Università
Ca'Foscari
Venezia

Corso di Laurea
magistrale

in Scienze del
Linguaggio

Doppio Diploma in Italienstudien con Goethe Universität Frankfurt am Main

Tesi di Laurea

A corpus analysis of femi(ni)cide narration in Italy and Germany in 2019

Relatore
Ch.mo Prof. Gianluca Lebani

Supervisore Tesi
Ch.ma Prof.ssa Cecilia Poletto (Goethe-Universität Frankfurt am Main)

Supervisore dell'attività svolta all'estero
Ch.ma Prof.ssa Irene Caloi (Goethe-Universität Frankfurt am Main)
Ch.mo Prof. Roland Hinterhölzl

Laureanda
Emilia Milano
Matricola 883068

Anno Accademico
2020 / 2021

Table of contents

Acknowledgments

I would like to thank my supervisor, Professor Gianluca Lebani, who spent a lot of time and patience trying to lead me through this path and helped me realising this project since the real beginning. I also would like to thank my assistant supervisor, Professor Cecilia Poletto, who came to my aid when I needed her. I wish to thank *Feminizidmap* group, which provided me with the lists of German femi(ni)cide cases and with some precious advice, and Dr Lucia Busso for her help and explanations. Finally, all my gratitude goes to my family and friends, who supported me and encouraged me during the whole Master.

Abstract

87.000 è il numero di donne uccise nel 2017 presente nel report delle Nazioni Unite (United Nations Office on Drugs and Crime [UNODC], 2018) riguardo l'incidenza del femminicidio del mondo. Di queste donne, 50.000 sono state uccise da partner o da familiari. Secondo lo stesso studio, il numero di donne uccise da partner o familiari nel 2012 è di 48.000, un dato che sembra destinato a crescere sempre di più. Questi omicidi legati al genere – *gender related killings* – sono registrati in tutti e cinque i continenti, con un'incidenza superiore in Africa – 3,1 donne ogni 100.000 – mentre il tasso stimato per l'Europa risulta essere di molto inferiore– 0,7 donne ogni 100.000. La violenza all'interno della coppia e l'omicidio che ne può seguire non sono sempre da considerarsi come legati alla violenza di genere, ma possono anche essere legati anche alla violenza domestica. La differenza riscontrata fra i due tipi di violenza è che i casi di femminicidio sono preceduti da episodi di violenza di genere, legata alla percezione (anche inconsapevole) della donna come qualcosa di proprio, influenzato dal diverso potere che il genere maschile e quello femminile hanno nella cultura occidentale. È proprio questo suo essere legato a dinamiche di potere e culturali e non a conflitti occasionali a far sì che il fenomeno del femminicidio sia un fenomeno in aumento (Johnson, 2015).

Il fenomeno del femminicidio, e della violenza di genere in generale, è solitamente considerato come un fenomeno universale, condiviso da diverse società e culture. Risulta quindi interessante analizzare se anche la sua narrazione può essere considerata universale. Il seguente lavoro si prefigura di analizzare la narrazione del fenomeno del femminicidio nella stampa italiana e tedesca nel 2019 attraverso metodi di linguistica computazionale.

I due Paesi di riferimento di questo studio sono culture che presentano delle differenze ma anche diversi punti in comune, probabilmente condizionati anche dalla vicinanza geografica, dalla moneta comune e dalla comune appartenenza all'Unione Europea. Ci si aspetta quindi di non trovare differenze sostanziali nella narrazione del fenomeno, e che le differenze riscontrate non siano attribuibili a una diversa percezione del fenomeno o a forti differenze culturali, ma a diverse caratteristiche linguistiche.

L'analisi linguistica è stata svolta su un corpus bilingue creato appositamente per questo studio e formato da articoli pubblicati su cinque giornali italiani e su cinque giornali tedeschi. I giornali italiani sono *La Repubblica*, *La Stampa*, *Il Giornale*, *Il Fatto Quotidiano*, e *Il Resto del Carlino* e i giornali tedeschi sono *Die Zeit*, *Süddeutsche Zeitung*, *Die Welt*, *Frankfurter Allgemeine Zeitung*, *Bild*. Dalla versione online di questi

giornali, sono stati scaricati articoli pubblicati nel 2019 e riguardanti casi di femminicidio avvenuti nel 2019. Questi casi sono stati individuati grazie alle liste di casi di femminicidio avvenuti in Italia forniti dalla *Casa delle donne per non subire violenza, ONLUS* di Bologna e dall'*Osservatorio sul femminicidio* di *ProsMedia*, e dalla lista di casi avvenuti in Germania fornita da *Feminizidmap.* Per creare il corpus, sono stati collezionati in un file di testo gli indirizzi URL degli articoli selezionati e inseriti in BootCat, un programma per la creazione di corpora basati su documenti collezionati da Internet. I metadati associati ai documenti del corpus includono la data di pubblicazione dell'articolo, il giornale e la sezione di appartenenza e alcuni dati per individuare il caso, come il nome della vittima e la relazione che il colpevole aveva con la vittima. Il corpus che ne risulta è un corpus non bilanciato, in quanto il sotto corpus tedesco è formato da una quantità di documenti maggiore rispetto al sotto corpus italiano. Questo sbilanciamento fra le dimensioni dei due sotto corpora è causato dal modo in cui si è deciso di raccogliere il materiale, partendo da liste di vittime di femminicidio che sono state registrate nei due Paesi nel 2019.

Le analisi eseguite sul corpus sono, principalmente, l'analisi dei topic ricorrenti e delle metafore usate per parlare di casi di cronaca classificati come casi di femminicidio.

L'algoritmo usato per la parte dello studio dedicata al topic modeling, quindi all'analisi dei topic ricorrenti nelle due sezioni del corpus, è l'algoritmo *Latent Dirichlet Allocation* (LDA), sviluppato da Blei et al. (2003) con lo scopo di estrarre i temi collaterali presenti nel documento sottoposto all'algoritmo. Dopo diverse prove condotte con numeri di topic da quattro a dodici, che si sono rivelati inadatti in base a metodi di valutazione quali la coerenza intratopica e librerie di visualizzazione dei topic, il numero di topic da estrarre è stato impostato a quattro per ogni sotto corpus. I risultati ottenuti dai due sotto corpora riportano molte somiglianze: infatti, i gruppi di topic estratti da entrambi i sottogruppi contengono termini che si riferiscono alle stesse sfere semantiche, come termini legati al crimine stesso, alle relazioni fra le persone coinvolte nel crimine, ai luoghi del crimine e alle forze dell'ordine. Un termine presente nei topic tedeschi ma non in quelli italiani è *colpevole*, mentre tutti gli altri termini sono quasi gli stessi, o comunque appartengono allo stesso ambito semantico. Altra peculiarità osservata è che un intero topic è dedicato alla figura della donna/vittima, nel sotto corpus italiano, mentre questo è assente nel sottocorpus tedesco, dove comunque la figura della donna risulta essere molto rilevante, dato che il termine donna – *Frau* – è presente in tre topic su quattro e, nel topic in cui questo termine non è presente, sono comunque rilevanti i termini legati alla figura della donna – *Mutter, Mädchen* (“madre”, “ragazza”). Inoltre, nei topic estratti dal sotto corpus tedesco, i termini legati alle relazioni interpersonali implicate nel crimine e ai ruoli propri del crimine hanno più rilevanza rispetto ai topic estratti dal sotto corpus italiano. Infatti, un topic estratto dal sotto corpus tedesco è dedicato ai termini legati alla famiglia e ai ruoli familiari, mentre un altro topic è dedicato ai ruoli implicati nel crimine.

La metodologia utilizzata per estrarre le metafore segue la metodologia usata in Busso, Combei e Tordini (2019) per l’analisi delle metafore, ed è basata sullo studio delle collocazioni di alcune parole individuate come caratterizzanti di ogni sotto corpus e che fungono da possibili termini target. Dallo studio delle collocazioni è stato possibile individuare alcune metafore e, quindi, i termini sorgente, su cui è stata basata la seconda parte delle analisi. Infatti, dopo aver analizzato le collocazioni dei termini target, anche le collocazioni dei termini sorgente sono state prese in analisi, per poter stimare il loro uso metaforico e considerare se usate anche con altri termini target. Per risalire alle metafore concettuali, le metafore linguistiche sono state raggruppate in base al dominio sorgente e poi in ulteriori sottogruppi in base al dominio target.

Allo stesso modo, le metafore concettuali che emergono dalla seconda analisi contengono metafore presenti in entrambi i sotto corpora, e metafore presenti solamente in un sotto corpus. Le metafore concettuali condivise fra il corpus italiano e il corpus tedesco sono quelle che riguardano il crimine e il colpevole del reato, avendo come dominio sorgente ANIMALE o INUMANO. Le metafore che non sono condivise sono quelle metafore che non riguardano direttamente il crimine, ma aspetti collaterali e presenti nella narrazione come un litigio, una malattia psichiatrica o i social media. Metafore riguardanti litigi e malattie psichiatriche sono presenti solamente nel sotto corpus italiano, mentre il tedesco tende a usare termini propri dell’ambito di riferimento quando si presenta un litigio o una malattia mentale all’interno della narrazione. Al contrario, un lessico meno specifico viene utilizzato quando si racconta dell’interazione sui social media. Infatti, metafore che possono essere riportate a una metafora concettuale come SOCIAL MEDIA ARE A PLACE FOR MOURNING sono presenti nel sotto corpus tedesco, ma assenti nel sotto corpus italiano, che usa invece termini propri dell’ambito dei social media.

Dall’analisi dei risultati ottenuti è quindi possibile dedurre che non ci sono differenze sostanziali nel modo di percepire il fenomeno a livello mediatico, almeno per quanto riguarda la narrazione in articoli di cronaca, e che le uniche differenze riscontrate sono differenze che rimandano a peculiarità linguistiche differenti, come l’uso di termini generalmente più specifici nella lingua tedesca, o la tendenza riscontrata di usare termini propri dell'ambito dei social media per quanto riguarda la lingua italiana.

I risultati ottenuti riportano quasi una ambivalenza interna alla narrazione del fenomeno del femminicidio. Da un lato ci sono i risultati ottenuti dal primo esperimento che sembrano essere la rappresentazione di una visione neutrale e oggettiva del fenomeno, di una narrazione che si limita a riportare quanto accade senza aggiungere nulla del background culturale di chi scrive. D'altro canto, l'analisi delle metafore concettuali sembra suggerire che le similitudini riscontrate rimandano a un modo analogo di percepire il fenomeno del femminicidio e della violenza di genere all'interno delle due culture; l'oggettività linguistica riscontrata nell'analisi dei topic lascia invece spazio a espressioni linguistiche e metafore concettuali che sembrano non riconoscere la sistematicità del fenomeno, cercando quasi di annullare la responsabilità del colpevole e azzerando la figura della donna protagonista dell'evento criminale.

Introduction

The phenomenon of femi(ni)cide[1] is considered a worldwide spread phenomenon which is meant to increase further. The phenomenon of femi(ni)cide as a type of homicide related to the gender of the victim and to unbalanced relations of power in Western society is recognised only in 2012 on a report of the UN Special Rapporteur Ms Rashida Manjoo (Weil, 2018), where the phenomenon of femi(ni)cide is described as the worst form of violence against women, experienced in a sequence of violent acts. A year later, the Vienna Declaration describes this phenomenon as the killing of a woman because of her gender. The gender-based feature of this type of homicide is already theorised by Johnson (1995), where a differentiation between two types of couple violence occurs, namely *common couple violence* and *patriarchal terrorism*. The term patriarchal terrorism seems to perfectly illustrate the idea of a systematic violence based on the gender stereotypes at the heart of Western society. The violence preceding the death of the victim is a systematic and intentional violence, not a sporadic response to a certain event, and its roots are in the patriarchal culture itself. Femi(ni)cide is, in fact, usually intended as something which does not happen in isolation but it is simply a type of manifestation of a culture which accepts ‘behaviors that are degrading and detrimental to women, including many forms of violence’ (Rodriguez, 2010). Conditions for femi(ni)cide to happens already lay in the patriarchal and sexist society where people live in. However, femi(ni)cide is not only restricted to Western society. Indeed, United Nations Office on Drugs and Crime (UNODC, 2018) recorded a total of 87,000 murders of women in 2017 all over the world, the 58% of whom was killed by intimate partner or a family member. On a global level, the number of the killings of women by the hand of a lover or a relative registered are extremely higher in respect to the number of men intentionally killed by intimate partners or family members (82% women, 18% men). This disparity is clear also in a report about crimes committed by intimate partner in 2019 complied the German Police Office (Bundeskriminalamt, 2020). This report contains information about violent crimes occurred by the hand of the partner in 2019 in Germany. In the year taken into account, the victims are 141,792, 114,903 of whom are women. This document reports that 301 women died by the hand of their partner, against 93 men, while women victims of severely physically injured are 11,991, against 5,169 men, women victim of stalking

[1] This term is used to refer to gender-related killings of women. It is a term composed both of *femicide* and *feminicide,* words used to describe the killing of a woman because she is a woman. *Femicide* has been firstly introduced in English speaking countries, while feminicide is a translation of the Spanish translation *feminicidio* (Pinelo, 2018). The use of this hybrid word is a reference to both ideologies.

are 28,906 (against 3,571 men), and the number of women deprived of their freedom are 1,514 (against 183 men). A similar document was drafted for the year 2015 from the same office omitting any allusion to gender-related violence and femi(ni)cide to describe these killings and crimes which exponentially occur more against women than against men, a reference which is absent in the 2019 report too. This feature is underlined in Pinelo (2017), where it is reported the number of women killed in that year by partners or ex partners (311), which corresponds to the 82% of the victims of couple violence killings reported in the document concerning 2015. Moreover, Commission on Crime Prevention and Criminal Justice clearly states that gender violence exists in Germany and is experienced by the 40% of the women at least once in a lifetime, but that femi(ni)cide 'is not a phenomenon which can be found in Germany' (2014, Annex, p.1). As already noted in Pinelo's article, these numbers state that the phenomenon of gender violence in general and, more specifically, of femi(ni)cide cannot be denied and the absence of the term femi(ni)cide in the German documents and the late introduction of the term in official contexts in 2012 (ten years ago with the Istanbul Convention) suggest that a lot of work is still to be done to comprehend this phenomenon and to prevent it. In the last years, new associations are arising to study the phenomenon, raising awareness about femi(ni)cide and help to prevent it. Some of these associations are creating databases reporting the cases of femi(ni)cide – as those consulted to create the corpus for this thesis – to strength awareness about femi(ni)cide. As reported in Weil and Naudi (2018), collecting and analysing data about the phenomenon are considered to be important weapons to prevent gender-based violence and therefore gender-related killings by the CEDAW (Convention on the Elimination of All Forms of Discrimination against Women) general recommendations and by the Council of Europe Convention on Combating and Eliminating Violence against Women and Domestic Violence. Another important tool to raise awareness about the theme of femi(ni)cide is the media (Weil and Naudi, 2018). In this perspective, the current study aims to analyse newspaper articles about femi(ni)cide in order to understand the perception media have of the phenomenon and if their point of view may indicate different ways of understanding the phenomenon on behalf of the two Countries in analysis, Italy and Germany. Moreover, a work on femi(ni)cide awareness seem to be necessary also because, as suggested by Weil and Naudi (2019), and by the German documents illustrated above (Bundeskriminalamt, 2016, 2020), femi(ni)cide seems to be quite difficult to recognise, while it is easier to identify gender violence.

This thesis is, therefore, a cross-linguistic study whose aim is to analyse what characterises the narration of the phenomenon of femi(ni)cide. More specifically, it analyses newspaper articles published in Italy and Germany during the same year – 2019 – and compares the results to detect whether there are any similarities or differences and if these differences may be caused by different ways of treating the phenomenon or to dissimilarities due to the two different languages and cultures. This goal is achieved through the use of text mining techniques, such as topic modelling algorithms and collocations analysis. In fact, this analysis proposes to mainly study two aspects of the narration, namely the recurrent themes characterising the texts and the metaphors used in the newspaper articles. The choice of the two techniques to apply is determined by the interest in grasping what collateral themes are connected to the main theme of femi(ni)cide and what are the mental associations laying behind the narration of this type of violence and the people involved in it.

The choice of this theme, femi(ni)cide, is due by its universality, as the phenomenon is usually described as something proper of almost any society. The universality of the phenomenon allows for the possibility to collect texts in different languages about it, and hence to create a bilingual corpus and to analyse the way the phenomenon is treated by mass media in different cultures.

To this purpose we collected a bilingual corpus formed by Italian and German newspapers published in 2019 and concerning femi(ni)cide cases happened in 2019. It is composed of ten newspapers in total, five for the Italian part of the corpus – *La Repubblica*, *La Stampa*, *Il Giornale*, *Il Fatto Quotidiano*, and *Il Resto del Carlino*[2] –, and five for the German subcorpus – *Die Zeit, Süddeutsche Zeitung, Die Welt, Frankfurter Allgemeine Zeitung,* and *Bild*[3]. The data used to collect the documents for the corpus are taken from *Casa delle donne per non subire violenza, ONLUS*, Bologna and from *Osservatorio sul femminicidio – ProsMedia*, for what concern**s** Italian femi(ni)cide cases, while data concerning German cases are taken from a database provided *Feminizidmap.*

[2] https://www.repubblica.it/
https://www.lastampa.it/
https://www.ilgiornale.it/
https://www.ilfattoquotidiano.it/
https://www.ilrestodelcarlino.it/

[3]https://www.zeit.de/
https://www.sueddeutsche.de/
https://www.welt.de/
https://www.faz.net/
https://www.bild.de/

Topics extraction is realised through an unsupervised algorithm, *Latent Dirichlet Allocation* algorithm, which returns a selected number of groups of words – topics – which usually appear in the same documents. The algorithm is applied twice, once on the Italian subcorpus and once on the German subcorpus, returning two sets of four topics. The number of topics to be extracted is empirically decided, through a close analysis of the set of topics and the use of evaluation measures, e.g., pyLDAvis and Umass Measure. The algorithm application is a tool employed to understand which collateral discourses are implied in the narration of femi(ni)cide. Results show that discourses involved in the narration of femi(ni)cide in Germany and Italy are still related to the main theme of the articles, the crime.

The main theory beyond metaphors analysis is Conceptual Metaphor Theory, developed by Lakoff and Johnson (1980/2003). This theory states that metaphors are not simply a linguistic phenomenon, but already shapes our way of reasoning in our mind. Therefore, according to this theory, every linguistic metaphor should be attributed to a broader conceptual metaphor which influences our way of perceiving a certain object or event in terms of something else and, thus, leads us to talk about it using a metaphorical language. Identifying the conceptual metaphors involved in the narration of femi(ni)cide may help understanding how the phenomenon is perceived and the mental connection behind it. In order to understand the understanding of the phenomenon, linguistics metaphors concerning people involved in the crime event, both general roles as *victim* or relationships names as *wife* or *husband*, are extracted. Due to the similarities between the two Countries analysed, it is not expected to find strong differences in the narration of the phenomenon by the media. It is indeed expected to find affinity in the narration, mainly for what concerns the patriarchal background of Western societies. However, these patriarchal values should be filtered by the style of news journalism, which does not give much space to personal opinions but usually required a language the more impartial as possible. Nonetheless, opics concerning the intrinsic different worth of men and women in patriarchal societies and also conceptual metaphors related to this idea are expected to be found.

The following work is divided into five chapters. The first two chapters work as theoretical introduction to the text mining techniques used and the theory beyond them. The third chapter is an introduction to web-based corpora and to the creation of the Italian-German corpus used in this study. The fourth and the fifth chapters report the conducted experiments and the results obtained.

The first chapter reports the state of the art of conceptual metaphor theory and metaphors extraction from corpora. Initially, theories about metaphors were more speculative, but already from the new millennium, metaphor studies became to be more empirical, involving the aid of corpora. Among the studies reported in this chapter, also studies based on bilingual corpora and on corpora with newspaper articles concerning femi(ni)cide are reported.

The second chapter focuses on topic modelling algorithm in general and on the development of the algorithm used in this study, latent Dirichlet allocation. Among the studies reported in this chapter, also studies concerning topics extraction on a corpus of femi(ni)cide newspapers articles and a close reading analysis of articles of the same types are reported.

The third chapter concerns the creation of the corpus for the study. The first half of the chapter introduces the idea of corpus and its use in linguistic studies, paying more attention on web-based corpora; corpora created with the aid of online language material. The second part of the chapter focuses on the description of the methods used to create the corpus for the studies and its features. A statistical description and a lexical analysis of the corpus are provided at the end of the chapter.

Chapter four reports the experiment conducted to extract topics with the LDA algorithm. It describes the preliminary processes of cleaning the documents and deciding the parameter to create the LDA model. In the second half of the chapter, results are reported, a paragraph for each subcorpus, and a discussion of the results follows.

The fifth chapter reports the experiment conducted to extract metaphors. The first paragraphs focus on the description of the tool and the methods used to extract metaphors, while the second part presents the obtained results, a paragraph for each subcorpus, and the final discussion.

1. (Conceptual) Metaphor

Every person is familiar with the concept of metaphor, probably the first meeting with this word happens during the elementary school, when the teacher introduces for the first time the figures of speech. The first explanation everyone hears mirrors the classic idea of metaphor: an ornamental device that concerns language and nothing more. Classical theories on metaphors (Mácha, 2016) are based on the assumption that everyday language is literal and only some new expressions are metaphorical. This means that metaphors are conceived as new and poetic expressions, that are only used outside conventional language sphere. Since Aristotle, metaphor was intended as a figure which allowed to transfer a meaning from one object/word to another. As cited in Levin (1982), Aristotle provided the following example to explain the metaphor as a process of analogy:

> ‘For instance, a cup is to Dionysus what a shield is to Ares; so he will call the cup 'Dionysus' s shield' and the shield 'Ares' cup.'’ (Poetics, Aristotle in Levin, 1992, p.24).

