought.* Amsterdam: John Benjamins, 77–90.

Fellbaum, C. (1998). WordNet: An electronic lexical database. Cambridge, MA: MIT Press.

Frisson, S., and Pickering, M. J. (1999). The processing of metonymy: evidence from eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25: 1366–83.

Gibbs Jr, R. W., Gibbs, R. W., & Gibbs, J. (1994). *The poetics of mind: Figurative thought, language, and understanding.* Cambridge University Press.

Graves, A. (2013). Generating sequences with recurrent neural networks. *arXiv preprint arXiv*:1308.0850.

Grice, H. P. (1975). Logic and conversation. In *Speech acts* (pp. 41-58). Brill.

Gritta, M., Pilehvar, M. T., Limsopatham, N., & Collier, N. (2017). Vancouver welcomes you! minimalist location metonymy resolution. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1248-1259).

Guresen, E., & Kayakutlu, G. (2011). Definition of artificial neural networks with comparison to other networks. *Procedia Computer Science*, 3, 426-433.

Handl, S. (2011). *The Conventionality of Figurative Language: A Usage-Based Study.* Tubingen: Narr Verlag.

Harabagiu, S. (1998). Deriving metonymic coercions from WordNet. In *Usage of WordNet in Natural Language Processing Systems*.

Harabagiu, S. M. (2008). Questions and intentions. *Advances in Open Domain Question Answering*, 99-147.

Hewitt, J. & Manning, C.D. (2019). A Structural Probe for Finding Syntax in Word Representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4129–4138.

Hilpert, M. (2006). Keeping an eye on the data: Metonymies and their patterns. *Trends in linguistics studies and monographs*, 171, 123. In Stefanowitsch, A., & Gries, S. T. (Eds.). (2007). *Corpus-based approaches to metaphor and metonymy* (Vol. 171). Walter de Gruyter.

Illina, I., & Fohr, D. (2023). Semantic Information Investigation for Transformer-based Rescoring of N-best Speech Recognition. In *LTC 2023*.

Jurafsky, D. & Martin, J. H. (2023). *Speech and Language Processing*.

Kadan, A., Gangan, M. P., & Abraham, S. S. (2023). REDAffectiveLM: Leveraging Affect Enriched Embedding and Transformer-based Neural Language Model for Readers' Emotion Detection. *arXiv preprint arXiv:2301.08995*.

Kamei, S. I., & Wakao, T. (1992). Metonymy: Reassessment, survey of acceptability, and its treatment in a machine translation system. In *30th annual meeting of the association for computational linguistics* (pp. 309-311).

Kienpointner, M. (2011). Figures of speech. *Discursive pragmatics*, 102-119.

Kövecses, Z. (2006). *Language, Mind and Culture: A Practical Introduction*. Oxford University Press.

Lakoff, G. (1987). *Women, Fire and Dangerous Things: What Categories Reveal about the Mind.* University of Chicago Press.

Lakoff, G., & Johnson, M. (1980). The metaphorical structure of the human conceptual system. *Cognitive science*, 4(2), 195-208.

Lakoff, G., & Johnson, M. (2008). *Metaphors we live by.* University of Chicago press.

Langacker, R.W. (1987). *Foundations of Cognitive Grammar*, Vol. I: *Theoretical Prerequisites*. Stanford University Press.

Langacker, R.W. (1993). Reference-point constructions. *Cognitive Linguistics* 4: 1–38.

Lapata, M., & Lascarides, A. (2003). A probabilistic account of logical metonymy. *Computational Linguistics*, 29(2), 261-315.

Larsen-Freeman, D., and Cameron, L. (2008). *Complex Systems and Applied Linguistics*. Oxford University Press.

Lenci, A. (2011). Composing and updating verb argument expectations: A distributional semantic model. In *Proceedings of the 2nd workshop on cognitive modeling and computational linguistics* (pp. 58-66).

Li, H., Vasardani, M., Tomko, M., & Baldwin, T. (2020). Target word masking for location metonymy resolution. *arXiv preprint arXiv*:2010.16097.

Li, C., Zhang, J., Li, B., & Xu, Z. (2023). Key Information Extraction Method Study for Road Traffic Accidents via Integration of Rules and SkipGram-BERT. *In Advances in Intelligent Systems, Computer Science and Digital Economics IV* (pp. 658-672).