This example starts from a parallelism between the two gods and the objects typically associated with them. This is an analogy, as the objects have the same worth for the two owners. Therefore, calling Dionysus’ cup his shield, implies a non-literal use of the word *shield,* which is recognised as a poetic metaphor. Classical theories on metaphors do not try to explain this phenomenon as something related to thought, but only to language, and a particular type of language, namely the language of poetry and fiction.

In the last decades, the idea of metaphor shifted from something exclusively linguistic to something related to thought. Lakoff and Johnson (1980/2003) illustrates *Conceptual Metaphor Theory*. According to this theory, human thought is metaphorical. This theory assumes that verbal metaphors are not merely linguistic, but they already exist in human brain. Other scholars follow this way of conceiving metaphors, underlying metaphors’ role in the way people perceive the world (Gibbs 2011, Cserép 2014).

Examples provided by Lakoff and Johnson (1980/2003) aim to explain how this process works. Metaphors like ARGUMENT IS WAR and TIME IS MONEY may be unconsciously used by the speaker in everyday situation. These two concepts, that are at the basis of many linguistics expressions, are considered conceptual metaphors. They are formed by two domains, a target domain – the domain metaphorically described – and a source domain – the domain used to describe the target domain, usually a concrete domain. In these two examples, ARGUMENT and TIME are the target domain, in fact they are the most abstract topics in the metaphor, while WAR and MONEY are the source domain. This means that word referring to the semantic field of war and money are used to talk about more abstract concept as argument and time. Therefore, the aforementioned conceptual metaphors may be linguistically translated into expressions as 'I demolished his argument' and 'you're wasting my time' (Lakoff and Johnson, 1980/2003).

Usually, metaphors are the easiest way people know to talk about something. This is because target and source domains are linked, and the activation of the source domain may bring to the activation of the target domain. For example, when the node of the source domain WAR is activated, the node of the target domain ARGUMENT is rapidly activated. Thinking metaphorically is therefore easy and almost automatic as thinking simultaneously of items from the same domain (this theme will be better discussed in paragraph 1.2). Furthermore, there are cases where the only possible way to talk about something is in metaphorical terms. As reported by Deignan (2005), it is quite impossible to talk about life in different terms than the metaphor LIFE IS A JOURNEY. In her book 'Metaphor and Corpus Linguistics', the linguist underlies that 'for many metaphorical expressions there are no literal paraphrases, and certainly none that are "exactly and literally what we mean".' (Deignan, 2005, p. 17), a point of view shared in the community of Conceptual Metaphor researchers, especially for what concerns abstract topics (Sweetser, 1990/2012, Kövecses, 1991). An example may be the CONDUIT METAPHOR, identified by Reddy (1979/1993 in Deignan, 2005) and further analysed by Lakoff and Johnson (1980/2003). This is a complex conceptual metaphor used by speakers when talking about language and may be illustrated as: ideas (objects) are put into words (containers) and sent (throughout a conduit) to the hearer who puts them out of the container. The CONDUIT METAPHOR may be individuated into the linguistic metaphor 'it's difficult to put my ideas into words', in which the conception that words are containers and ideas need to be inserted in it in order to be understood by the hearer is clearly reported. Moreover, this conceptual metaphor is no more perceived as such and is almost the only available way to talk about language.

As conceptual metaphors are grounded in the way people experience the world, it is unavoidable that the body plays an important role. Conceiving abstract concept as more concrete things leads also to think about more abstract concepts in term of the bodily experience people have in everyday life. Since human body is universal, conceptual metaphors using the body as source domain are potentially universal, 'this explains why many conceptual metaphors, such as KNOWING IS SEEING, can be found in a large number of genetically unrelated languages' (Kövecses, 2017). Therefore, linguistic metaphors as 'The argument is clear' (Lakoff and Johnsons, 1980/2003) may be found in different languages because the conceptual metaphor KNOWING IS SEEING is not linked to a specific culture**,** but to universal bodily experience (e.g., it is possible to hear utterances like *che idea brillante* in Italian too).

However, conceptual metaphors based on bodily experiences should be considered as potentially universal, as they may underwent changes according to the cultures and to more specific contexts. Some metaphors may be found in a specific culture, while others may be context induced and therefore not present in the brain of all the speakers of a certain culture. Hence, according to Kövecses (2017), there are three main sources for conceptual metaphors, namely the body, cultural specificity, and more general context.

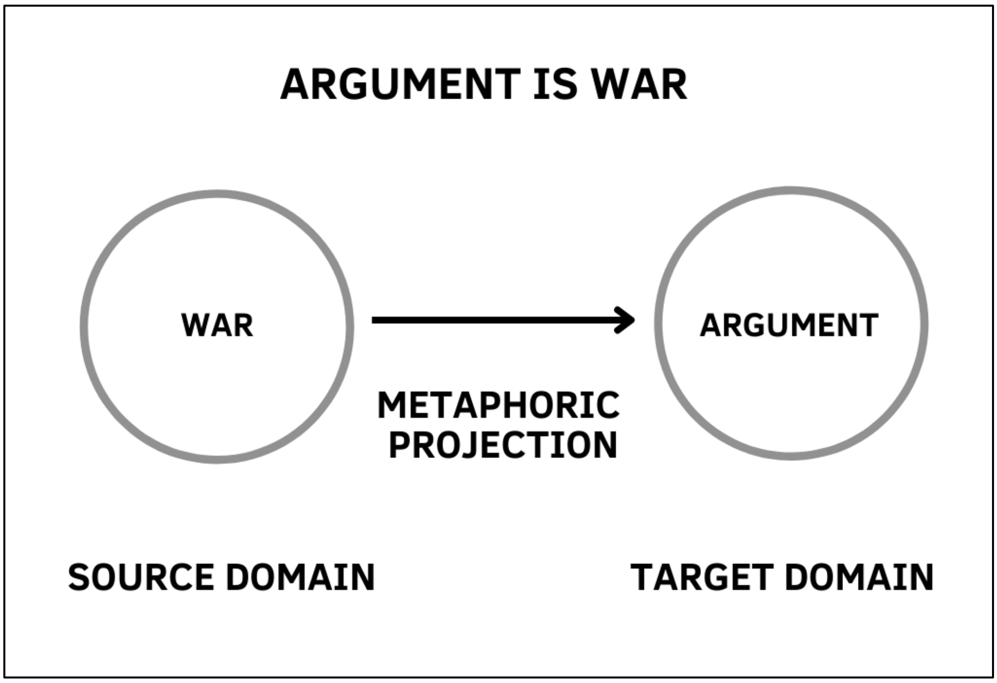


Figure 1.1: the metaphorical projection goes from the source domain – WAR – to the target domain – ARGUMENT.

1.1 Main types of metaphors

In the first theory developed by Lakoff and Johnson (1980/2003), there are three types of conceptual metaphors, namely structural, orientational and ontological.

Examples provided so far – KNOWING IS SEEING, ARGUMENT IS WAR, TIME IS MONEY – are commonly categorised as *structural metaphors*, metaphors in which one concept is structured in term of another. For instance, the linguistic metaphor 'I demolished his argument' (Lakoff and Johnsons, 1980/2003) structures ARGUMENT in term of WAR, using the verb *to demolish*, whose semantic field is that of war**.** *Orientational metaphors*, on the other way around, organize a system of concepts in term of another**.** This type of conceptual metaphors is usually characterised by bodily experience and cultural peculiarities and makes one concept understandable in terms of spatial orientation. Orientational metaphors are HAPPY IS UP and SAD IS DOWN. According to Lakoff and Johnson (1980), these two conceptual metaphors may be justified by the posture of human body when in a status of happiness or sadness. When people are happy, they commonly assume an erect posture, while when they are feeling sad, they assume a droop posture**.** Therefore, it is common to hear 'I am feeling up', when someone is feeling happy, and the linguistic metaphor 'I am feeling down' when someone is sad. In general, when a culture thinks of happiness as something 'up', all the emotions, feelings, or, more broadly, abstract concept perceived as positive by that culture will be thought at as something 'up' – as the conceptual metaphors HEALTH IS UP and HAVING CONTROL IS UP illustrated in Lakoff and Johnson (1980/2003). On the contrary, if happiness is seen in this way, all the negative feelings would be conceived as something 'down', in opposite to all the positive concepts – SICKNESS IS DOWN and BEING SUBJECT TO CONTROL IS DOWN. The fact that these conceptual metaphors are influenced both by the bodily experience and by the culture, brings to the assumption that the shifting of values caused by a new situation mirrors in the shifting of orientational metaphors. Example provided by Lakoff and Johnson (1980/2003) to illustrate the influence of culture on conceptual metaphors is the purchase of a new car. In times when economy flourishes, peoples are guided by the metaphor BIGGER IS BETTER, that leads them to by a new car which is as bigger as possible. In a period of financial crisis, on the contrary, SAVING MONEY IS BETTER takes the place of the previous conceptual metaphor. The orientation of the two metaphors is the same, namely 'better', that may be spatially translated as 'up' – GOOD IS UP – but this 'good' changes as times do. Conceptual metaphor system stays coherent with the cultural period people live and changes with it.

Ontological metaphors are another type of metaphors used to think about abstract concepts in terms of more concrete objects. As structural and orientational metaphors, also ontological metaphors derive by people's bodily experience of the world and are evidence that it is easier to think, and talk, about concrete things rather than abstract. Events, ideas, emotions are seen as entities or substances in order to fulfil purposes such as referring to them, quantifying them, or identifying some of their peculiarities. As for orientational metaphors, also utterances containing ontological metaphors may be difficult to be recognised as metaphorical expressions. A widespread metaphor is the container metaphor, namely conceiving something abstract as a container to be filled. VISUAL FIELDS ARE CONTAINERS may be an example, and it is not perceived as such, so when the speaker utters 'he's out of sight', it is probably that the pronunciation of this utterance happens without an explicit knowledge of the metaphorical entity in it. According to Lakoff and Johnson (1980/2003), personification is an extension of ontological metaphor. The sentence 'inflation is eating up our profits' does not only mirror that people thinks about an abstract concept in term of an entity, but also that this entity is humanised.

Lakoff and Johnson (2003) revise this categorisation affirming that a metaphor may have more than one of these three features. Each metaphor is at the same time structural – as every conceptual metaphor maps structure to structure –, ontological – as every conceptual metaphor creates target domain entities –, and many of them are orientational – as they map orientational image schemas. The conceptual metaphor MIND IS A CONTAINER, for instance, may be analysed both as structural, as it maps the structure of the target domain – MIND – and the structure of the source domain – CONTAINER –, and as ontological, as the abstract idea of mind is concretised through the image of a container in order to talk about it.

Since all conceptual metaphors are structural, all conceptual metaphors imply a mapping between the source and the target domain. This mapping occurs at the brain level and allows speakers to understand 'one domain of experience […] in terms of another' (Kövecses 2017, p. 13), giving rise to linguistic metaphors. How this process works at the brain level will be illustrated in the next section.

1.2 Conceptual metaphors in our brain

This paragraph will focus on the process which takes place in human brain to form conceptual – and therefore linguistic – metaphors.

This metaphorical mapping is created by the simultaneous activation of nodes belonging to different brain regions and the creation of a new path, shorter to the already existing connections, that goes form the source domain node to the target domain node.

As illustrated in Lakoff (2014), the first step which brings to the creation of the metaphor is the activation of an already present neural path connecting the neurons of the two activated nodes. As long as the activation is kept, neural links between the two domains are created and increase in strength. A contemporary regular activation of the two nodes causes the creation of a shorter path between them, namely the metaphor. This is an asymmetrical activation pattern as it goes from the source domain to the target domain. Taking as example the conceptual metaphor THE MIND IS A CONTAINER, it is possible to understand this phenomenon. In this case, the two activated nodes are the node concerning *mind* and the node concerning *container*. It is easy to talk about the mind as a container because the regular simultaneous activation of the two nodes causes the creation of a path shorter than the already existing paths, which gives the possibility to rapidly connect the two domains. This activation is, therefore, an asymmetrical relation, as the resultant circuit goes from the container node to the mind node. It is in fact possible to think and talk about the mind – the target domain – in terms of a container – the source domain – but not the opposite. A similar assumption was already made by Lackoff (1990 in Deignan 2005), in the theory known as the *Invariance Hypothesis*, investigating the linguistic expressions of some conceptual metaphors. This theory states that the image-schematic structure preserved in the metaphor is that of the source domain (Gibbs, 2011). Examples provided by Gibbs (2011) to explain this hypothesis include the image schema of *source-path-goal*. The experience of moving from one place to another creates a recurring pattern which is later projected on more abstract domains. The conceptual metaphor arising from this schema is PURPOSES ARE DESTINATION, linguistically conveyed as 'I got sidetracked on my way to getting a PhD' (Gibbs 2011, p.536). In the linguistic metaphor, schemas belonging to the source domain – DESTINATION – are applied to the target domain – PURPOSES – so that the PhD becomes a destination to be reached.

A feature of conceptual metaphors is the ability to activate more than one source domain from one conceptual target domain. This phenomenon is presented by Narayanan's *Spike-timing dependent plasticity* theory (in Lakoff 2014). The phenomenon may be justified by the existence of different conceptual metaphors sharing the same target domain, e.g., LOVE IS A JOURNEY, LOVE IS A PATIENT and LOVE IS MADNESS (Lakoff and Johnson 1980/2003, p.85) share the same target domain – LOVE – but are built with different source domains – JOURNEY, PATIENT, MADNESS.

Moreover, at the brain level, it is possible to make a distinction between metaphors, namely between primary and complex metaphors. According to Lakoff (2014), primary metaphors are circuits that maps a primitive neural schema into another primitive neural schema. In other words, primary metaphors directly link two nodes – as it happens for the conceptual metaphor THE MIND IS A CONTAINER already described in this paragraph. On the other hand, complex metaphors are metaphors composed of two or more primary metaphors. THEORIES ARE BUILDING may be considered a complex metaphor as it is composed by the primary metaphors PERSISTING IS REMAINING ERECT and STRUCTURE IS PHYSICAL STRUCTURE (Gibbs, 2011). A further distinction between these two kinds of conceptual metaphors is that primary metaphors are directly rooted in bodily experience – as PERSISTING IS REMAINING ERECT – while complex conceptual metaphors are more likely to have a cultural and social motivation, other the physical motivation provided by the primary metaphors they are built on.

1.3 Blending Theory

Although Conceptual Metaphor Theory received positive feedback in the academic environment, with many scholars who decided to further investigate this theory and its application (Gibbs, 2011, Cserép, 2014**),** it still has a weak point. The *Invariance Hypothesis* has been long debated as it does not recognise any role to the target domains in the final structure of the final metaphor. Therefore, another theory that studies figurative language developed by Fauconnier and Turner in 2002 arose, namely the Blending Theory**.** The focus of this theory is how conceptual structures are combined in language use. The blended space is a mental space in which elements of other mental spaces combine together. Differently to what stated by Conceptual Metaphor Theory, blending needs two or more input spaces to occur and it is not an asymmetrical mapping but 'both inputs project into another structure, called the blend' (Dancygier, 2017). An example of blending may be 'silver tsunami', expression used by the media to talk about the huge number of retiring seniors and the possible impact this phenomenon could have on the economy. According to Dancygier (2017), this blend has three inputs: seniors, tsunami, and economy. The projection arising from the first two inputs – seniors and tsunami – is a partial projection, as only some particular features are projected in order to form this blend, namely old people have grey hair and are in retirement and a tsunami is a catastrophic event. The results of this catastrophic event will be seen in a different domain, therefore input 3 is required to complete the blend. The emergent structure of this new configuration of concepts is therefore a blend where the action of old people (retirement) may cause a negative event (the catastrophic wave of the tsunami) in the economic field. At the end of this process a backward projection takes place in order to add this new conceptual structure to the inputs involved. However, this new meaning is only true in the blending. Therefore, when thinking about aging or old people, none would see it as negative for the society, but when in this blend.

In analysing metaphors, Conceptual Metaphor Theory e Blanding Theory should not be seen in opposition to each other, but rather as two methods to study figurative language. As suggested by Deignan (2006), linguistic metaphors may sometimes be not so static as it is required by Conceptual Metaphor Theory, according to which the structure of the source domain is imposed to the target domain. In her study on the literal and metaphorical use of lexical items belonging to the source domains of FIRE, PLANTS, MOVEMENT and ANIMALS, she noticed that many metaphors do not maintain the structure of the source domain but seem to be influenced from both inputs. Examples provided in Deignan (2006) show how a lexical element may change part of speech from its literal to its metaphorical use. A noun, *blossom,* may be used as a verb when in a metaphorical expression: ‘Venture capitalists provide the vital infusion of funds to help budding capitalists blossom’ (Deignan 2006, p. 113). According to the Invariance Hypothesis, the source domain controls the final metaphorical structure. In this example, however, the structure differs from that of the source domain, as in the source domain of plants, when literally used, *blossom* is more likely to be a noun than a verb. Therefore, both the influence of the source domain and of the target domain is found.

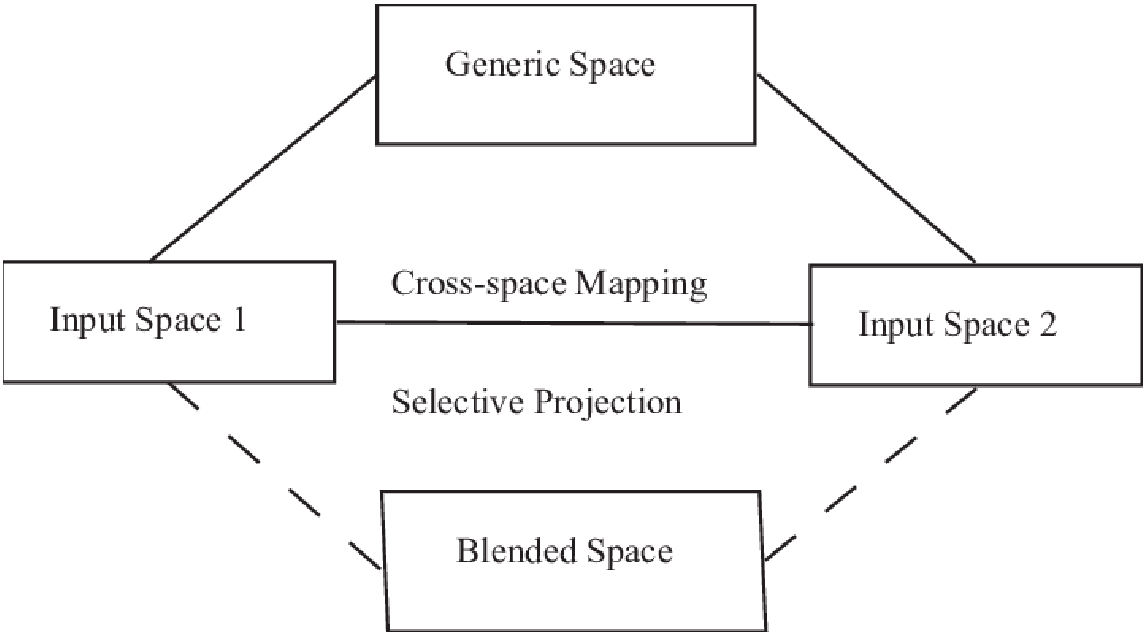


Figure 1.2: a graphical illustration of blending. The blended space is an outcome derived from the partial projection of two input spaces in a new space, the blend. (Somov and Voinov, 2017, p. 618).

1.4 Conceptual metaphors in language

A useful way to study conceptual metaphors is searching for their lexical clues in corpora. As illustrated in Deignan (2006**),** this type of studies may have two main aims, namely analysing how metaphors are used to define something in particular or to convey an ideology – these studies may be part of the Critical Discourse Analysis tradition –, or to study how metaphors are used to construct shared understandings in interaction. As the aim of this thesis is more comparable to Critical Discourse Analysis studies, from this point onward, research aiming to understand the role of metaphors in interaction among people will not be illustrated, but whether corpus analysis may justify the existence of conceptual metaphor and how metaphors are used to describe a particular theme or to convey a particular ideology.

Santa Ana (1999) investigates metaphors used to talk about race in connection to immigration in US in *Los Angeles Time* articles. From a corpus made of 107 articles, ten per cent of 1900 metaphors are about immigrants, the majority of which depicts them in a negative way, as IMMIGRANTS ARE DEBASED PEOPLE, IMMIGRANTS ARE WEEDS and IMMIGRANTS ARE COMMODITIES. The main conceptual metaphor that arises from his search is IMMIGRANTS ARE ANIMALS – e.g. 'The truth is, employers hungering for really cheap labor hunt out the foreign workers' (Santa Ana, 1999, p. 201). This conceptual metaphor is further investigated in Deignan (2006), where the scholar searches for the occurrences of the verb *hunt* and its inflections to analyse the connotations taken by the metaphorical use of this verb. When metaphorically used, the occurrences may be divided into four groups, namely *good hunts bad*, *good hunts good*, *bad hunts bad,* and *bad hunts good.* From this analysis, it is not possible to detect whether immigrants are depicted in a positive or negative way. In fact, Deignan (2006) points out that it is difficult to define a conceptual metaphor as negative or positive as it has many different possible mappings, and it is influenced by the context of use. However, it is possible to agree with Santa Ana (1999) results that immigrants are not presented as humans but as entities to be used.

As the metaphorical mapping is from one domain to another, it should be expected that the same word is used in the same literary metaphor, or with the same aim, regardless of its wordform. Instead, it has been noticed that inflection affects the metaphor, and the same word may be used as a reference to one conceptual metaphor when in the singular form, and a completely different conceptual metaphor when used in the plural form. Deignan (2005) extracts linguistic metaphors with the noun *rock*, both in its singular and plural form, from the Bank of English (HarperCollins, 2005). Results show that the singular form of the noun is used in a positive way, as to show support to someone or linked to the idea of building something. On the other hand, the plural form of *rock* is used in a negative way, probably with a relation to the low physical position of rocks on earth and therefore as a result of the conceptual metaphor DOWN IS BAD. The author affirms that these linguistic metaphors are influenced by the literal meanings of the lemma in analysis. The metaphorical use of *rock* alludes to the biblical episodes in which is said that the house of God is built on a rock. Hence, the metaphorical use of the singular form still keeps this idea of strength and stability, e.g., 'his rock-solid defence of traditional values' (Deignan, 2005, p. 218) On the other hand, the negative values of the plural form *rocks* seems to come from the literal use of *rocks* used to indicate those rocks under the see, which are dangerous for sailing.

The difference emerged in the use of different wordform of the same lemma seems to be in contrast with Conceptual Metaphor Theory, as the mapping is not between two domains, but it is between one wordform and a target domain. These findings better fit with Blending Theory, probably also because Lakoff and Johnson (1980/2003)'s theory is not a usage-based theory but a *speculative* theory. They do not ground their study on a corpus analysis, as the studies illustrated so far (Deignan, 2006, Deignan, 2005, Santa Ana, 1999), but they speculate on language according to their personal experience about metaphors they would say, they heard or read. This different kind of approach seems to be a limit when dealing with metaphor extracted from corpora, as it is highly likely that the two scholars did not take into account all the types of metaphors. This lack of usage-based methodology is filled from other scholars, as in the last years almost every study on metaphors is corpus based (Stefanowitsch, 2021).

An analysis similar to the study conducted by Deignan (2006) is conducted in Stefanowitsch (2021). The researcher based his work on the results obtained in Deignan (2006) about the linguistic metaphors with *flame* and *flames*. Again, the singular form presents more positive connotations than the plural form. On a 59-million-word section of the Bank of English (HarperCollins, 2005) she extracted 529 occurrences of *flame*, among which 95 had a metaphorical use. From these 95 metaphors with the singular wordform, only 5 had a negative prosody, while 90 were positive, such as those referring to lover – '[…] Ross spots his old flame Rachel in the congregation.' (Deignan, 2006, p. 116) – or belief and determination – 'Gascoigne does his best to keep that flame alight.' (Deignan, 2006, p. 116). On the other hand, of the 642 citations of the plural form *flames*, only 58 are metaphorical. 24 occurrences seem to have positive prosodies and the use of *flames* is linked to similar area as the positive use of *flame*, such as lover – 'I'd watch it when old flames try to be friends' (Deignan, 2006, p. 117) – and passion – 'will the flames of passion be fanned?' (Deignan, 2006, p. 117). The other 34 occurrences of *flames* have negative prosodies as they refer to being criticised – 'Weren't this band the band of the month last month? Now they're just shot down in flames?' (Deignan, 2006, p. 117) – being in a disastrous situation – '…his future crashing in flames.' (Deignan, 2006, p. 117) – or other negative feelings and experience – 'flame of intolerance' (Deignan, 2006, p.117). Stefanowitsch (2021) notices a fallacy of this experiment, namely the lack of separate analysis of the occurrences of the singular and plural form. This test has been made by the linguist extracting 20 occurrences of the literally used *flame* and 20 of the literally used *flames* from the British National Corpus (The British National Corpus, version 3 (BNC XML Edition). 2007). It was found out that negative connotations are more frequent with the plural form than with the singular wordform – the majority of negative connotation of the singular wordform arises when it is used as a collective noun. Therefore, the final results are consistent with Deignan (2006)'s suggestion of isomorphism between literal and metaphorical use of some words.

As already observed so far, studies on metaphor conducted with a speculative method risk to result unsatisfying, at least to a certain extend. In fact, as mentioned above, results obtained with corpus analysis seem to be inconsistent with Conceptual Metaphor Theory, as this theory was firstly hypothesised without the aid of any corpus observation. It is hence possible that some characteristics remained unnoticed. Studying metaphors on corpora gives the possibility to observed feature that, otherwise, would remain unknown, as the possibility to obtain different metaphors from different wordforms of the same lemma (Deignan 2006). A corpus-based approach may be also useful to understand the main attitude of a culture towards a theme and whether it is mirrored in the metaphors they use (Santa Ana 1999).

The next section should be helpful in order to understand how metaphors can be extracted form corpora. In the following paragraph two main methodologies will be illustrated, together with some examples.

1.5 Extracting metaphors from corpora

There are three main methods to extract metaphors from corpora, namely starting from the source domain, from the target domain, or from introductory expressions. As it is not expected to find expressions such as 'metaphorically speaking' or 'so to speak' in newspaper articles[4], only studies applying the first two methods will be reported in the following paragraphs.

1.5.1 Source domain studies

In this section, methodologies used to extract metaphors based on the source domain are illustrated. These studies take words from the source domain and analyse the context they are used in, in order to understand if they have a metaphorical meaning. In other words, to extract linguistic expressions of the conceptual metaphor TIME IS MONEY, researchers will look for words belonging to the semantic field of the source domain, MONEY. They will then analyse the context of each occurrence in order to understand whether the term is used metaphorically or not and if the linguistic metaphor belongs to the conceptual metaphor TIME IS MONEY.

A study provided in Stefanowitsch (2021) follows the methodology used by Deignan (1999) to analyse the isomorphism between literal and metaphorical language. The corpus used is the British National Corpus Baby (The BNC Baby, version 2. 2005) and, likewise Deignan's study, the focus is on the base form of adjectives (*cold, hot, cool, warm*) used as modifier of target domains nouns and nouns used metaphorically. The categories individuated by Deignan are presented with dictionary-like definitions and examples. Extracted occurrences are analysed similarly in Stefanowitsch (2021), according to five metaphor categories: ACTIVITY, AFFECTION, TEMPERAMENT, SYNESTHESIA and EVALUATION. The first category, *activity,* is found only for *cold* and *hot,* justified by the scholar as its use of the binary relation between the two opponents. Conceptual metaphors of this type are HIGH ACTIVITY IS HEAT and LOW ACTIVITY IS COLDNESS, linguistically expressed as 'cold war' and 'hot topic' (Stefanowitsch 2021, p.399). *Affection* is found for all the words except for *cool,* a phenomenon still unexplained, and is individuated in conceptual metaphors as AFFECTION IS HEAT and INDIFFERENCE IS COLDNESS. On the other hand, the *temperament* category is found in all the adjectives except for *warm,* explained as a possible speakers' confusion between AFFECTION IS TEMPERATURE and TEMPERAMENT IS TEMPERATURE. Conceptual metaphors linked to TEMPERAMENT are

[4] Research is conducted on a Sketch Engine composed of the online version of five Italian newspapers (*Libero, Il Resto del Carlino, La Repubblica, Il Fatto Quotidiano, Il Corriere della Sera*) and a blog (*huffingtonpost.it*) – a subcorpus of the available Italian Web 2016 (itTenTen16). Over 6,480,865 tokens, *metaforicamente* occurs only seven time. The other expression searched is *per così dire*, as it may also functions as an introduction to metaphors. The word *dire* occurs 2,965 times, only 11 occurences of the word are in the expression *per così dire*.

EMOTIONAL BEHAVIOUR IS HEAT and RATIONAL BEHAVIOUR IS COLDNESS, which are translated into the linguistic metaphors ‘cool head, cold facts, hot temper’ (Stefanowitsch 2021, p.400). The category *evaluation* is not equally distributed across the words in analysis, probably because it is not a unique conceptual metaphor, but may be a derivation from other conceptual metaphors (e.g., *cool person* may derive from TEMPERAMENT IS TEMPERATURE).

A further study illustrated in Stefanowitsch (2021) which aims to extract linguistic metaphors using source domain words as starting point – other than Deignan (1999, 2006) – is Stefanowitsch (2005). Its purpose is to find out evidence for the function of linguistic metaphors with literal paraphrase (*down of* NP/*beginning of* NP) in the British National Corpus. For each pair investigated, the metaphorical variant is supposed to be used with nouns referring to entities that are more complex. Results show that both expressions are typically used with words referring to time or events. However, a difference was found in these expressions. Words used with the literal *beginning of* refer to events with clear boundaries and a definite time span, such as *year, century, chapter, war* (Stefanowitsch 2021, p.407), while words used with the metaphorical *down of* refer to events without clear boundaries and duration, e.g., *civilisation, history, time, era* (Stefanowitsch 2021, p.407).

1.5.2 Target domain studies

Other than through words from the source domains, it is possible to extract metaphors also taking into account words from the target domains. According to Stefanowitsch (2021), this is the case especially when dealing with metaphorical patterns, namely ‘multi-word expression from a given source domain (SD) into which one or more specific lexical item from a given target domain (TD) have been inserted’ (Stefanowitsch 2021, p. 410). Taking into consideration the conceptual metaphor EMOTIONS ARE SUBSTANCES IN A CONTAINER, the linguistic metaphor *he is filled with anger*, for instance, contains words both from the source – the verb *to fill* is a linguistic expression of SUBSTANCES IN A CONTAINER – and from the target domain – the emotion *anger*. The technique proposed by Stefanowitsch (2021) is to select one or more word referring to a given target domain, retrieve all their occurrences from the corpus, and then identifying whether they are used in relation to their domain or to different domains. Final steps are to define the source domains and group the linguistic expressions into wider possible conceptual metaphors. This methodology is applied by Stefanowitsch (2021) in the analysis of linguistic metaphors linked to emotions in cross cultural studies. The study reported in Stefanowitsch (2021) looks for the word *happiness* in a British English corpus – LOB corpus (University of Oxford, Lancaster-Oslo-Bergen corpus of modern English (LOB), 1978) – and an Indian English corpus – Kolhapur corpus (Shivaji University, 1986). Comparing the main metaphors obtained from the two corpora, the PURSUIT and SEARCH metaphors – as [pursuit of NP_{EMOT}] or [seek NP_{EMOT}], e.g., *pursuit of happiness* and *seek happiness* – are equally present in both corpora, while the TRANSFER metaphor is more frequent in the Kolhapur corpus. Form this analysis he concludes that it is possible to grasp cultural differences through metaphors analysis.

Another case study using a similar methodology in Stefanowitsch (2021) aims to extract metaphors that express emotion gradeability, namely how intense an emotion is according to the metaphor used to express it. So, they look for syntagma containing the name of an emotion modified by an expression which has the role of increasing this emotion, e.g,. *explosive desire*. The term analysed are *anger, desire, disgust, fear, happiness, pride, sadness, shame.* The queries used to extract metaphors look for expressions such as [full of NP_{EMOT}] – when dealing with the conceptual metaphor EMOTIONS ARE SUBSTANCES IN A CONTAINER – or [NP_{EXP} burst/erupt/explode with NP_{EMOT}] (Stefanowitsch 2021, p. 418) – probably linked to the same conceptual metaphor, but here the substance fails to stay in the container.

A similar method is adopted by Baker (2006). Searching for the metaphorical uses of *allegation* in the British National Corpus, he firstly looked for collocations, and then for concordances of *allegation* and its plural form. The chosen grammatical patterns are: [noun] of allegation(s), [verb] by allegation(s) and allegations(s) as [adjective/noun]. After putting the relevant concordance lines of metaphor into groups, it has been noticed that it is possible to extract more metaphors with this methodology than with the collocation method. Some of the conceptual metaphors arisen from this analysis are HEAVY IS IMPORTANT (linguistically expressed as 'lend weight to allegations') and ARGUMENTS ARE BUILDING (linguistically represented as ‘allegations as “groundless”’) or ARGUMENT IS A CONTAINER (from the linguistic metaphor ‘waves of allegations’). For a more ambiguous expression, *trot out,* further research has been conducted to understand in which context it appears in the British National Corpus. Looking for the collocations of the verb in the corpus, it is possible to clarify the meaning of the expression – in the British National Corpus, *trot out* is metaphorically used to refer to the ideas or speech of a person, e.g., ‘Lyle, an affable philosopher, who will trot out an occasional quotation […]’ (Baker 2006, p. 172).

Another possible method is the keyword analysis suggested by Partington (1998) and reported in Stefanowitsch (2021), according to which the application of this kind of analysis on a specialised corpus based on the target domain will reveal the dominant source domains. To illustrate this method, Stefanowitsch (2021) use a subcorpus of the British National Corpus Baby (The BNC Baby, version 2. 2005) focused on the domain of ECONOMY. Then, he identifies the keywords of the subcorpus making a comparison with a subcorpus of the genre *newspapers*. The keywords extracted are all linked to the field of economics, except for *Midlesbrough* and *rise*. The latter word recall a down to top movement, a field that is found when taking into account the top 200 keywords. In fact, eleven of these keywords belong to the field of vertical movement, while twelve terms are linked to the semantic field of health – recovery – and increasing in size – expansion, growth. Expanding the analysis of keywords to 200 allows to notice that the vertical movement is an important source domain for economics. Then, a study of the concordances of the verb *rise* and its inflections has been made to check whether the verb was used metaphorically. In the 20 analysed occurrences, the verb indicates the metaphor MORE IS UP, rather than literally referring to the vertical movement, and the surrounding words make it possible to see where the metaphor is applied, namely to economic concepts as *prices*, *sales*, *profits* (Stefanowitsch 2021, p.425). Therefore, following this methodology, metaphors for a given target domain – ECONOMY, in this case – can be detected with a keyword analysis applied to a specialised corpus from that specific domain. Given the nature of the corpora required, this method is not always recommended, also because it is difficult to create specialised corpus for some domains, especially abstract domains as the domain of EMOTIONS (Stefanowitsch 2021).

1.5.3 Metaphors in gender violence narration

A study on metaphors in femi(ni)cide narration is Busso, Combei and Tordini (2019). Their aim is to analyse the lexicon of woman as a victim in a multimodal corpus (both written than oral) under different aspects. They analysed their corpus in many ways, one of this is extracting metaphors. The written corpus is composed of four Italian national newspaper articles published from September 2016 to June 2017. The metaphor detection part is divided into five main steps. The first step consists in searching the keywords in the corpus. This research included both keywords already known – as the researchers already used them to look for the articles used to build the corpus – and new keywords found through lexical analysis. This first simple and more raw research is followed by the observation of the cooccurrences and collocation of the keywords in analysis. A further step is that of perfecting the research by looking for phrases as *x is y*, and of noun phrases pre- or postmodified by an adjective. The third step of the research is based on the results of the preceding steps. They made a keyword in context research – they looked for a word, extracted its occurrences and analysing the contexts in which this word was found. The words they looked for resulted from steps one and two, form the source domain resulting from them, such as *preda* and *malato/a* belonging to the source domains of ANIMALS and DISEASE. In the fourth step they concentrated on complex syntactic structures, taking into account also lexical verbs as *sbocciare.* As final step, they organised all the metaphors in hierarchical structures to identify broader metaphors, and then checked them on MetaNet (Petruck, 2018).

Results show that women are represented as hunted animals, but men do not have the fault to hunt women, they remove responsibility from them – they are beasts, and beasts do not have control over their instincts. Another type of metaphor identified is bounded to new technologies, mainly internet, which is characterized as predator or as a court, as it has the power to kill – *massacra* – and the role of judging – *gogna del web* (Busso et al., 2019, p.273). After checking on MetaNet, they realised that they found out both conventional metaphors, such as ARGUMENT IS WAR, PEOPLE ARE ANIMALS, PEOPLE ARE PLANTS, and novel metaphors, as the metaphors related to the world of internet and social media – WEB IS A LOCATION, WEB IS A KANGAROO COURT – but also concerning violence and relationships – VIOLENCE IS A GAME, RELATIONSHIP AS A CLOSE LOCATION. From this study it is possible to conclude that linguistic metaphors used to talk about gender violence are expressions both of conventional conceptual metaphor, as ARGUMENT IS WAR – which was already cited in Lakoff and Johnsons (1980/2003) – and of conceptual metaphor never observed before in the literature – as VIOLENCE IS A GAME. The confirmation of old metaphors and the observation of new metaphors strengthens the opinion of the need of corpus-based studies to analyse the actual use of conceptual metaphors.

1.6 Cross-linguistic studies on metaphors

Another relevant line of research for the purpose of this thesis is that of cross-linguistic corpus studies on metaphors. These studies aim to study the same theme on two or more corpora of different languages and cultures. Studies of this type are reported in Deignan (2005) and here some of them will be briefly illustrated.

Deignan, Gabrys and Solska (1997) aims to analyse correspondences of conceptual metaphors in English and Polish. They sort linguistic metaphors extracted from the Bank of English (HarperCollins, 2005) in groups according to their source domain and ask Polish native speaker with a high English proficiency to translate them. Results show that in some cases the same mappings are present, sometimes with different realizations, and there are also cases in which a direct correspondence lacks, and a translation from English to Polish without paraphrasing is impossible. The conceptual metaphor RELATIONSHIPS ARE BUILDINGS was translated from English to polish without any problem, keeping the same linguistic realisation. On the other hand, the conceptual metaphor IDEAS ARE FOOD still exists in both cultures, but with different linguistic materials. In fact, the English *half-baked* is translated in Polish with a word that can be translated as *unripe.* This word – *niedojrzale* – on the opposite of its English counterpart, may be used metaphorically, as the Polish speakers did. Metaphors that caused difficulties in translation are *bring something (a fact, situation) home to someone* and *drive a message/idea home* (Deignan et al. 1997, p. 355), where they tried to translate *home* literally, as this metaphor does not exist in their L1.

Boers and Demecheleer (1997 in Deignan 2005) analyse metaphors in economics discourse in English, French and Flemish. Using frequency counts of the metaphors with different source domains, they found out that almost the same source domains are used, but with different frequency, according to the stereotypes of each nation.

The results of these two studies may be justified by Gibbs (1999 in Deignan 2005) assumption that universal categories are ‘culturally filtered’ (Deignan 2005, p.100).

A study that seems to justify differently the dissimilarity of metaphors used to talk about the same topic is Semino (2002). Comparing Italian and English newspaper corpora about the Euro, during its introduction in 1999, she finds out that their way of speaking about it reflects their different attitude about it. Italy joined the Eurozone and the source domains extracted from the Italian corpus are ‘JOURNEYS, SPORT, WAR and EXAMINATIONS’ (Deignan 2005, p.100), which reflect the desire and the fear of this change. On the other hand, linguistic metaphors in the English corpus as *lock* and *one size fits all* aim to stress the negative factors of inflexibility and shared currency. Therefore, differences in conceptual metaphors concerning the same topic are not only due to cultural differences, but also to differing ideas and attitudes towards the topic.

Following the studies described in this paragraph, a cross-linguistic study on femi(ni)cide would be interesting to understand whether the phenomenon is universal or culturally bound. This phenomenon is usually described as universal, as it is based on the asymmetrical power relation between men and women which characterised occidental society (Pinelo 2018, Johnson 1995). Moreover, this type of violence is also defined 'patriarchal terrorism' (Johnson, 1995) in order to distinguish it from common couple violence, a violence that could be perpetrated by both partners of the couple against the other partner and it is not justified nor pushed by the patriarchal society. The languages taken into consideration in this thesis – Italian and German – are both European languages, therefore it is possible to assume that the two cultures do have differences between them, but also many traits in common. Due to the many similarities – both states belong to the European Union, both are secular states, both have the same currency – it is not expected to find strong cultural differences in the metaphors used, although differences between the two cultures exist, and can be ascribed to their different history and geography. Therefore, what is expected to be found is similar conceptual metaphors both in the Italian and the German corpus, both because of the universality of the phenomenon and to the similarities between the two cultures.

2. Latent Dirichlet Allocation

This chapter will provide a brief introduction to the topic modelling task, with a main focus on the latent Dirichlet allocation algorithm (Blei, Ng & Jordan, 2003), along with an illustration of how topic modelling can be used for discourse analysis.

2.1 Introduction to topic analysis

Before starting with any further explanation, it should be useful to understand what *topic* means. In general, a topic may be the main theme of a text or a portion of text. For example, it is possible to affirm that the topic of the newspaper articles analysed in this thesis is femi(ni)cide. The leading idea behind these algorithms is that there may be more than one topic in a text. In this approach, topics are intended as the multiple themes in a document, not only as its main subject. For instance, topics as prostitution or web and social media reality may characterise a text about femi(ni)cide (Busso et al., 2019). A similar intuition is found in Navarro-Colorado (2018). Analysing a corpus of Spanish poems of the Golden Age period, he found evidence for affirming that topics extracted should be understood as the way the poet chose to write about other things rather than the main themes of the poems. He uses the literary term *Leitmotiv* to distinguish this different understanding of topic from the common idea of topic, namely the theme(s) that may be found in one or more texts. Therefore, a text may contain one or more topics, that researchers may extract following one of the techniques illustrated in this chapter.

Figure 2.1, taken from Zhai and Massung (2016, p. 331), illustrates the main ideas of extracting topics from a certain number of texts. An initial corpus is required, then the number of topics is decided, and it is found out which documents cover which topic and in which quantity.