Littlemore, J. (2015). *Metonymy*. Cambridge University Press.

Littlemore, J., May, A., & Arizono, S. (2018). The interpretation of metonymy by Japanese learners of English. *Applying Cognitive Linguistics: Figurative language in use, constructions and typology*, 99, 51.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv*:1907.11692.

Lord, C., Elsabbagh, M., Baird, G., & Veenstra-Vanderweele, J. (2018). Autism spectrum disorder. *The lancet*, 392(10146), 508-520.

Lowder, M.W., and Gordon, P. C. (2013). It's hard to offend the college: effects of sentence structure on figurative-language processing. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 39 (4): 993–1011.

Manning, C. D. (2008). *Introduction to information retrieval*. Syngress Publishing.

Markert, K., & Hahn, U. (2002). Understanding metonymies in discourse. *Artificial intelligence*, 135(1-2), 145-198.

Markert, K., & Nissim, M. (2002). Metonymy resolution as a classification task. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)* (pp. 204-213).

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5, 115-133.

McKinney, W. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference* (Vol. 445, No. 1, pp. 51-56).

McRae, K., & Matsuki, K. (2009). People use their knowledge of common events to understand language, and do so as quickly as possible. *Language and linguistics compass*, 3(6), 1417-1429.

Mikolov, T., Karafiát, M., Burget, L., Cernocký, J., & Khudanpur, S. (2010). Recurrent neural network based language model. In I*nterspeech* (Vol. 2, No. 3, pp. 1045-1048).

Nerlich, B., Clarke, D., and Todd, Z. (1999). 'Mummy, I like being a sandwich': metonymy in language acquisition. In K.-U. Panther and G. Radden (eds.) *Metonymy in Language and Thought.* Amsterdam: John Benjamins, 361–84.

Nguyen, D. T., & Tran, T. (2023). Natural language generation from Universal Dependencies using data augmentation and pre-trained language models. *International Journal of Intelligent Information and Database Systems*, 16(1), 89-105.

Nissim, M., & Markert, K. (2003). Syntactic features and word similarity for supervised metonymy resolution. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics* (pp. 56-63).

Markert, K., & Nissim, M. (2007). Metonymy resolution at SemEval I: Guidelines for participants. In *Proceedings of the ACL 2007 Conference*.

Panther, K.-U., and Thornburg, L. (1998). A cognitive approach to inferencing in conversation. *Journal of Pragmatics*, 30: 755–69.

Pedinotti, P., & Lenci, A. (2020). Don't invite BERT to drink a bottle: Modeling the interpretation of metonymies using BERT and distributional representations. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 6831-6837).

Peirsman, Y., and Geeraerts, D. (2006). Metonymy as a prototypical category. *Cognitive Linguistics*, 17 (3): 269–316.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Radden, G., & Kövecses, Z. (1999). Towards a theory of metonymy. *Metonymy in language and thought*, *4*, 17-60.

Rambelli, G., Chersoni, E., Lenci, A., Blache, P., & Huang, C. R. (2020). Comparing probabilistic, distributional and transformer-based models on logical metonymy interpretation. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing (AACL-IJCNLP).*

Rapp, A. M., Erb M., Grodd, W., Bartels, M., and Markert, K. (2011). Neurological correlates of metonymy resolution. *Brain and Language*, 119 (3): 196–205.

Ray, S. (2019). A quick review of machine learning algorithms. In *2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon)* (pp. 35-39).

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8, 842-866.

Ruiz de Mendoza Ibáñez, F. J., and Diez Velasco, O. I. (2002). Patterns of conceptual interaction. In R. Dirven and R. Porings (eds.) (2003). *Metaphor and Metonymy in Comparison and Contrast.* Berlin and New York: Mouton de Gruyter, 489–532.

Rundblad, G., and Annaz, D. (2010a). Metaphor and metonymy comprehension: receptive vocabulary and conceptual knowledge. *British Journal of Developmental Psychology*, 28: 547–63.

Rundblad, G., and Annaz, D. (2010b). The atypical development of metaphor and metonymy comprehension in children with autism. *Autism*, 14 (1): 29–46.