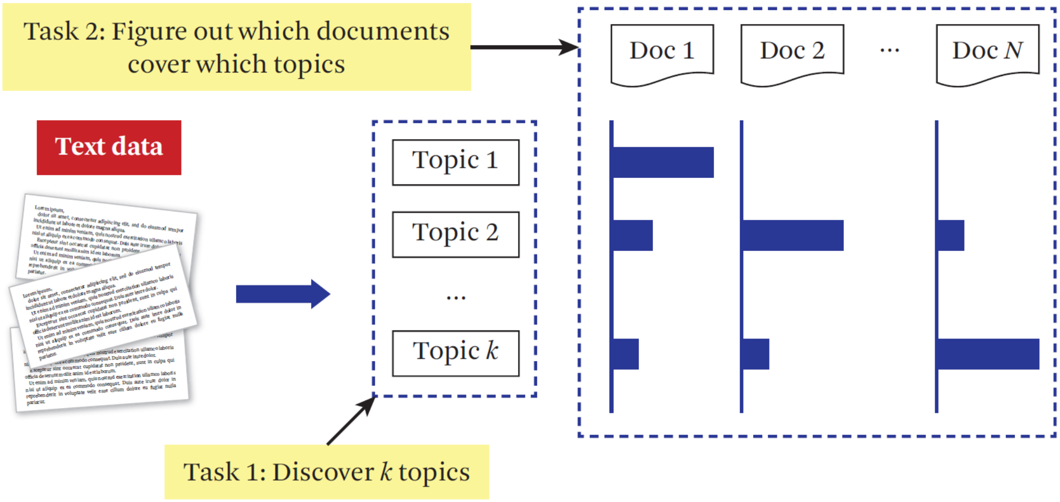

Task 2: Figure out which documents cover which topics
Text data
Topic 1
Topic 2
...
Topic k
Doc 1
Doc 2
...
Doc N
Task 1: Discover k topics

Figure 2.1: a scheme explaining the main procedure in topic extraction (Zhai & Massung, 2016, p. 331)

As illustrated in Zhai and Massung (2016), there are several ways to extract topics, according to how the notion of topic is understood. Topics may be intended as terms, namely as a word or a phrase to be found in the text(s). Following this type of methodology, topic is a word/phrase, and the researcher examines the distribution of one or more words in the corpus to look for the distribution of one or more topics – e.g., the term *sports* may be a topic, as well as the phrase *gender violence*. According to this strategy, looking for both terms in the corpus means to look for two different topics. When topics are intended as terms, the first step is to select these terms. This methodology involves finding out the frequencies of the words in the documents and picking the terms with the highest frequency. These words are now seen as topics. After the selection of the topic terms, the occurrences of each term in each document are calculated. This procedure, even if relatively simple, gives rise to several problems, bounded both to the ambiguity of words and to the impossibility to describe complex, specialised topics. The ambiguity of topics is caused by the ambiguity of words. This means that there are several words which can have more than one meaning and, while researchers are looking for a certain meaning, they might unconsciously accept the same word used with a different meaning. For example, when looking for the topic word *book* with the meaning of 'a set of printed pages that are fastened inside a cover so that you can turn them and read them'[5], researchers may accidentally accept the same term – *book* – but used with the verbal meaning of 'to reserve'. The second problem arises when dealing with more specialised

[5] Definition taken from the Oxford Learner's Dictionary (https://www.oxfordlearnersdictionaries.com)

texts that are about a subcategorization of the searched topic and do not use that specific word. As reported in Zhai and Massung (2016), a text about basketball that does not use the word *sports* results not to cover the topic *sports*, even though it is about sport. In order to solve these problems, topics may be intended as a distribution over words, namely a topic is not a single term anymore, but a list of words. Using more than one word to describe topics, allows to describe more complicated topics. Moreover, the words forming the topic have a probability weight, making thus easier to create a topic, bring together related words, and estimate the coverage of the topic. This approach also solves the problem of ambiguity, as probabilities of the same word may be found in more than one topic. Therefore, the term *sports* would appear in a list of words together with *game, basketball, ball* and so on (see Figure 2.2), while the term *book* could be hypothetically found in two different lists of topic words, according to the two different contexts of use. Another possible solution may be the use of distributional models (Lenci 2018, Boleda 2020) as topic model algorithms. The next paragraph introduces this type of algorithms, with a focus on latent Dirichlet allocation (Blei et al., 2003).

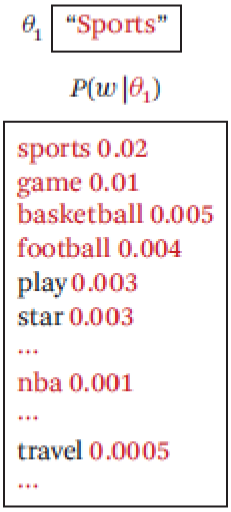


Figure 2.2: topic 'sports' extracted with topic model algorithm. Image taken from Zhai and Massung (2016, p. 336)

2.2 Latent Dirichlet Allocation

Distributional models, as topic model algorithms, are part of the broader field of distributional semantics. Distributional semantics is ‘a usage-based model of meaning, based on the assumption that the statistical distribution of linguistic items in context plays a key role in characterizing their semantic behaviour’ (Lenci, 2018, p.152). Distributional models are based on the distributional hypothesis, which assumes that words used in similar linguistic contexts have similar meanings. In other words, distributional models mathematically encode this hypothesis, translating each word of a given corpus in a probability value and represent words as vectors in a vector space, which are nearer if they appear in similar context and more distant when they are used in different context. This is, in general term, what is at the heart of probabilistic topic model.

As illustrated in Blei (2012), probabilistic topic model is a set of algorithms that analyses the words of written texts in order to find out the themes that characterise them, how these themes are connected and if and how they change over time.

An example of topic model algorithm is probabilistic latent semantic analysis (PLSA: Hofmann, 1999). This algorithm aims to mine multiple topics from text documents and to compute the coverage of each topic in each document. The point of weakness of this algorithm is its ability to function only with the training documents, it cannot extract a topic from an unseen document. Therefore, this algorithm may be applied only to the texts already used to train it. Training an algorithm before each use may result time consuming and requires more skills than simply run an algorithm. This problem led to the development of the latent Dirichlet allocation model (LDA: Blei et al., 2003).

As probabilistic latent semantic analysis, latent Dirichlet allocation aims to discover to which proportion documents of a given corpus contain a set of topics. As other distributional models, this algorithm is based on the contextual use of words, assuming that if certain words appear together in different documents, they belong to the same topic. The number of topics is not defined a priori, but has to be empirically decided by the researchers, depending on the aim of the study and the type of texts they are working with. What differentiate LDA to PLSA and gives to LDA the possibility to extract topic from unseen documents is that it is based on prior Dirichlet distribution, which defines a distribution over a vector of probabilities of topics. As illustrated in Blei et al. (2003), this unsupervised algorithm works on three levels: corpus level, document level, and word level. The corpus level is represented by α and β, variables θ represent the document level and are one per document, while variable z and w are word-based variables and are one for each word in every documents. As it is based on the naive Bayes assumption, it uses a bag of word as the starting point of its computation. This means that each document becomes an unordered set of words where the relationship among them is ignored and only their frequency in the document is kept (Jurafsky and Martin, 2020, ch. 4). In a few words, this algorithm works as follows: it runs through the words of each document, computes their probability and groups them together in the same topic according to whether they are used in the same documents. This process may be used to detect two or more topics from the corpus, according to the research requirement.

In the next section the opinions of scholars about the use of this algorithm in discourse analysis studies will be presented.

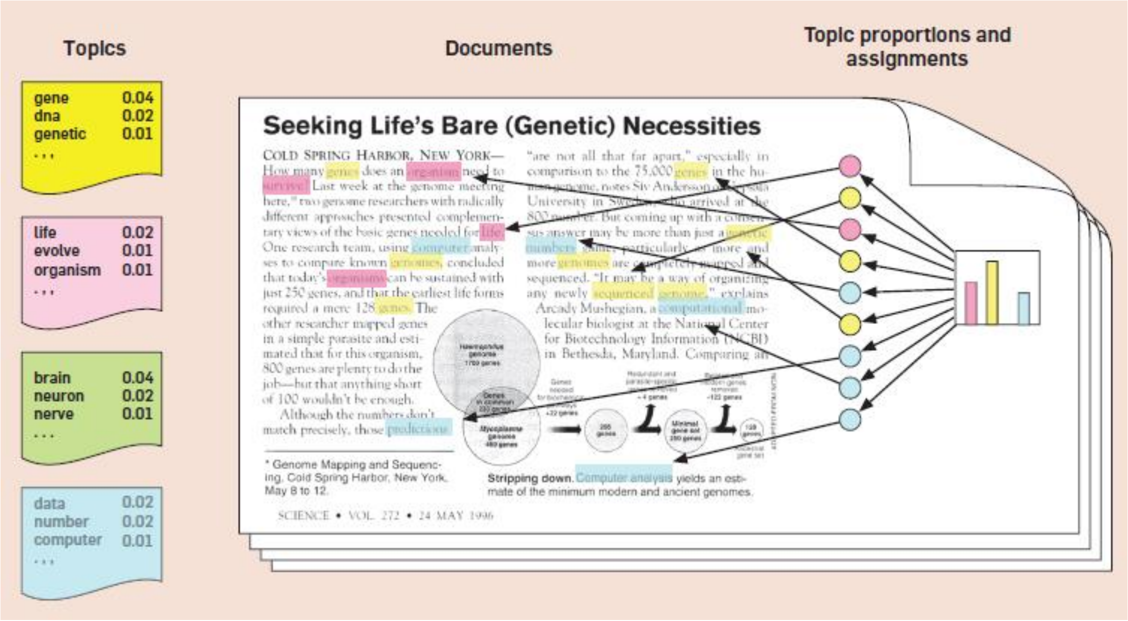


Figure 2.3: how LDA works. A certain number of topics is assumed to exist for all the texts in the corpus. The algorithm, then, chooses a distribution over topics and choose a topic assignment for each word. (Blei, 2012, p. 79).

2.3 Latent Dirichlet allocation in discourse analysis

In recent years, topic model algorithms have been used in the field of discourse analysis instead of the two main methodologies, namely *interpretivist text analysis* and *systematic qualitative coding*. As illustrated in Jo (2019), both methodologies involve a human reader and, therefore, the risk of a subjective judgment and a restricted number of texts to be analysed.

In studies where the interpretivist text analysis methodology is applied, researchers make a careful reading of the material and identify the characteristics and meanings of the data (Foucault, 2013). By carefully reading a large amount of linguistic material, Foucault analysed how madness was perceived in Renaissance period and how it is perceived in modern times, making a comparison between them. Following this methodology, researchers work alone, reading and analysing texts on their own. Systematic qualitative coding works similarly, with multiple human readers who analyse the same text and judge it according to a set of already fixed criteria. Research that follows this methodology is Slater et al. (2008). The study investigates the cancer discourse with five main research questions. The first research question aims to detect in which proportion cancer topics are covered in U.S media – newspapers, magazines, and television news. The second research question aims to investigate how frequently different cancer sites are covered in U.S. media. The third research question investigates how cancer site and cancer topic are related in U.S media. The fourth research question asks whether there is a correlation between the cancer site coverage in U.S. media and its incidence rate, while the fifth research question aims to investigate the correlation between the cancer site coverage and its mortality rate in the United States. In order to answer these questions, they chose six trained readers who read the different texts and classified them according to criteria provided by the researchers – nine topics for cancer topics and 25 types of cancer for cancer sites. After this part, the researchers analysed the obtain results to check how similar they were. In this research, coders read the texts to detect their main topic. As more than one reader analyses the same text, the possibility of subjectivity is avoided when their results are consistent. However, the problem of a limited amount of data still persists.

A way to avoid both the risk of subjective judgment and the restricted material to be analysed may be the application of text mining methods such as topic modeling algorithms in discourse analysis studies. These methods give the possibility to also solve another problem, that of reproducibility of the experiment, as it is not possible to reproduce results with the two close reading methodologies cited above.

According to Jo (2019), from the topics extracted by LDA algorithm it is possible to analyse the information used in language to express discourse, namely 'frequently appearing words […] and network structures between them' (Jo 2019, p. 332). Frequently appearing words are those words with and high probability rate which appear frequently in (at least) one portion of documents. On the other hand, the network structures may be seen as an incomplete information, as words assigned to the same topic often appear in the same document. This peculiarity of topic model algorithms – extracting more than one topic from the same text – is another point of strength for discourse analysis studies. Indeed, it is unlikely that a text treats only one theme, it usually mirrors more than one discourse, even if in a latent way, and LDA is able to detect them.

However, it also has weaknesses, as measuring the relation between words. This is one of the greatest points of weakness of LDA applied to discourse analysis. Conceiving a text as a bag of word cancel almost all the relations between words, such as cooccurrences in the same sentence and the word order, and take into account only the smooth relations of appearing, most of the time, in the same documents. This represents a loss for discourse analysis studies.

This negative feature is underlined also in Brookes and McEnery (2019). The two discourse analysts set up an experiment to investigate whether topic model algorithm, especially LDA, are suitable for discourse analysis studies. Their experiment is based on a corpus of 228,113 online comments patients gave about the National Health Service in England. First, they extract topics from this corpus, and then analyse with a close reading 20 documents of the corpus, making a comparison between the two results. Their aim is to comprehend whether the list of words (topics) obtained with LDA may function as a reliable indicator of the themes in the texts and the thematic coherence of the topics.

The number of topics to be extracted is set 20 and the words to be seen for each topic is 20 too. The text is pre-processed with the exclusion of functions words but not with lemmatisation – a typical step when dealing with prose text (Navarro-Colorado, 2018) – as it is not considered to be useful for discourse analysis. The comparison of the results obtained from the two parts of the experiment leads the authors to conclude that it is possible to infer the themes of a text with LDA topics, but not in an accurate way. Besides avoiding lemmatisation, they come up with the need to avoid the exclusion function words and punctuation to prevent a loss of meaning. Function words and punctuation are useful to understand whether the feedback was negative or positive and the change of attitude of the patient while writing the text. For what concerns the thematic coherence, they find out that only six topics out of 20 were in the 20 texts they analysed in the second part of the experiment. However, this low level of coherence between the extracted topics and the 20 documents analysed with close reading methods may be due to the fact they the authors had no idea about the content of the texts, and they may have inferred the theme of each topic wrongly. They analysed a huge amount of comments given by patients of the English National Health Service, which could have been about any theme – such as pregnancy, cancer, domestic accident and so forth. Nevertheless, they still consider LDA a good way to study discourse, as long as it is not used in isolation, but as a starting point for other qualitative studies.

Therefore, given the conclusion these scholars arrived to, it seems that using LDA on a specialised corpus, as the corpus used in this thesis, may still be a good idea. Unlike the above-mentioned study, the main theme of these newspaper articles is known – femi(ni)cide –, and topics extraction would give an idea of the collateral discourses used by the media when talking about this theme.

2.4 Topic analysis of femi(ni)cide

Looking for topic model algorithms applied to femi(ni)cide corpora gave a few results. A study where topic model algorithms are applied to a corpus with newspaper articles about gender violence is Busso et al. (2019). Their aim is to analyse the lexicon of woman as a victim in a multimodal corpus (both written than oral) and, in order to fulfil this aim, they extract metaphors, topics, and carry out a sentiment analysis on this corpus. The written corpus is composed of articles from four Italian national newspaper (*Corriere della Sera*, *La Stampa*, *Il Fatto Quotidiano* and *La Repubblica*) published from September 2016 to June 2017. They worked on a corpus with 17,840 lemmas and chose to extract ten topics. These topics belong to the semantic area of crimes (how it happened, and the weapons used), indicate places and people involved, or are legal terms. Topics related to gender violence prevention, prostitution and web were also found. From this research they conclude that journalists prefer to write in a simple and objective way, reporting all the elements involved in the crime, and giving a marginal role to the victims. Another study that should be useful to report here is Gius and Lalli (2014). Even if a close reading methodology is preferred to the application of topic modeling algorithms, they analyse how intimate femicide is reported in Italian newspapers. Their corpus counts 166 articles published in three national newspapers (*Corriere della Sera*, *La Repubblica*, *La Stampa*), both online and in print version, in 2012. These articles are about 53 cases of femi(ni)cide, as in 2012 occurred 124 cases, of which only 72 committed by an intimate partner, but not all of them appeared in the national press. A thematic analysis of the corpus revealed that the articles can be grouped according to three principal images evocated during the narration, namely ‘Romantic Love’, ‘Loss of Control’ and ‘Other Contextual Elements’ – in this group, articles that justify the crime through social condition, culture, mental illness, substance abuse and similar are gathered. The two sociologists conclude that the use of romantic love and loss of control in the narrative about femi(ni)cide contributes to the idea that these acts of violence are 'either unpredictable isolated acts of love preservation or […] the result of an irrational rage' (Gius and Lalli, 2014, p. 68). In both cases, the perpetrator is considered in a moment of emotional instability, removing responsibility from him, using words as 'raptus' or 'jealousy'. A similar outcome is achieved when the fault of the crime goes to the victim, who should have understood what the perpetrator was capable of. An explicit condemnation of the perpetrators is found only when they are characterised as 'monster' or 'beast', when they are de-humanised. In this analysis, they find out that Italian press is still bound to gender stereotype and avoid the theme of patriarchal society and the asymmetrical power relationship between victims and perpetrators.

Therefore, the main results obtained by these two studies are that the Italian press does not put the victim at the centre of the narration and avoid considering the power relationship between the victim and the murderer, it objectively reports elements involved in the crime, and is still bound to gender stereotypes.

Part of this study aims to extract topics from a corpus built with Italian and German texts through the application of the LDA algorithm.

The corpus used in this study is a collection of newspaper articles from 2019, both from Italian and German newspapers. The two subsets of the corpus will be analysed separately and then the results will be compared, in order to understand how the attitude towards the phenomenon of femi(ni)cide varies in these two countries. It was difficult to find topic analysis studies on German written production about femi(ni)cide, hence the results will be analysed taking into account these two studies on Italian newspaper articles on the theme. For what concerns the Italian corpus, different results may be justified both by the different years considered – as the corpus of Gius and Lalli (2014) is composed by 2012 newspaper articles and Busso et al. (2019)'s corpus by 2016/2017 newspaper articles – and by the different newspapers – as the Italian corpus contains, other than the newspapers already used in the two previous studies (*La Repubblica, Il Fatto Quotidiano, La Stampa*), also *Il Resto del Carlino* and *Il Giornale*. Regarding the German corpus, due to the absence of previous studies on this language, different results should be justified mainly as a different attitude towards the theme and the different years of the corpora. The results obtained with the topics extraction will be compared between the two subcorpora.

3. Corpora and web-based corpora

As already introduced in the previous chapters, this project is a corpus-based study. Therefore, in order to better contextualise the study, a general introduction to what a corpus is and a specific introduction to the corpus used in the study is provided.

According to Gries and Newman (2013, p. 258), a ‘corpus is a machine-readable collection of language used in authentic settings/contexts’. Hence, a corpus is a set of language material, which is not originally created to be part of the corpus, and that is gathered and encoded to be read by a machine – a computer. A further definition, which can be suitable for the aim of this chapter and works as an extension of the previous definition, is in Kilgarriff and Grefenstette (2003, p. 2), where a corpus is still considered a collection of language, but they also specified its aim, that of being ‘an object of language of literary study’.

The corpus used for this study is in line with both the previous definitions. In fact, this corpus may be considered as a set of language created for a natural context, as defined by Gries and Newman (2013), since the texts collected for the corpus have the original purpose of communicating and describing events. Their creation is not bound to the creation of the corpus, rather they were published on different Italian and German newspapers in 2019. Moreover, it is ‘a machine-readable collection’ as the language material is encoded in order for the computer to read it and work on it. Finally, it also respects the definition given by Kilgarriff and Grefenstette (2003), as it is the object of a linguistic study. Therefore, this corpus is a collection of natural occurring language, built to be read by a machine and for a purpose, namely the linguistic study.

A corpus may be classified by genre, medium, historical variation, language, text integrity and annotation (Lenci et al., 2005).

Classifying a corpus by genre means to define whether the corpus is a specialised or a general corpus. For instance, the British National Corpus (BNC Consortium, 2007) is a general corpus as it contains texts from different genres, such as newspapers and academic texts, spoken and written documents, while an example of a specialised corpus is The Michigan Corpus of Academic Spoken English (MICASE: Simpson et al., 2002), as it is composed only of transcribed academic speeches. The medium category concerns the medium of the corpus documents. Namely, a corpus can be considered written if all the texts are written, or it can be considered a corpus of transcribed speech if the texts were firstly oral texts and then transcribed to form the corpus, while it can be classified as mixed or multimedia when it is composed both by written and transcribed speech or, for example, by written and audio files, e.g. the British National Corpus (BNC Consortium, 2007) is mixed as it contains samples of written and spoken English, while the Brown Corpus (Francis and Kucera, 1979) is written as it includes written texts only. For what concerns the historical variation, a corpus may be diachronic, when it is based on texts from different periods of time, or synchronic, if it is created with texts of the same period. For example, the Corpus of Historical American English (Davies, 2010) is diachronic as it consists of texts published from 1810 to 2009, while the Brown Corpus (Francis and Kucera, 1979) is synchronic as it contains American English texts published in 1961**.** In addition, the documents of a corpus may all be created in the same language, and thus building a monolingual corpus, or in two or more languages. When dealing with more than one language, a corpus may be classified in two ways: or as parallel corpus – if it presents original documents in one language while the others are the exact translations of those documents – or as a comparable corpus – if the documents are on the same topic but born independently, not as a translation of a common original text. An example of parallel corpus is the European Parliament Proceedings Parallel Corpus (Koehn, 2005), a corpus made of extracts from the proceedings of the European Parliament. The corpus contains texts in 21 languages, which are the translations of the English original documents. Then, an example of comparable corpus is The International Corpus of English (Greenbaum, 1991), a project which includes different states where English is used as national language – both as first or second official language –, with the goal of documenting different national and regional varieties of English.[6]

Finally, the last category concerns the annotations added to the corpus documents, namely whether the corpus presents raw texts, or if the documents have additional information, such as structural information – as the name of the author, the title or the chapter number – or linguistic annotations – as the tagging of the part of speech of the tokens.

3.1 Web as/for corpus [7]

In the last century, the idea of corpus was linked to an image of something static and immutable. As an example of these more classic corpora, the British National Corpus

[6] It concerns varieties from 27 states and geographical regions, namely Australia, Bahamas, Canada, East Africa (Kenya, Malawi, Tanzania), Fiji, Ghana, Gibraltar, Great Britain, Hong Kong, India, Ireland, Jamaica, Malta, Malaysia, Namibia, New Zealand, Nigeria, Pakistan, Philippines, Puerto Rico, Scotland, Singapore, South Africa, Sri Lanka, Trinidad and Tobago, Uganda, USA.

[7] The definition is taken from Gatto (2014, p. 37) as the web may be used both as a source for directly study language and as a source for downloading files to create online corpora.

(BNC Consortium, 2007) may be taken into account. This corpus was built between 1991 and 1994 and its documents are still the same. As declared on the official website of the BNC, three different versions have been published without the adding of new texts, they simply revised the corpus before the publication of each edition.

However, in the new millennium, the idea of corpus switched from an immutable object to a body of texts which is not static at all, thanks to the spread of the use of the world wide web. According to Kilgarriff (2001, p.343) ‘the corpus of the new millennium is the web’, as, due to the new texts added every day, it seems a rich resource to investigate different types of linguistic phenomena.

The possibility for the web to be used as a corpus is further developed in Gatto (2014), where four different ways of conceiving the web as a source to build corpora are reported. Three of these methods may be broadly summarised as using the web as the substitute of a corpus, and they work retrieving and analysing linguistic phenomena directly from the web. The fourth method, instead, is designated as ‘the Web as a corpus shop’ (Gatto, 2014, p. 37) and it is the method employed in this thesis to build the corpus. Considering the web as a shop means to select and download texts from the web to create an offline corpus to be used with a particular purpose. Therefore, the final result is a collection of online documents available offline. Following this method, the web works as a resource to download texts.

According to Kilgarriff (2001) and Gatto (2014), the web is a great and precious resource for linguistic studies based on corpora. At the same time, however, it may be a dangerous environment to observe linguistic phenomena. This is because, using commercial search engines, no lemmatisation or part of speech tagging is provided, there is a limited search syntax, and its results are for webpages and not for instances (Kilgarriff, 2007). Moreover, because of its characteristics, it may become dangerous to consider the web as a corpus. Working with the web as a corpus may be risky because of its indefiniteness, as nobody knows the exact number of texts available online, and because the web is not projected following any linguistic idea. (Sinclaire 2005 in Gatto, 2014).

3.2 Introduction to BootCaT

Taking in consideration these reflections, the most suitable way to use the web to investigate linguistic phenomena seems to be using it as a resource to download texts. A useful tool for this purpose is *BootCaT* (Baroni and Bernardini, 2004), that also is the tool used to create the corpus for this study.

BootCaT stands for *Bootstrapping Corpora and Terms* and is a freely available toolkit to build corpora from the web. The original idea is that of creating a corpus starting from a few initial words. The main procedure described in Baroni and Bernardini (2004) is illustrated in figure 3.1 and may be explained as followed. Firstly, BootCaT receives a set of selected words and searches them through the web, downloading a corpus and terms from the web. From these results is then possible to extract a new list of unigrams to be used as seeds to repeat the process, in order to create a new and, probably, more precise corpus. Then, it extracts multi-word terms from the created corpus and list of words. After a theoretical description, Baroni and Bernardini (2004) also provide practical examples of this procedure. They create a corpus and a term list based on the words of an English psychiatric article – Fleisher et al., 2002 – which are not in the Brown corpus (Kuˇcera and Francis, 1967). These six words come from the abstract of the article and are *dissociative, epilepsy, interventions, posttraumatic, pseudoseizures, ptsd.* They repeated the process twice. The first time, BootCaT gave a corpus of about 396,000 words. Form this corpus, they extracted 40 terms which worked as new seeds for the second iteration launched to build the second corpus. Finally, they merged the two corpora to create a final corpus. They then extracted a unigram term list based on term and document frequency and used this list, and the corpus, to extract multi-words expressions. At the end of this process, they evaluated the corpora and the term lists following three steps. Firstly, they evaluated 30 webpages chosen by chance for each corpus. From these webpages, informativeness, relatedness to the topic and reliability are evaluated. The evaluation is positive when the people who read the webpage find it to be informative, reliable, and coherent enough in topic and register with the theme of the corpus. A second step is the evaluation of unigram and multiword terms. A list of 100 unigram terms and a list of 100 multiword terms is extracted from both subcorpora to be evaluated according to their relevance and if they are written correctly. The evaluation is positive when these terms are relative to the topic and are well-formed. The last step is the manual collection of the terms both from the English text and from the Italian translation and a comparison of these lists with those retrieved by BootCaT. For what concerns the first evaluation step, Italian webpages seemed to be more informative than English pages, even if sometimes out of topic. On the other hand, the English unigram list resulted to be better formed than the Italian list, as for the multi-word term lists.

The illustrated procedure is not the only possibility offered by BootCaT to create a corpus. The user has the alternatives to directly provide the software with a list of tuples, with a text file with a list of URLs, or provide the software with a folder from which documents should be extracted to create the corpus.

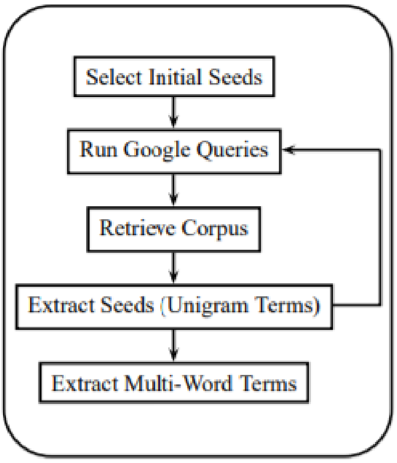


Figure 3.1: the BootCat process to collect material from the web to build a corpus. (Baroni and Bernardini, 2004, p. 1314).

3.3 The Italian-German corpus

This research is carried on a comparable corpus made of an Italian and a German part. The two corpora will be presented separately, after an introduction of the common aspects – according to the corpus features already presented in the first paragraph of this chapter.

This corpus may be defined a comparable corpus as the topic of the two sub-corpora is the same – femi(ni)cide cases reported by newspaper articles. Both the Italian and the German corpus are composed of newspaper articles reporting femi(ni)cide episodes happened in 2019 and published in the same year. For each sub-corpus, articles are taken from five different national newspapers – five Italian national newspapers for the Italian corpus and five national newspapers for the German corpus.

Here a description of how the selection of the newspapers took place. The first step concerned the choice of the Italian newspapers, and after them the German newspapers were selected. In general, the point was to choose newspapers with different political ideas. A first parameter was whether it is possible to download the full text of the needed articles or not. According to the possibility of downloading the whole text, a further selection was made, restricting the number of newspapers to five, namely *La Repubblica*, *La Stampa*, *Il Giornale*, *Il Fatto Quotidiano*, *Il Resto del Carlino.* The selection of the

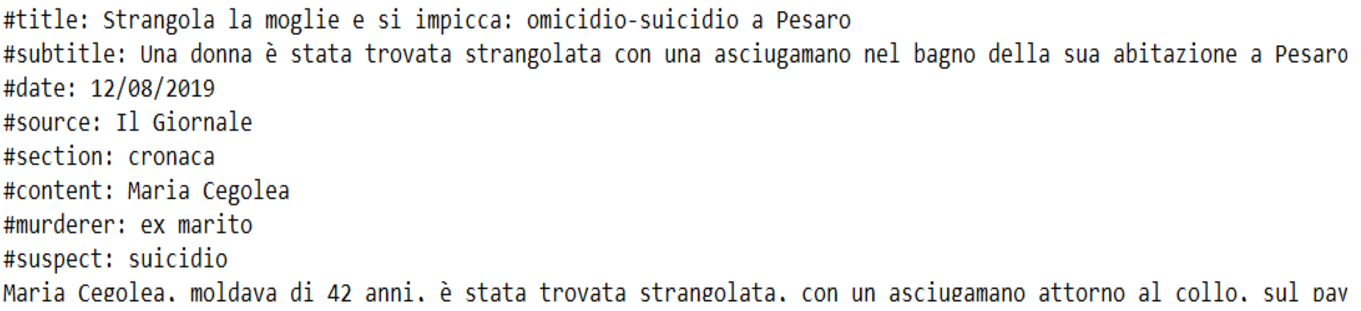
German newspapers has been made through a sort of parallelism with the Italian newspapers. Thus, *Die Zeit* was chosen as it has a left-wing political opinion similar to *La Repubblica*, *Süddeutsche Zeitung* as it has a liberal opinion similar to *La Stampa*, *Die Welt* because it is conservative as *Il Giornale*, *Frankfurter Allgemeine Zeitung* as it has an editorial policy similar to that of *Il Fatto Quotidiano*, and *Bild* as it is right-wing as *Il Resto del Carlino*.

Hence, the corpus may be defined as comparable – when considering the two languages and the content of the two sub-corpora –, synchronic – as all the documents are published in the same year –, specialised – as it contains only newspaper articles –, written – as all the documents are originally written – and it is formed with integral texts.

The corpus is then enriched by manually annotating, for each document, the following structural information:

- Title of the article
- Subtitle of the article
- Date of publication
- Source – which is the newspaper
- Section – which is the section of the newspaper
- Content – the case of femi(ni)cide reported
- Murderer – the relationship the murderer had with the victim
- Suspect – whether the murderer was accused[8], sentenced or committed suicide
- Paragraph title.

How this information is added to the text is shown in figure 3.2, a picture of a document from the Italian section of the corpus. I manually added these tags to the documents of the corpus.



```
#title: Strangola la moglie e si impicca: omicidio-suicidio a Pesaro
#subtitle: Una donna è stata trovata strangolata con una asciugamano nel bagno della sua abitazione a Pesaro
#date: 12/08/2019
#source: Il Giornale
#section: cronaca
#content: Maria Cegolea
#murderer: ex marito
#suspect: suicidio
Maria Cegolea, moldava di 42 anni, è stata trovata strangolata, con un asciugamano attorno al collo, sul pav
```

Figure 3.2: This figure is a screenshot of one document of the corpus, showing the structural annotation added to each document of the corpus.

[8] The tag *accusato* (accused) is used both in cases where there is no information about the verdict, both in cases where the suspect was released

A further shared characteristic between the two subsets of the corpus is the pre-processing of the texts, a procedure required in order for the texts to be used to extract metaphors and topics, the two goals of this study.

The pre-processing of the texts begins with the download of the needed tools, such as the Stanza library (Qi et al., 2020), JSON (Williams, 2015) and os[9]. JSON – JavaScript Object Notation – is a text format encoded in UTF-8 and os is a module used to interact with the operation system of the machine. Stanza is a Python natural language processing toolkit that supports 66 human languages, among which Italian and German. One of the advantages of this toolkit is the possibility to give raw texts as input and receive annotated texts as result. Its pipeline has also been tested and evaluated on 112 datasets in order to give a high-level accuracy (Qi et al., 2020). In this research, Stanza is used to tokenise, identify multi-word tokens and tag each token with its part-of-speech. Stanza application is the last of four main steps. The previous main steps are the deletion of meaningless words, proper nouns, and structural information. As first step, a list of stop-words made of lexical verbs, prepositions, articles, and common adjectives and adverbs is downloaded from the Stopwords-ISO package by Diaz and Suriyawongkul[10] available on GitHub, which contains both an Italian and a German list. This list of stop-words is then transformed into a set and enriched with the name of the victims and the perpetrators of the crime, where available. Before the deletion of the set from the texts, a list with the raw texts without the lines containing the structural information is created, with the help of the os functions which allows to read files stored into operation system folders. This list then undergoes the tokenisation process through Stanza and the words contained in the set – meaningless words and names of victims and perpetrators – are deleted. At the end, this list is saved as a JSON file and becomes the corpus of the study. This process is made twice, once for the Italian corpus and once for the German corpus.

The preprocessing is the same for both tasks – topics extraction and metaphors extraction –, the only difference in preparing the documents is in the exclusion of function words and proper nouns. As already illustrated, a list of function words – such as prepositions, articles, conjunctions, and so on – is deleted from the documents, together with the proper nouns of victims and perpetrators. This passage is done only when preparing the

[9] Lib/os.py

[10] https://github.com/stopwords-iso/stopwords-iso

documents for topic modeling, while, when preparing the texts for metaphors extraction, only the additional information are excluded.

After the illustration of the common features between the two corpora, a more detailed presentation of the two corpora follows in the next paragraphs.

3.3.1 Italian corpus

The Italian corpus consists of newspapers articles from *La Repubblica*, *La Stampa*, *Il Giornale*, *Il Fatto Quotidiano*, *Il Resto del Carlino* written and published in 2019. In order to collect the material, two datasets have been checked, compiled with the name of the victims of femi(ni)cide and other information. The need to consult two databases depends on the unofficial nature of the data. Actually, the name of the victims in these datasets are obtained by research on newspaper articles. Therefore, it is possible that there are some cases missing in both lists. The two datasets are provided by *Casa delle donne per non subire violenza, ONLUS* and *ProsMedia – Osservatorio sul femminicidio –*, which together collected 95 cases of women killed by partners, ex partners, friends, relatives, or acquaintances. One of the 95 instances reported in these databases has been excluded from the study as the killer murdered all the members of the family, not only the woman. *Casa delle donne per non subire violenza, ONLUS*[11] is an Italian association born in 1990 with the aim of opposing gender violence and giving aid to the victims through different projects, as the creation of an anti-violence centre and safe houses to host victims of gender violence. The database provided by *Casa delle donne per non subire violenza, ONLUS* does not only gives the victim names, but also other data as the date of the killing, the town where it occurred, the age of the victim, the type of relationship between the victim and the murderer, the cause, and the number of children the woman had. It reports 95 cases.

The other dataset is provided by *ProsMedia – Osservatorio sul femminicidio*[12] –, an association born in 2008 with a focus on media analysis and media education. It has many goals, both in a pedagogic and scientific direction, among which the goal of opposing gender violence, cyber bullying and hate speech. Similarly to the database provided by the ONLUS, this database also reports date, place, name of the victim, the type of relationship between the victim and the murderer and the name of the murderer. It has 86 cases.

[11] https://www.casadonne.it/

[12] https://prosmedia.org/

The research for the articles started from the data acquired from these lists. Two main methodologies were used. The first method is that of writing in the Google search space the complete name of the victims, in inverted commas, sometimes with the adding of the word *femminicidio* or with the city were the crime occurred and/or the year – 2019 –, and, to specify the newspaper, the addition of the part *in: website of the newspaper* – for instance, *"Name Surname" and feminicide in: repubblica.it*. The other method is to directly look for the articles on the website of the newspaper, in its search bar, typing the complete name of the victim in inverted commas, sometimes with the year of the crime – 2019 – and the city where it happened.

After having identified the articles, every link was copied and pasted in a text file, for a total of five text files, one for each newspaper. Each one of these files is then passed to BootCaT, uploaded in the section that allows to upload a file of links and returns text files reporting the text downloaded from each webpage. These texts are then tagged according to the structural annotation described in the previous section (paragraph 3.3) – title, subtitle, date, source, section, content, murderer, suspect.

The Italian subcorpus contains five newspapers and 94 femi(ni)cide cases, for a total of 512 articles and 190,229 tokens. In a separate notebook, a code is written to count how many femi(ni)cide cases are reported in all the newspapers in analysis and how many cases are not reported at all. Only nine cases out of 94 are reported in all five newspapers, while there is no article at all regarding 17 cases of femi(ni)cide.

3.3.2 German corpus

The creation of the German corpus is slightly different from that of the Italian corpus. Similarly to the Italian corpus, also for this part of the corpus an initial list with the cases of femi(ni)cide has been crucial. Unfortunately, it was not possible to find an online database with the required data. The only solution seemed to be to personally contact the German associations active in the field of gender violence. After a few negative answers because no association had a similar list, people behind the project *Feminizidmap* decided to share their database. *Feminizidmap*[13] is a German project born in 2018 with the intention of documenting the killings of women and girls at the hands of men in Germany. Their database contains data as the name of the victims, the age of the victims, the age of the murderers, the kind of relation the murderer had with the victim,

[13] https://feminizidmap.org/

the type of femi(ni)cide – e.g., intimate femi(ni)cide, non-intimate femi(ni)cide, female infanticide**,** and so forth –, and the place where the crime took place. All these data are unofficial data, as for the Italian lists, and are gathered by the media, thus there are same data that are absent for some cases – for instance, no name or no age is reported for some victims. The number of the cases reported in this list is 175. Nine cases are excluded from the study as the killing concerns all the family members, not only the female member, or because the killing occurred as a consequence of a robbery.

After receiving this list, a member of the *Feminizidmap* project also shared the newspaper articles she downloaded for another project. They are from three of the five newspapers taken into analysis in this study – *Bild*, *Frankfurter Allgemeine Zeitung*, and *Süddeutsche Zeitung*. The files were sent in PDF format. As they should be text files to be investigated, they were firstly changed from PDF to image, and then from image to txt format. The transition from image to txt file is made thanks to the OCR engine Tesseract (Smith, 2007). This tool can recognise more than 100 languages, among which Italian and German, and takes images as input to return them as text in different formats. Tesseract version 4.10 for windows 64 bit is here used to transform images into txt files. In a separate notebook, version 4.10 of Tesseract with the function of returning texts lines from an image is downloaded and installed in an assigned folder. Then, a function is written with the aim of reading the images of the German articles, transforming them in txt files and creating a folder where these files are then deposited.

For the other two newspapers, the methods illustrated for collecting the Italian articles is used. The only difference is that the name of the victims is not part of the query used to search for the articles, but only the name of the city where the crime occurred, the year when it occurred and the age of the victim. This choice is due by the fact that the victim's name is hardly ever used by the German press and, when it is reported, the surname is omitted most of the time, or it is reported only its initial letter.

After collecting all the needed articles, they are then tagged with the structural annotation already illustrated – title, subtitle, date, source, section, content, murderer, suspect. The German subcorpus contains five newspapers and 166 femi(ni)cide cases, for a total of 1,010 articles and 207,313 tokens. In a separate notebook, a code was written to count how many femi(ni)cide cases are reported in all the newspapers in analysis and how many cases are not reported at all. Only two cases out of 166 are reported in all five newspapers, while there is no article at all regarding 16 cases of femi(ni)cide.

3.4 The Italian-German corpus: statistic description

The obtained corpus results to be unbalanced, as the German subcorpus contains more tokens than the Italian subcorpus – 207,313 tokens for the German corpus against the 190,229 of the Italian subcorpus. This differentiation in length is due by the method used to collect the documents for the corpus, namely creating the corpus looking for newspaper articles about femi(ni)cide cases, which happened in a different number in the two states. Also the distribution of femi(ni)cide cases over newspapers results to be different in the two subcorpora. As it can be seen form the graphs (figure 3.3 and figure 3.4), 18% of Italian cases (17) are not present in any newspaper, against the 10% of the German cases (16). On the other hand, 10% of Italian cases (9) are covered by all five newspapers, while only 1% of the German cases (2) are reported by all the five German newspapers taken into account. The majority of German femi(ni)cide cases (94 cases, which are the 57% of the German cases) are reported by two newspapers, while the number of Italian newspapers reporting more cases is three and they report 24% of the Italian cases occurred in 2019 (9).

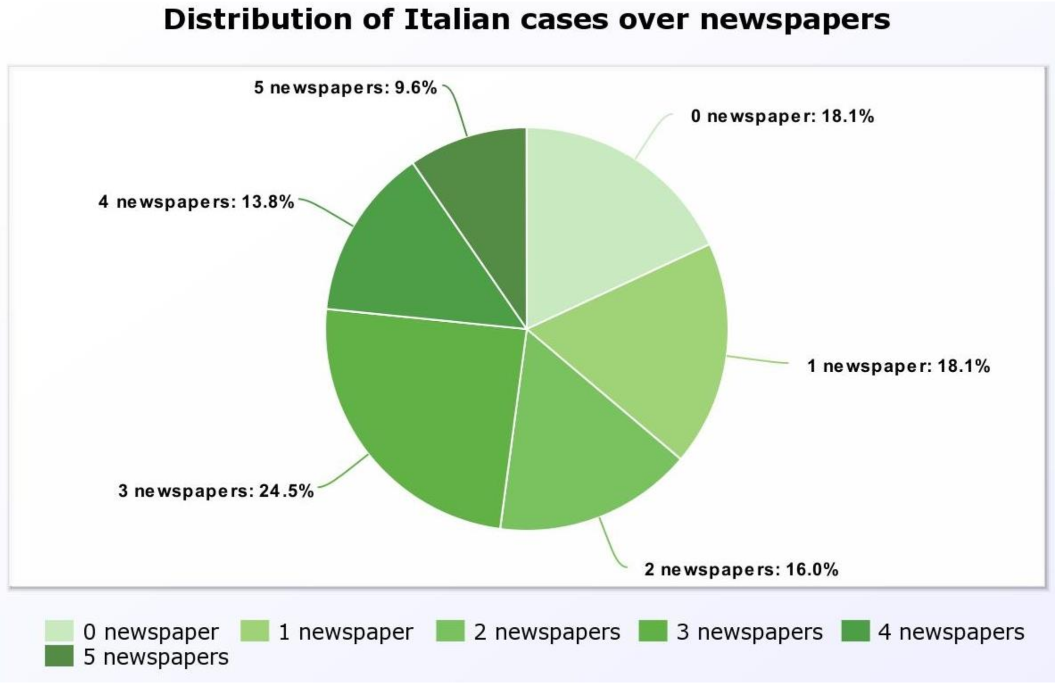


Figure 3.3: A graph showing the distribution of Italian femi(ni)cide cases over the newspapers in the corpus.