Sperber, D., and Wilson, D. (1987). Precis of relevance: communication and cognition. *Behavioral and Brain Sciences*, 10, 697–754.

Sperber, D., & Wilson, D. (2002). Pragmatics, modularity and mind-reading. *Mind & language*, 17(1-2), 3-23.

Sperber, D., and Wilson, D. (2004). Relevance theory. In G. Ward and L. Horn (eds.) *Handbook of Pragmatics.* Oxford: Blackwell, 607–32.

Stallard, D. (1993). Two kinds of metonymy. In *31st Annual Meeting of the Association for Computational Linguistics* (pp. 87-94).

Steen, G. (2009). Deliberate metaphor affords conscious metaphorical cognition. *Cognitive Semiotics*, 5(1-2), 179-197.

Tager-Flusberg, H. (2006). Defining Language Phenotypes in Autism, *Clinical Neuroscience Research* 6: 219–24.

Tenney, I., Das, D., & Pavlick, E. (2019). BERT rediscovers the classical NLP pipeline. *arXiv preprint arXiv:1905.05950*.

Utt, J., Lenci, A., Padó, S., & Zarcone, A. (2013). The curious case of metonymic verbs: A distributional characterization. In *IWCS 2013 Workshop Towards a Formal Distributional Semantics* (pp. 30-39). ACL (Association for Computational Linguistics).

Van Herwegen, J., Dimitrious, D. and Rundblad, G. (2013). Development of novel metaphor and metonymy comprehension in typically developing children and Williams syndrome. *Research in Developmental Disabilities*, 34 (4): 1300–11.

Warren, B. (1999). Aspects of referential metonymy. In K.-U. Panther and G. Radden (eds.) *Metonymy in Language and Thought*. Amsterdam: John Benjamins, 121–37.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations* (pp. 38-45).

Wu, Z., Chen, Y., Kao, B., and Liu, Q. (2020). Perturbed Masking: Parameter-free Probing for Analyzing and Interpreting BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4166–4176.

Zhang, S., Fan, R., Liu, Y., Chen, S., Liu, Q., & Zeng, W. (2023). Applications of Transformer-based Language Models in Bioinformatics: A Survey. *Bioinformatics Advances.*

Zhao, W. (2014). Anaphora Resolution in Stream-of-Consciousness Discourse: A Metonymic Account*. International Journal of English Linguistics, 4(3), 34.*

Zhi, C. (2020). An empirical study of the scope of Spanish to Chinese machine translation. A GNMT case study concerning metaphorical and metonymic expressions. Círculo de Lingüistica Aplicada a la Comunicación, (83), 1-25.

Zarcone, A., Utt, J., & Padó, S. (2012). Modeling covert event retrieval in logical metonymy: probabilistic and distributional accounts. In *Proceedings of the 3rd Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2012)* (pp. 70-79).

Acknowledgements

At the end of this thesis, it seems to be only fair to thank those who through their support made this achievement possible.

Firstly, I'd like to thank the supervisor of this thesis, Prof. Lebani, because it was only thanks to his guidance, suggestions, and support that I managed to complete this project. I am truly grateful for all the time you dedicated to guide me through this work. Along with the professor, I'd like to mention Dr. Dall'Igna: I cannot express how thankful I am for all the "technical" support you offered me and for not blocking me on WhatsApp, despite all the times I annoyed you with my messages.

Secondly, I'd like to thank my family who helped me becoming who I am today: you showed me the importance of putting effort into whatever I am doing as well as the meaning of unconditional love. A special mention goes to my parents: I will be forever grateful for how you believed in me from the very beginning and supported me in any decision, even when this meant letting me go. You have given me the strength to reach my goals.

Last but not least, my gratitude goes to my friends, those who were there from the beginning and those that I had the chance to meet along the way. So, I wish to say a huge thank you to my main support system: Andrea, Elisabetta, Francesca, Francesca, Gaia, Mario, Sara, Sara, and Teresa. Anyone who has you in their life can be considered lucky: you never failed to show me love and support despite distance and different life paths and for that I will never thank you enough. I'd like to mention a few great addictions as well I met during my time in Konstanz: Carmen, Giulia, Maria, Melissa, and Valeria. Thanks to them, a foreign country did not seem so "foreign", instead they made Konstanz feel like a second home.