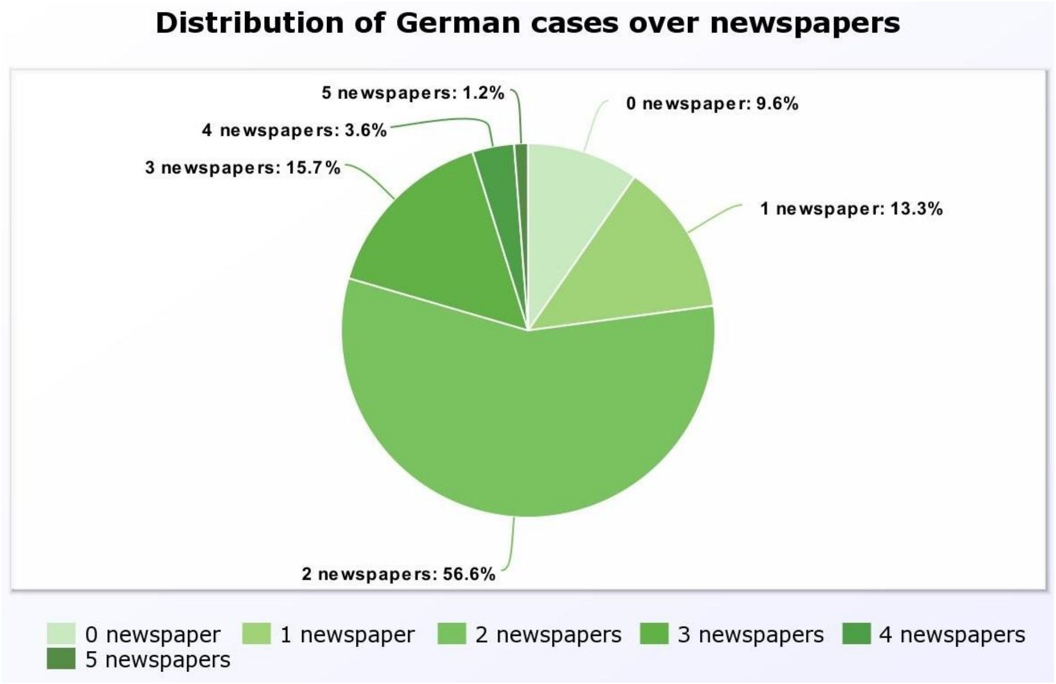


Figure 3.4: A graph showing the distribution of German femi(ni)cide cases over the newspapers in the corpus.

3.5 The Italian-German corpus: lexical analysis

After this statistic description, a preliminary lexical analysis on the two subcorpora may reveal something more on the content of the two subcorpora.

From the word clouds of the top 50 words of the two subcorpora (figure 3.5 and figure 3.6) it arises that the main words are those belonging to the semantic field of crime and the people involved in the event. In both subcorpora, the words *police* – *Polizei, carabinieri* – and woman – *Frau, donna* – seem to be the more frequent, but also words concerning the investigation process and the death of the victim appear. The words are extracted from the same corpus version used for topics extraction, namely with no closed class words and no proper nouns. It seems that no other word is included except those words which are strictly related to the crime event.

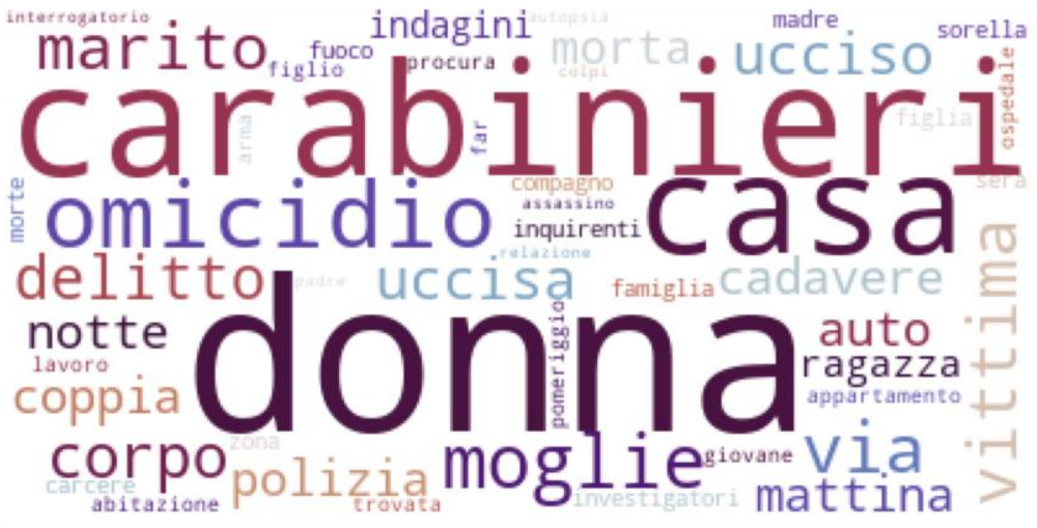


Figure 3.5: Word cloud including the top 50 words of the Italian subcorpus.

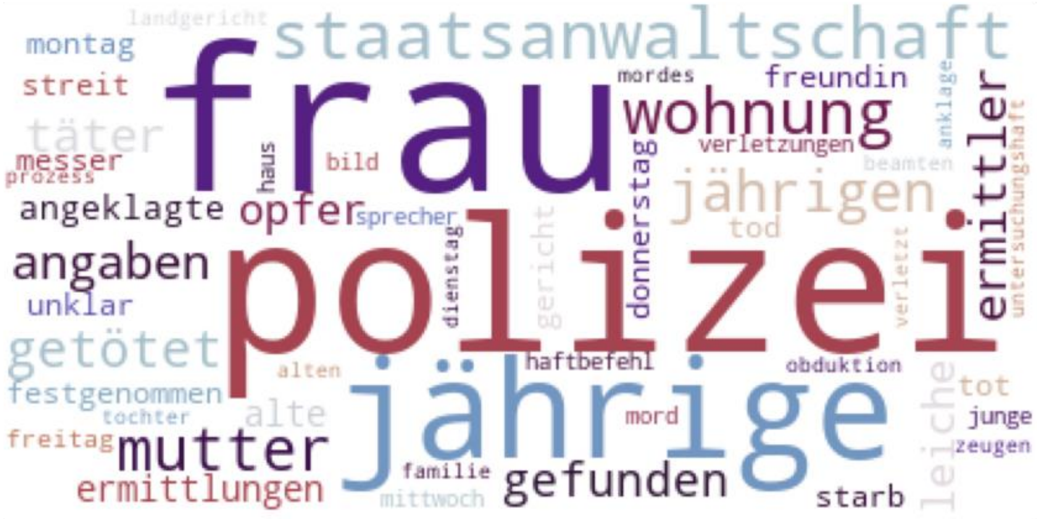


Figure 3.6: Word cloud of the top 50 words of the German subcorpus.

Collocations of the most frequent words are extracted in order to better understand the content of the corpus. The processed corpora are uploaded on Sketch Engine[14] (Kilgarriff et al. 2014, Kilgarriff et al. 2004) to extract collocations and create the following graphs through the function *word sketch*. Sketch Engine is an online tool which allows to analyse both ready to use corpora and new corpora created by the user. Among its main functions, the functions employed for this study are: the possibility to tag the corpus with the part of speech of each token; the possibility to search for concordance and/or collocations of the needed lemmas/words, using different types of parameters, as the possibility to specify the lemma, the wordform or the tag of the needed token; an analysis of the corpus

[14] http://www.sketchengine.eu

(e.g., information about the vocabulary size); the possibility to create some charts based on the relationships the words have in the corpus.

Collocations of *donna, carabinieri, uomo* and *vittima* ('woman', 'police', 'man', 'victim') are extracted from the Italian subcorpus.

The type *donna* cooccurs with words connected to the violent act – *violenza, scomparsa, vittima* ('violence', 'missing', 'victim') –, family members – *marito, figlio, compagno* ('husband', 'son', 'partner') – and also concerning her sexual orientation – *lesbica* ('lesbian'). Collocations of the term *uomo* are extracted as this word represents the counterpart of the woman in this type of event, as the murder usually is a man. Therefore, collocations of *uomo* are extracted to see the different contexts of use of the two words. Cooccurrences of the type *uomo* do not present any point in common with *donna*, and they are mainly focused on law enforcement and a word which seems to be unexpected according to the other words extracted, *belva* ('beast'). These results are confirmed by plotting the *word sketch difference* of *donna* and *uomo*. *Word sketch difference* is a function of Sketch Engine which allows to compare two words according to their collocations in the corpus. No word connected to violence appears near *uomo*, but they occur near *donna.* To further deepen this lexical analysis, collocations of the type *vittima* ('victim') are analysed. Terms appearing in its collocations mainly concern violence. The term femi(ni)cide appears, together with the word *carnefice* ('executioner') and words which make clear that the victim is a woman – *picchiata* and *uccisa* ('beaten' and 'killed' with the feminine ending -*a*). The other word which appears to be the most used in the Italian corpus is *carabinieri*, which mainly cooccurs with terms referring to law enforcement. Schematic representations of these four words and their collocations are added below.

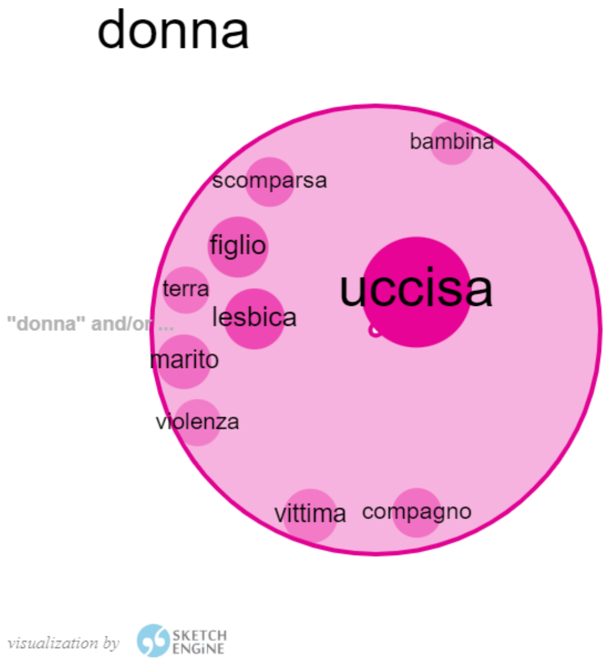


Figure 3.7: Bubble chart of the collocations of *donna*.

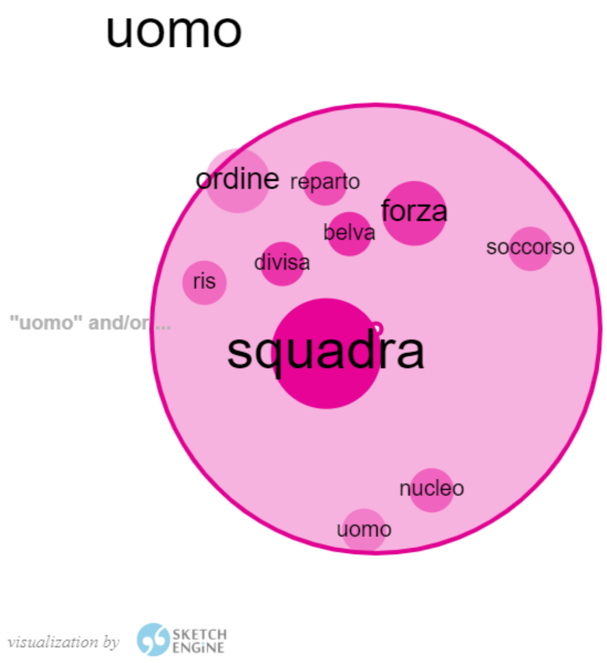


Figure 3.8: Bubble chart of the collocations of *uomo*.

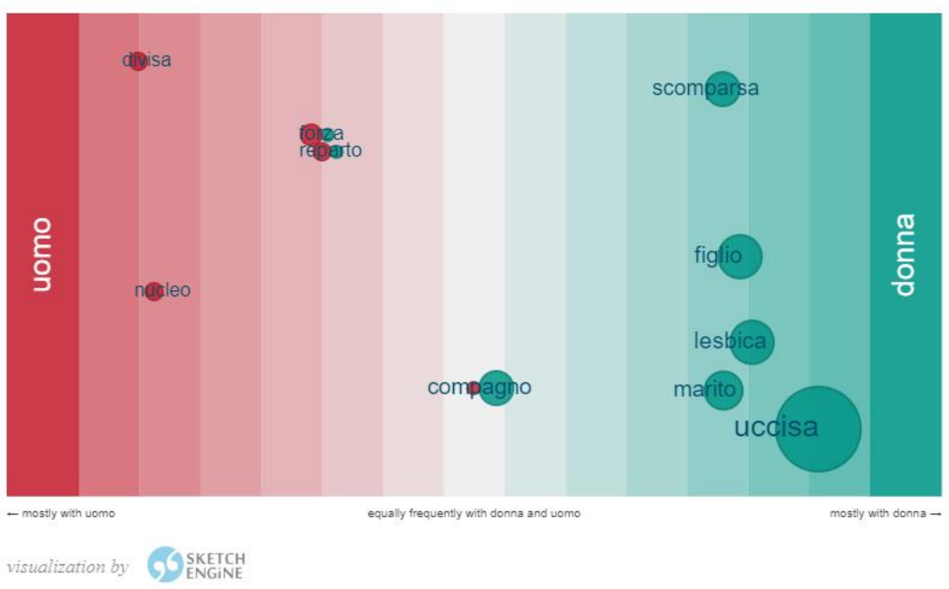


Figure 3.9: Word sketch difference representation of the type *uomo* and the type *donna.*

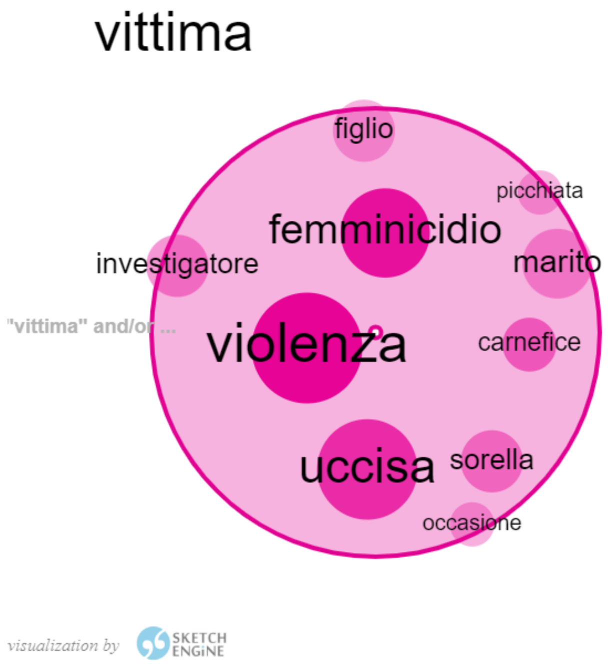


Figure 3.10: Bubble chart of the collocations of *vittima.*

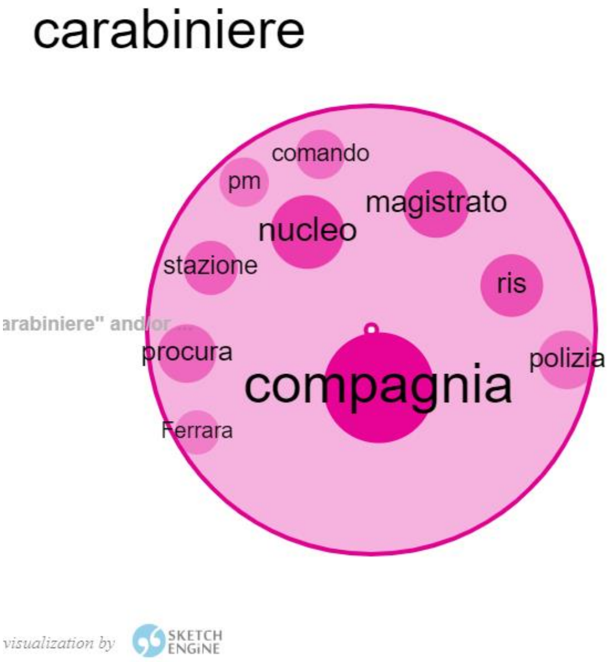


Figure 3.11: Bubble chart of the collocations of *carabiniere.*

A similar analysis is conducted for the German subcorpus too. Collocations of the same words extracted from the Italian subcorpus are extracted for the German equivalent. The type *Frau* (‘woman’) cooccurs with words referring to death and crime. The majority of the collocations belong to the semantic field of death, as the verbs *sterben* (‘to die’), *getötet* (‘killed’) or the noun *Leiche* (‘corpse’). The counterpart role in the crime event, *Mann* (‘man’), cooccurs with words concerning crime – *Untersuchungshaft, Pflichtverteidiger* (‘preventive detention’, ‘public defender’) – and mental health – *psychiatrisch* (‘psychiatric’). These different types of collocations arise also with the *word sketch difference* function, since words nearer the term *Frau* are those concerning death and violence – *tot, Leiche, verletzen* (‘dead’, ‘corpse’, ‘injure’) – but also age – *alt, jung, jährig* (‘old’, ‘young’, ‘years old’) –, while words nearer to *Mann* concerns legal action – *Pflichtverteidiger.* The term *Opfer* (‘victim’) occurs with words centred on the people involved in the crime, concerning their gender – *männlich, weiblich* (‘male’, ‘feminine’) –, their origins – *stammen*, Kosovo (‘come from’) – and words typical of the crime – *Tat, Täter, verletzt* (‘criminal act’, ‘perpetrator’, ‘injured’). Also the term *Polizei* (‘police’) occurs with words connected to crime and death, such as *umbringen* and *töten* (‘to kill’).

A schematic representation of the numbers of the corpus split for the two subcorpora, namely their dimension and their vocabulary size, and the number of femi(ni)cide, newspapers and documents for each subcorpus (table 3.1) follows the charts concerning the German subcorpus. Then another chart is reported (table 3.2), showing the structural information added to the documents.

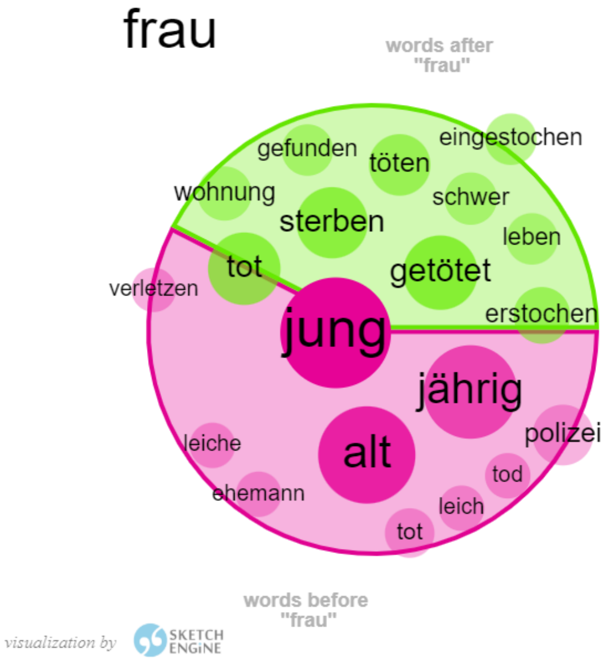


Figure 3.12: Bubble chart of the collocations of *Frau*.

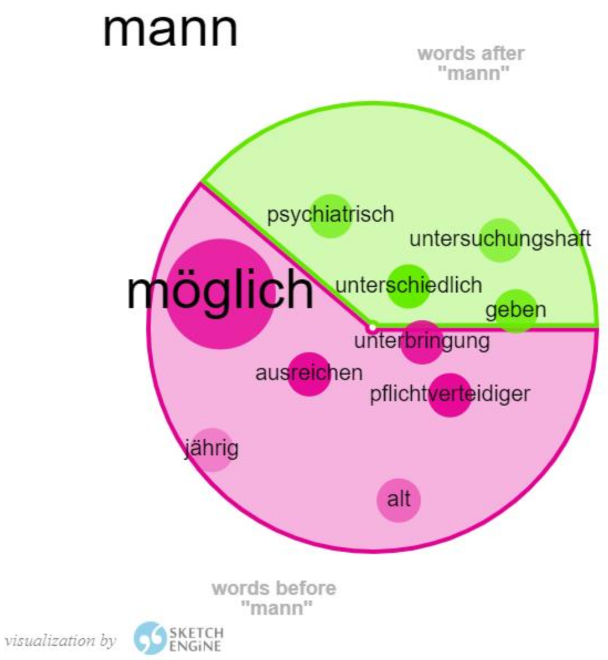


Figure 3.13: Bubble chart of the collocations of *Mann*.

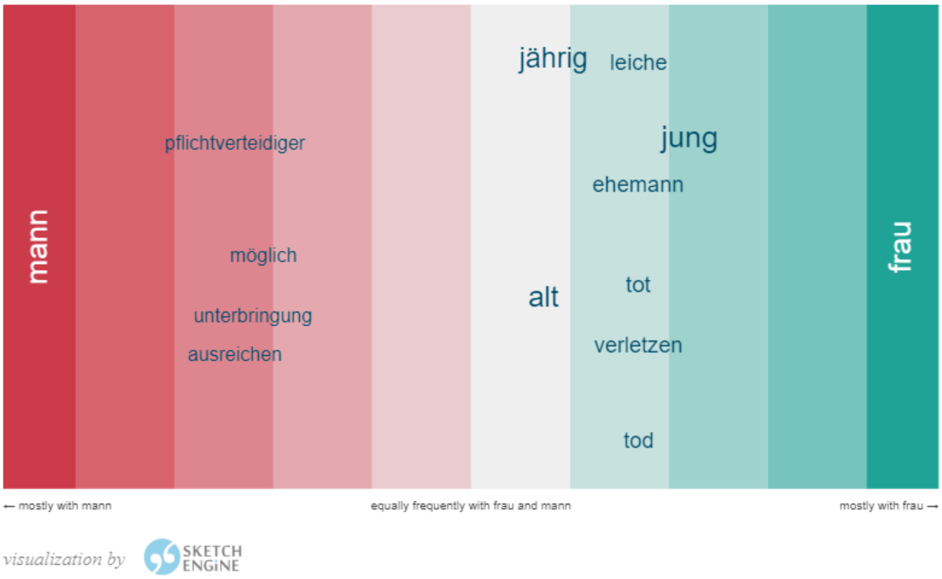


Figure 3.14: Word sketch difference representation of the type *Mann* and the type *Frau.*

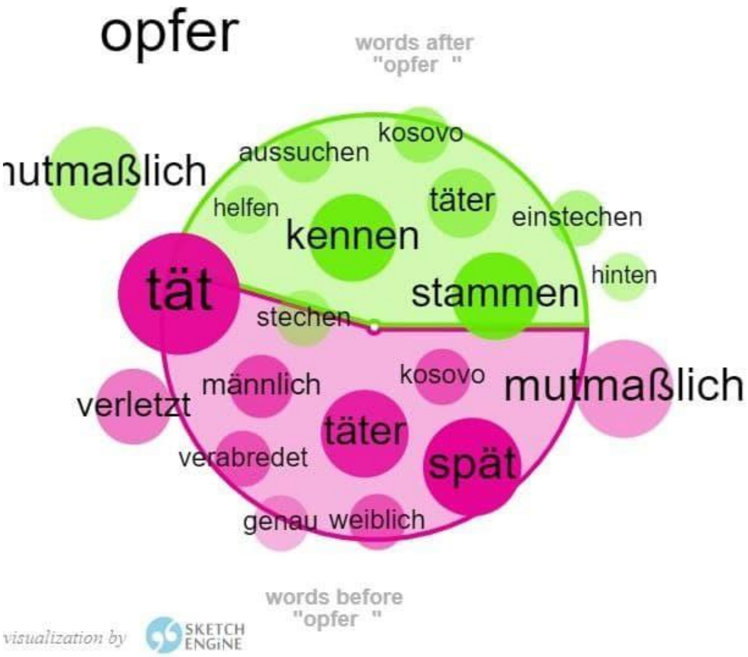


Figure 3.15: Bubble chart of the collocations of *Opfer*.

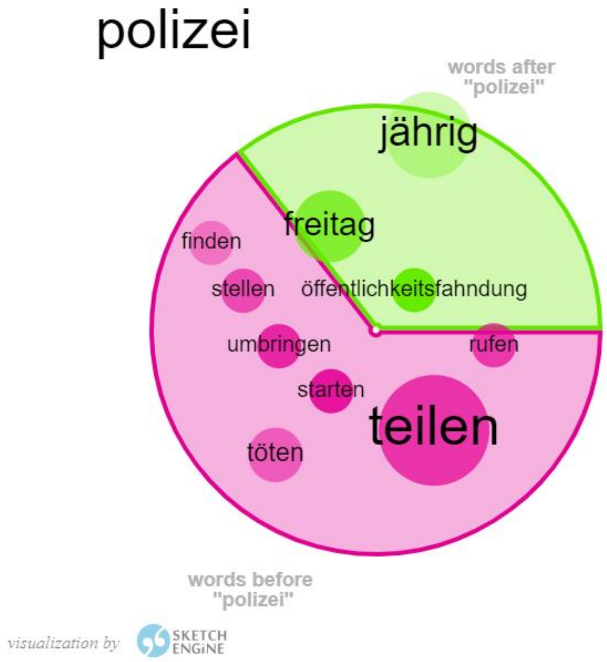


Figure 3.16: Bubble chart of the collocations of *Polizei*.

	ITALIAN CORPUS	**GERMAN CORPUS**
Number of femi(ni)cides	94	166
Number of newspapers	5	5
Number of articles	512	1,010
Number of types	99,746	131,536
Number of tokens	190,229	207,313

Table 3.1: A schematic representation of the numbers of femi(ni)cide cases, newspapers, articles, and words contained in both subcorpora.

STRUCTURAL INFORMATION	
Title	Title of the article
Subtitle	Subtitle of the article
Date	Date of publication
Source	Name of the newspaper
Section	Section of the newspaper
Content	Name of the victim
Murderer	Type of relationship the murderer had with the victim (tags: marito, ex marito, fidanzato, ex fidanzato, compagno, ex compagno, padre, fratello, nipote (di nonna), nipote (di zia), conoscente, vicino di casa, sconosciuto, unknown)
Suspect	What happened to the suspect (tags: condannato, accusato, suicidio, unknown)
Paragraph title	The title of the paragraph

Table 3.2: A schematic summary of the structural information added to the raw texts.

4. Topics extraction

As seen in chapter 2, latent Dirichlet allocation is an algorithm belonging to the family of topic modelling algorithms, based on the assumption that words appearing together in several documents belong to the same topic. Here we use this algorithm that, more precisely, works running through the words in each document, rating their probability and gathering them in a selected number of topics according to their cooccurrence in the same documents.

The documents used in this part of the study are already tokenised and function words, proper names and structural information are excluded from them. Although Brookes and McEnery (2019) suggests to avoid the exclusion of function words and punctuation to prevent some loss of meaning, this passage is included as the risk is that of finding useless topics formed by words belonging to closed classes only, as they usually have an higher frequency than words from open grammatical classes.

Another aspect of the text pre-processing to be noted here is the avoided step of lemmatisation. During texts pre-processing, lemmatisation is avoided as it can cause a loss of meaning. For what concerns metaphors, it is avoided because the same word but with different ending may be used in different types of metaphors. As already reported in chapter 1, experiment conducted by Deignan (2006) in extracting linguistic metaphors with two different wordforms of the same word (e.g., the singular form *flame* and the plural *flames*) demonstrates that the singular wordform is used in more positive ways, while the plural wordform usually has negative connotation. In fact, *flame* is used in sentence like '[…] Ross spots his old flame Rachel in the congregation.' (Deignan, 2006, p.116), where it is linked to the idea of love and lovers, while *flames* appears in metaphors as '…his future crashing in flames.' (Deignan, 2006, p. 117), where *flames* identifies a negative situation. For what concerns topics extraction, Brookes and McEnery (2019) suggests to avoid lemmatisation to avert a loss of meaning in topics, basing their assumption on the results of McEnery et al. (2015). In fact, in McEnery et al. (2015), a corpus analysis about the social media and press reaction to the death of an English soldier on behalf of two people with Muslim faith, the same word but with different ending – the singular form *Muslim* and its plural form *Muslims* – relate to two different discourses about Muslims. Applying topic modelling algorithm to a lemmatised corpus would bring to a richer set of topics, if considering the types, as each word would have only with its basic form, without any declination or conjugation. On the contrary, leaving each wordform in the corpus may cause to have more than one wordform of the same word in the same or in different topics, as they are read as different types by the algorithm. For some scopes, it would be better to use a lemmatised corpus, in order to avoid repetition of the same word in different wordforms. For this type of research, lemmatisation is not helpful as information conveyed by gender or plural markers may be meaningful for this type of study.

In addition to the pre-processing already shown, other steps are necessary to prepare texts to LDA algorithm application and are illustrated in the next section.

4.1 Preparing the texts

In order to prepare the text to apply the LDA algorithm, further separated notebooks are created, one for the Italian corpus and another for the German corpus, following the same steps for both.

Firstly, the pre-processed subcorpus is imported in the new notebook.

As a second step, Gensim (Řehůřek et al., 2010), pyLDAvis (Siviert and Shirley, 2014) and Matplotlib (Hunter, 2007) are imported. Gensim is an open-source Python library, which represents documents as semantic vectors and may be used to apply LDA. All the algorithms in Gensim are unsupervised, as LDA, and analyse the semantic structure of a corpus based on its statistical co-occurrences patterns. pyLDAvis – python library for interactive topic model visualisation – is a tool able to extract data from a LDA topic model and create an intertopic distance map, where topics are represented as circles on a two-dimensional plane, with the indication of the percentage of the corpus tokens contained in the topic. At the right of the two-dimensional plane, a bar chart is plotted, with a list of terms for the selected topic, with the indication of their frequency in the corpus and their estimated frequency in the topic. The position and the dimension of the circles on the intertopic distance map may be consulted to understand whether the model should be improved, increasing or reducing the number of topics. In fact, a model with circles in different sections of the plane and with a similar dimension is generally considered a good model, while a model with circles of different dimensions and placed in the same area of the two-dimensional plane is generally considered an improvable model. The last of the imported tool is Matplotlib, another tool for visualising data on plots, which in this study is used to create a visual representation of the words in a topic on separate bar plots.

The third step is the creation of a dictionary, containing the subcorpus, with Gensim tools. Gensim dictionary module creates a mapping between each token and its identification number. Therefore, a Gensim dictionary contains couples of tokens and identification numbers.

Then, the size of the dictionary is reduced filtering out words that occur in less than ten documents or in more than 80% of the documents. This step aims to exclude words appearing rarely or very commonly in the corpus. Initially, this value was set to less than 20 and more than 60%, but then changed because the topics extracted with this value are less coherent than those extracted from the dictionary with the selected values. Moreover, also from the pyLDAvis representation, the model with the dictionary with types occurring in more than ten and less than 80% of the documents seem to be more suitable because, unlikely the first selected values, there is no overlapping between two topics. Filtering out these words from the dictionary, the number of types in the Italian corpus is reduced to 1,326, while in the German corpus they are reduced to 1,468.

The fifth step is the creation of a bag of words from the reduced dictionary, as LDA algorithm does not keep track of the relation between words in a certain document but only of their frequency in a certain document. This step occurs after the creation of the dictionary as the bag of words corpus is a set of the integer numbers contained in the dictionary and their frequencies. In fact, it does not incorporate tokens orthographic form, but only their identification number.

The following step is the uploading of these two types of documents – dictionary and bag of words – in a folder, to avoid the creation of the two bags of words and dictionaries again, as it may be time consuming when working with a high number of texts.

It is now the time for the main step, the step concerning topics extraction. In order to create the LDA model, a few hyperparameters need to be specified, namely the number of topics, the bag of words and the dictionary to work with, the number of passes and a random state. The number of topics is empirically decided, as described below, and the bag of words and the dictionary included in the model are those created in the previous steps. The number of passes indicates how many times the algorithm runs over the corpus; a higher number of passes produces a more precise result. Then, the hyperparameter of the random state should be specified in form of a number, as it makes the process a non-random process, ensuring the reproducibility of the results. In this case, the random state is 67, and it is repeated any time the model is launched.

Initially, topics extraction is done only on the Italian corpus, with a low number of passes (150) and a number of topics from four to twelve. For each number of topics, pyLDAvis is plotted, to visually understand what number of topics could work better. From six to twelve, topics overlap on the intertopic distance map. The only two models with no topics overlapping on the two-dimensional plane are the model with four topics and the model with five topics. For this reason, all the other models are excluded, while these two are further developed. The algorithm is then launched again, both with the number of topics set to 4 and, on a separate notebook, with the number of topics set to 5. When relaunching the algorithm, the number of passes is increased up to 300, in order to have more accurate results.

As already reported at the beginning of the paragraph, these steps are done both with a dictionary containing words occurring in more than 20 documents and in less than 60% of the documents, and with a dictionary containing words occurring in more than ten documents and in less than 80% of the documents. The same procedure is applied to the German subcorpus, with the same number of passes, the same number of topics (four and five) and the same dictionary variants. Before applying this procedure, a brief test on this corpus is done, in order to check whether models with six, seven, eight, nine, ten, eleven or twelve topics should be suitable. LDA model with these numbers of topics are not suitable for the German subcorpus either, since they produce overlapping circles when plotting the pyLDAvis map. For this reason, the same LDA models applied to the Italian corpus, with the same condition, is run on the German corpus.

Evaluating the number of topics to set according to their position on the intertopic distance map, the models with four topics seem to be more suitable for both subcorpora. The Italian topics did not have any overlap, while a minimum overlap may be observed between two topics extracted from the German corpus. Also German topics extracted with the five topic models overlaps, but it is larger than those obtained with the four topics models. At the same way, the distance among the five topics is smaller than the distance among the four topics, which are situated in more distant areas of the plane. Nonetheless, a closer look to the five topic models' results is given. From a close reading analysis, these topics seem to be too broad, as each topic contains words from different semantic fields. Therefore, even if there is a little overlapping between two circles representing two German topics, the models with four topics seem to be more suitable.

After excluding the models with five topics, both because of the plotting on the intertopic distance map, both because of the close reading of the obtained topics, a close reading of the topics from the four topics models is made. The close reading of the results is done comparing the results obtained from the Italian corpus with a dictionary including words appearing in more than 20 documents and in less than 60% of the documents with those obtained from the Italian corpus with a dictionary including words appearing in more than ten documents and less than 80% of the documents, while results obtained from the German corpus with a dictionary including words appearing in more than 20 documents and less than 60% of the documents are compared with the topics obtained by the application of the algorithm to the German corpus with the other dictionary variant. These comparisons allowed to notice that the probability score of the terms in each topic is higher for the topics extracted from the corpora with the dictionary variant which excluded words appearing in less than ten documents and more than 80% of the documents, both for Italian and German results.

The model with the dictionary variant with words appearing in more than ten documents and in less than 80% of them seems to be preferable even when considering the intrinsic coherence of the topics. A function of Gensim package computes a measure – Intrinsic Umass measure – helpful to understand the degree of interpretability of a topic model. It is an average among the coherence of each topic, calculated as the sum of the logarithm of the ratio between the number of documents where two words cooccur and the number of documents containing the first of the two words. According to this value, topics extracted from corpora with words appearing in more than ten and less than 80% of all the documents of the corpus are better as their values are a little higher than those extracted from corpora with the other dictionary variant (-1.528 for the Italian model and -1.549 for the German model with the dictionary values set to no less than ten and no more than 80% of the documents, against -1.575 for the Italian model and -1.623 for the German model extracted from corpora with the other dictionary variant).

Both for the higher level of probability rate of the terms in each topics, both because of the higher level of coherence, but also because the topics obtained seem to be less dispersive from a semantic point of view – as topics extracted from corpora with the dictionary variant including words used in more than 20 and less than 60% of the documents comprehend many semantic fields –, the models with corpora including words appearing in more than ten documents and less than 80% of the documents are selected.

In order to provide a visual idea of the two selected models, their intertopic distance maps are reported in figure 4.1 and figure 4.2.

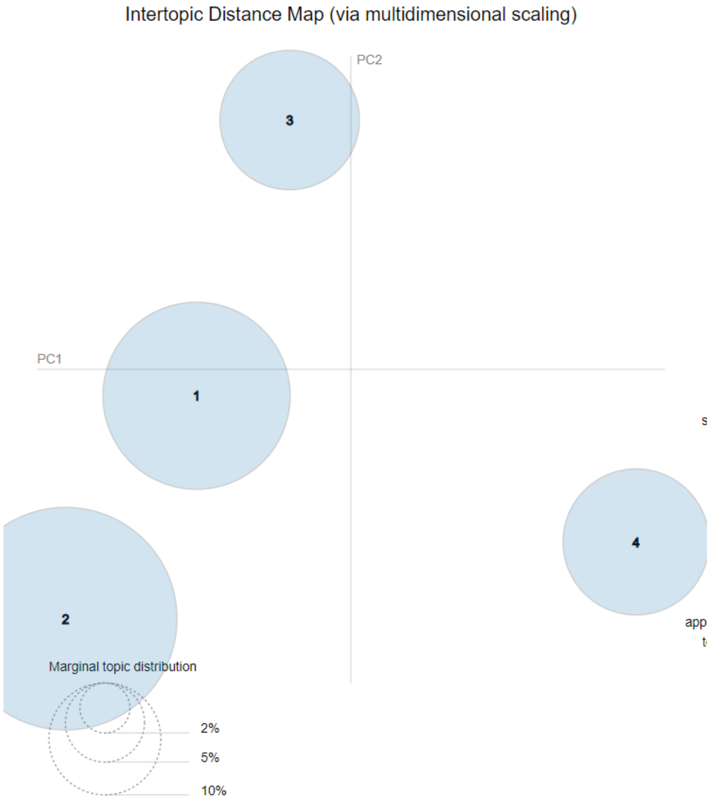


Figure 4.1: Intertopic Distance Map of the Italian model with four topics and the dictionary values set to no less than ten documents and no more than 80% of the documents.

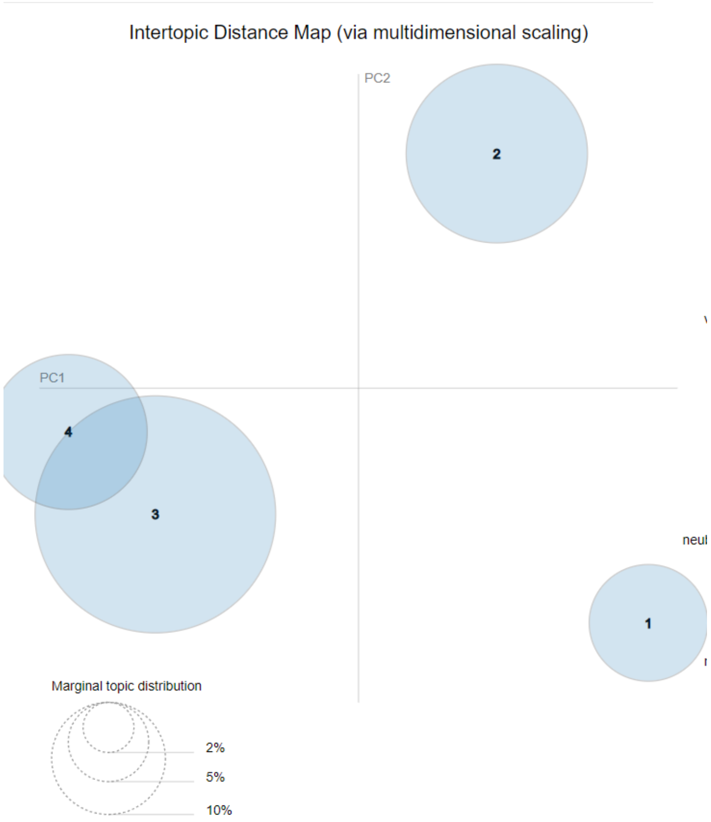


Figure 4.2: Intertopic Distance Map of the German model with four topics and the dictionary values set to no less than ten documents and no more than 80% of the documents.

4.2 Results – Italian subcorpus

This section illustrates the topics extracted from the Italian corpus with the four topics model and the dictionary set to no less than ten documents and more than 80% of the documents.

Reading the terms in the four extracted topics, they may be considered to describe two different moments, namely when the crime occurred and what happened after the crime. In fact, topic 1 mainly concerns the figure of the victim and topic 2 concerns people involved in the crime. On the other hand, topic 3 mainly concerns the finding of the corpse and topic 4 too is generally connected with the consequences of the crime. The four topics may be read in the bar plots reported below (in figure 4.3), which provide a list of the top ten associated words with an indication of their probability rate.

As it can be seen form the graphs, the probability rate of the terms is quite similar among all the four topics. The first top associated word of each topic seems to have the same probability rate, as well as the tenth associated word of each topic. In fact, the first top words of topic 1 and topic 3 have the same probability rate (0.020), which is similar to the probability rate of the first terms of topics 2 and 4 (0.021 for topic 2 and 0.016 for topic 4). Also the minimum probability rate is similar for all the topics, as the tenth words of the first and second topics have the same probability rate (0.007) so as the probability score of the tenth terms of topic 3 and topic 4 (0.008).

Taking into account their distribution over the subcorpus, topic 2 contains the majority of the corpus tokens (39.6%), then topic 1 contains a little less token (27.9%), and then topic 3 and topic 4 are smaller and contains quite the same percentage of the corpus tokens (15.5% for topic 3 and 17% for topic 4).

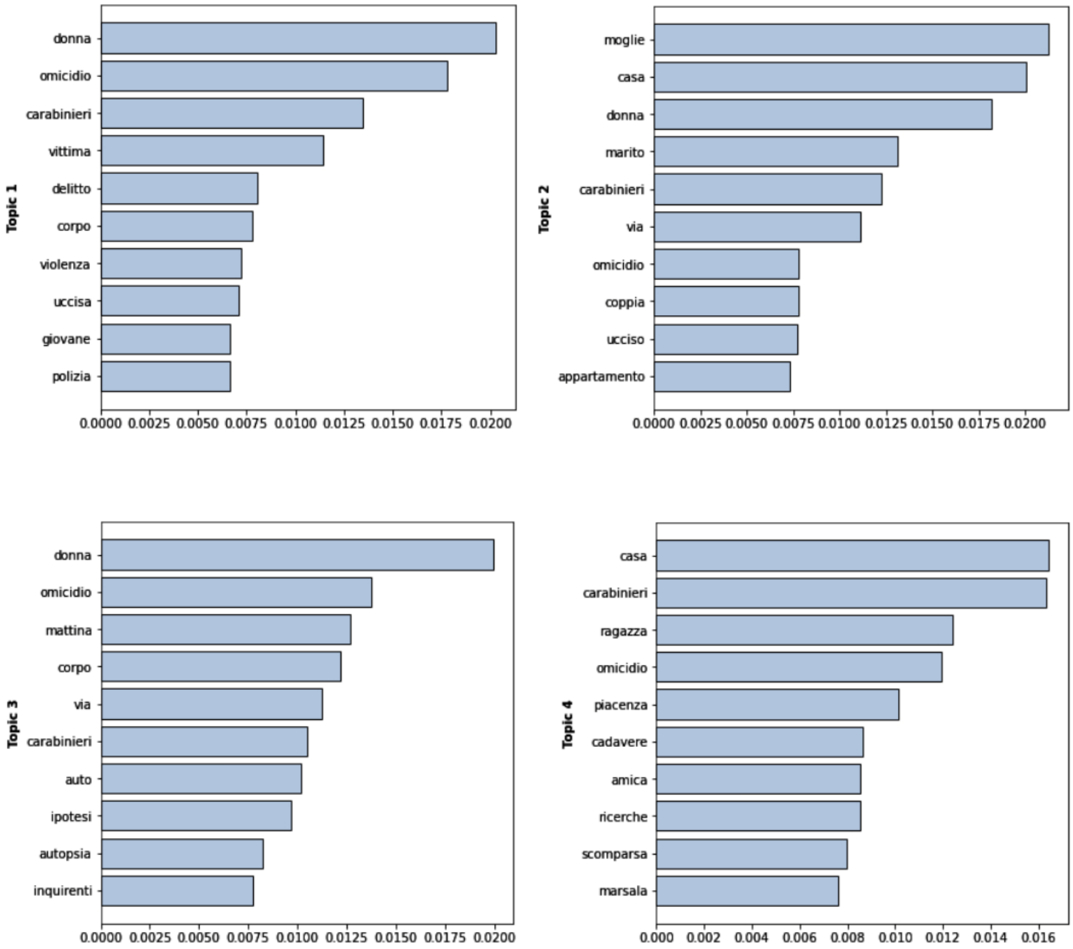


Figure 4.3: The bar plots contain the top ten associated words for each topic, sorted according to their probability rate.

4.3 Results – German subcorpus

After having described the result obtained from the Italian corpus, this section provides a description of those obtained from the LDA algorithm application to the German corpus.

The four extracted topics are mainly focused on the people involved in the crime and on juridical terms. Topic 1 mainly contains family terms while topic 4 contains the role names of people involved in the action of the crime. Topic 2 mainly concerns juridical entities involved after the crime and topic 3 is mainly linked to another consequence of the crime, namely death. The bar plots at the end of the paragraph report the top ten associated words for each topic with their probability level in the topic.

The terms with the highest probability rate are *Frau* from topic 3 (0.037) and *Polizei* from topic 4 (0.035), then there is the first term of topic 1, with a probability rate of 0.027, while the first term of topic 2 has the lowest probability rate (0.016). On the contrary, the probability level of the tenth word of each topic is similar, with the highest value of 0.011 (*Neubrandenburg* in topic 1) and the lowest value of 0.008 (*Tatort* in topic 4). The topic covering more tokens is topic 3 (44.7% of corpus tokens), which in the intertopic distance map overlaps with topic 4, which cover almost a half of the tokens of topic 3 (19.1% of tokens). Topic 1 covers the smallest percentage of tokens (10.8%) while topic 2 is the second topic for tokens covered (25.4% of the corpus tokens).

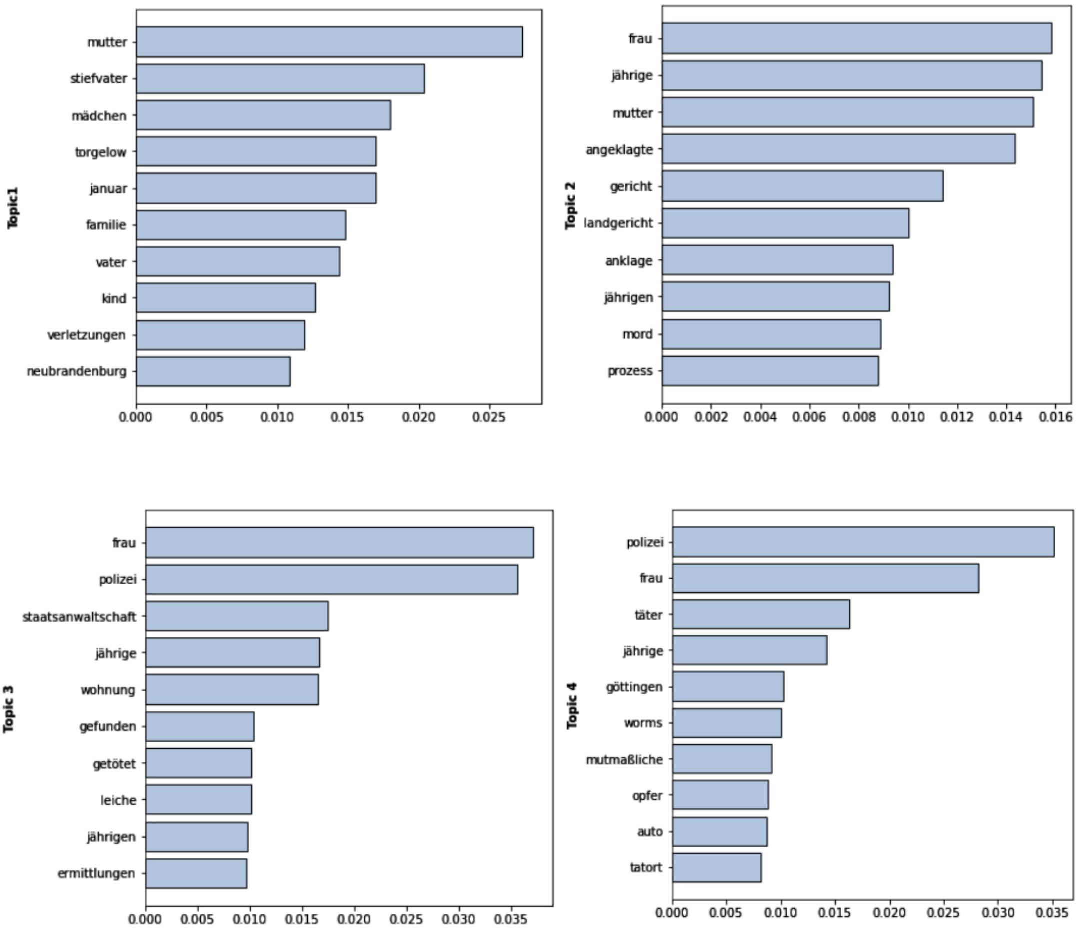


Figure 4.4: The bar plots contain the top ten associated words for each topic, sorted according to their probability rate.

4.4 Discussion

Even without an in-depth analysis it is possible to see that the topics extracted from the two subcorpora focuses on the event narrated in the newspapers; femi(ni)cide. All the terms forming the topics directly connects with the field of justice, with the idea of crime, death, the relationships among the people involved in the crime, or the place where the crime happened.

Two of the topics extracted from the German subcorpus centre on the people involved in the crime, both the generic role involved – as topic 4 with *Täter, Opfer, mutmaßliche* (‘perpetrator’, ‘victim’, ‘presumed’) – and nouns referring to family roles – *Mutter, Vater, Kind* (‘mother’, ‘father,’ ‘child’) in topic 1. Also topics extracted from the Italian corpus include terms connected with family roles – *moglie, marito, coppia* (‘wife’, ‘husband’, ‘couple’) in topic 2 – but they have no reference at all to the term *perpetrator.* Unlikely German topics, an entire topic seems to focus on the figure of the victim, namely topic 1, as it is composed by the terms *donna, vittima, uccisa* and *giovane* (‘woman’, ‘victim’, ‘killed’, ‘young’). Thanks to the lack of lemmatisation, it is possible to notice that the past participle of the verb *to kill* has a feminine ending. Therefore, it may be supposed that it is linked to *woman,* and that the victim of the criminal act is a woman.

Even though this divergence between the two set of topics, a similarity between the two extracted set of topics is the extraction of words belonging to the semantic field of right and justice. The German topic 2 contains words as *Gericht*, *Landgericht* and *Prozess* (‘court’, ‘regional court’, ‘trial’), while German topic 3 is formed by words as *Staatsanwaltschaft* and *Ermittlungen* (‘prosecution’, ‘inquiries’). Also Italian topic 3 contains words linked to the semantic field of right and justice, namely referring to the investigation process – *ipotesi, autopsia, inquirenti* (‘hypothesis’, ‘autopsy’, ‘investigators’). There are, furthermore, terms related both with people roles and with right and justice recurring in more than one topic, both for Italian and German topics. All four Italian topics contains the term *carabinieri* (‘Italian police’), also with a quite similar probability rate (0.013 in topic 1, 0.012 in topic 2, 0.011 in topic 3, 0.016 in topic 4). The equivalent German term *Polizei* does not appear so often, forming only two topics out of four. *Polizei* occurs only in the last two topics, which seem to group together words referring to a consequence of the crime, namely victim’s death – topic 3 – and to the roles involved in the crime – topic 4. Although this term may be found only in two topics out of four – unlike the Italian equivalent –, it has a higher level of probability than the Italian equivalent (0.036 for topic 3 and 0.035 in topic 4).

Other recurrent terms in these two sets of topics are those related to the figure of the woman. The German set of topics contains the term *Frau* in three topics (topic 2, topic 3 and topic 4) and it is the first top associated word for topic 2 and topic 3 (0.016 probability rate in topic 2 and 0.037 in topic 3) and the second top associated word in topic 4 (0.028 probability rate). Topic 1 does not include the word *Frau*, but it still has terms connected to the idea of woman, as *Mutter* and *Mädchen*, as first and third top associated word (0.027 probability rate for the first word and 0.018 for the third word). Thus, the Italian set of topics shows more terms linked to the idea of woman, at least one in each topic. This difference may also be a consequence of the inherent difference between the two languages. In this context, the Italian language is less ambiguous, because the German term *Frau* may be used to identify both a generic woman, both to identify someone's wife. This phenomenon is more unlikely to appear in the Italian language, where using *donna* to indicate a wife or girlfriend is suggested in very informal contexts only. Therefore, the German *Frau* should be intended both as wife and woman. These two meanings are usually split into two separate words in Italian**,** as it is possible to see in these topics too. The first term of topic 2 is, in fact, *moglie* (probability rate: 0.021) and its third term is *donna*, which is also the top associated term of topic 1 and topic 3 (with a probability rate of 0.020 in both). Topic 4 does not include any of these two words but still has terms connected to the idea of woman, such as *ragazza*, *amica* and *scomparsa*, which are declined with the feminine ending -*a* and, therefore, have a female referent. The word *ragazza* behaves similarly to the word *Frau*, as it may identify a generic young woman as well as someone's girlfriend.

In conclusion, it may be affirmed that the two sets of topics are quite similar as they both involve words connected to the crime – femi(ni)cide – reporting the perpetrator (*Täter*), the victim (*Opfer*, *vittima*), the kind of relationship involved (*Vater*, *marito, moglie*), the result of the crime (*Leiche, corpo, scomparsa*), and the legal procedure subsequent to the crime and the legal figures involved in it *(Polizei*, *Gericht*, *carabinieri*).

The obtained results are**,** to a certain extent, similar to those obtained by Busso et al. (2019). Of the ten topics extracted from a corpus composed of four Italian national newspaper articles (from *Corriere della Sera*, *La Stampa*, *Il Fatto Quotidiano* and *La Repubblica*) published from September 2016 to June 2017, seven of them are in line with the results obtained by this analysis. Only topic 1 – about the prevention of gender violence –, topic 6 – about prostitution – and topic 8 – about social media – concern themes which do not emerge from the topic analysis of the Italian and German subcorpora. The remaining topics relate to the semantic area of crime (how it happened, and the weapons used), identify places and people involved in the crime, or are connected to the semantic area of right and justice.

While it is possible to observe some similarities with Busso et al. (2019)'s results**,** the obtained results are completely different from those reported in Gius and Lalli (2014). In Gius and Lalli (2014), a corpus of 53 articles from three Italian national newspapers (*Corriere della Sera*, *La Repubblica*, *La Stampa*) is analysed with close reading methods. The three main images evocated during the narrations are 'Romantic Love', 'Loss of Control' and 'Other Contextual Elements' – articles that justify the crime through social condition, culture, mental illness, substance abuse. In the topics extraction on the Italian and German subcorpora, none of these themes come up to light. These differences may be justified mainly by two reasons, namely the different methodology used and the different years of publication of the articles composing the corpora. Gius and Lalli (2014) is a sociological study based only on close reading analysis by the researchers, no algorithm is involved, thus causing a more subjective analysis of the texts. The other reason is the different year of publication, as Gius and Lalli (2014) collected articles from 2012, seven years before the publication of the articles collected to build the Italian-German corpus. This variant is probably the cause of the differences emerged comparing the topics extracted by Busso et al. (2019) to those extracted from the Italian-German corpus. Busso et al. (2019) collected articles published from September 2016 to June 2017, while the Italian-German corpus is composed of articles published in the solar year 2019 relating to femi(ni)cide cases occurred in the same period. Basing the corpora on different years also involve taking into account different femi(ni)cide cases, articles narrating actions which took place in different ways. Moreover, as stated in Busso et al. (2019), their corpus reports articles from the 13th of September 2016 onwards, the day of the death of an Italian woman reported on newspapers as a victim of revenge porn. After this death, Italian journalists probably paid more attention on social media discourse when narrating femi(ni)cide cases and, as this is a well-known case in Italy, many articles have been published about it, underlying the social media part of the event. This may be the reason why there is a whole topic concerning social media in Busso et al. (2019) while there is no term at all concerning the social media area in the topics extracted from the Italian-German corpus.

In conclusion, the results of the topics analysis show an attention on behalf of the narration to report information about people involved in the crime, especially the type of relationship among them and the legal terms used to define them, information about the finding of the corpse, the juridical figures involved in the finding – police – and the procedure applied after the finding – court, process. These topics characterise both subcorpora, together with the feminine terms, which are recurrent in many topics, both with words identifying women, both with the grammatical gender of the ending of some words – e.g., *scomparsa, uccisa* (‘missing’ with the feminine ending -*a* and ‘killed with the feminine ending -*a*).

5. Metaphors extraction

Several studies (for a review, see Stefanowitsch, 2021) extract linguistic metaphors from corpora to analyse the conceptual metaphors beyond them. According to Lakoff and Johnson (1980/2003), verbal metaphors are linguistic representations of conceptual metaphors, which already are in our mind and shape our way of thinking and talking about anything.

This chapter aims to analyse whether metaphors are used to talk about femi(ni)cide phenomena and to which conceptual metaphors they may refer to. The methodology used to extract metaphors is similar to the methodology illustrated in Busso et al. (2019), already introduced in the chapter about conceptual metaphors (paragraph 1.5.3). Their analysis starts with the selection of keywords which may work as target domain in a possible metaphorical mapping, then they analyse the contexts in which they occur, find source domain terms and search for the contexts of use of these terms. After extracting the linguistic metaphors, they hierarchically group them and find the conceptual metaphor behind the linguistic expressions. After the detection of the conceptual metaphors used in the corpus, they consulted the metaphor repository MetaNet (Petruck, 2018), a project led by the International Computer Science Institute[15], to compare the obtained conceptual metaphors to the metaphors reported on the website. MetaNet is a repository of conceptual metaphors noticed analysing four languages – American English, Iranian Persian, Russian, and Mexican Spanish. Its available online version reports 685 conceptual metaphors and 576 frames and allows to search for conceptual metaphors in two different ways, namely directly looking for the conceptual metaphor or through its target or source domain – here reported as frame. For example, the conceptual metaphor PEOPLE ARE ANIMALS may be search by writing the entire conceptual metaphor on the webpage where the conceptual metaphors are recorded, and also through the source domain ANIMALS. Selecting the complete conceptual metaphor, PEOPLE ARE ANIMALS, a detailed page opens, where the metaphor type is reported – in this case, *composed/complex* –, and also references to studies which theorised the conceptual metaphor, related and entailed metaphors, the mappings characterising the conceptual metaphor and some linguistic examples – these fields are empty for this conceptual metaphor. In addition, source and target frame are specified, and also the language from which the metaphor is detected – PEOPLE is the target frame, ANIMALS the source frame and the language is American English. The other way to search for conceptual metaphors

[15] https://www.icsi.berkeley.edu/icsi/

in the MetaNet repository is by selecting a source or target frame. Selecting a frame, as ANIMALS, information about it are provided, such as the relevant lexical units of the domain, its relation with other frames and the metaphors in which the frame is used. For instance, some relevant lexical units of the frame are *animal* and *pig*, the frame ANIMALS is a subcase of ANIMATE ENTITY and BIRD is a subcase of ANIMALS, conceptual metaphors using this frame as target frame are not reported while conceptual metaphors using it as source frame are PEOPLE ARE ANIMALS, ADDRESSING CRIME IS CONTROLLING AN ANIMAL and UNCONTROLLED CRIME IS RAMPAGING ANIMAL.

The corpus used to study metaphors is the same built for LDA application, with the differentiations that names, punctuations, and function words are not excluded. This decision raises by the different nature of topics and metaphors. In fact, while for topics extraction only the single words are relevant and having a corpus with words belonging to functional grammatical classes and proper nouns may cause some topics to be meaningless for the purpose of the study, these items are indeed important to understand metaphors. Therefore, the corpus is the same for both tasks but has a slightly different pre-processing procedure.

Another differentiation between the two tasks is through what they are pursued. LDA algorithm is applied through the Python Genism library (Řehůřek et al., 2010), while the version of the subcorpora used to analyse metaphors are uploaded on Sketch Engine[16].

5.1 Metaphors extraction on the Italian subcorpus

The first step to extract metaphors from the corpus is to upload the needed corpus on Sketch Engine. During this step, the text is pre-processed with the Italian TreeTagger part-of-speech tagset, which tags the tokens according to their part of speech and some more specific grammatical categories – e.g., gender, number, tense and so on.

Our procedure is inspired by the methodology employed in Busso et al. (2019), here illustrated in figure 5.1. As a first step, a frequency list of the nouns is created. The first 30 nouns on this frequency list are then classified according to their semantic field. Nouns belonging to the groups concerning <crime>, <law enforcement>, <females>, <males>, <people>, become the seeds for metaphor extractions**.** These nouns are listed in table 5.1. The term femi(ni)cide – *femminicidio* – is added to these words.

[16] http://www.sketchengine.eu

The second step consists in analysing the contexts of use of these words, in all their possible wordforms and taking note of every possible metaphor or hint of metaphor or important word. For example, from the analysis of the wordforms of *uomo* the image of hunting appears to be a possible metaphor – e.g., *caccia all'uomo* ('man hunting').

The third step is a more fine-grained search through the extraction of some exact collocations of the starting seeds. Namely, these collocations are sentences where the seeds function as subject of the copular verb, or the alternation of an adjective and a seed and vice versa. The queries used for this step, here reported for the term *uomo*, are [lemma = "uomo" & tag = "NOUN"][tag = "ADJ"], [tag = "ADJ"][lemma = "uomo" & tag = "NOUN"], [lemma = "uomo" & tag = "NOUN"][lemma = "essere" & tag = "VER:fin"].

Both the second and the third steps allows to retrieve more metaphorical concepts which are looked for in the fourth steps. In fact, in the fourth step, all the metaphors are grouped according to their source domain and, after that, the collocations of the source domain terms are extracted in order to look for new target domains. For example, collocations of the wordforms of the verb *cacciare* ('to hunt') are searched for and analysed to detect whether it is metaphorically used with different target domain than *uomo* and how frequent it is metaphorically used.

At the end of these steps, the linguistics metaphors are grouped together based on their source domain and we tried to understand to which conceptual metaphor could refer to. Metaphors which are highly likely to be used in different contexts other than newspaper articles, such as linguistics metaphors referring to LIFE IS A COMPLEX STRUCTURE, are excluded from the results as they are proper to almost every discourse about life and not a characteristic feature of femi(ni)cide discourse or newspaper journalism. Conceptual metaphors concerning the theme of gender violence are then compared to the conceptual metaphors recorded on MetaNet (Petruck, 2018).

MetaNet is first checked based on the source frame – the semantic domain from which the metaphorical terms are taken, e.g., ANIMALS in the conceptual metaphor PEOPLE ARE ANIMALS – in order to see to which conceptual metaphors the collected source domain can belong to.

The same method illustrated in step three is conducted for the proper nouns too, in order to see whether victims and perpetrators are included in some metaphorical context. This step gave no results, and it is therefore plausible that no metaphor is used in combinations with the names of the characters of the crime.

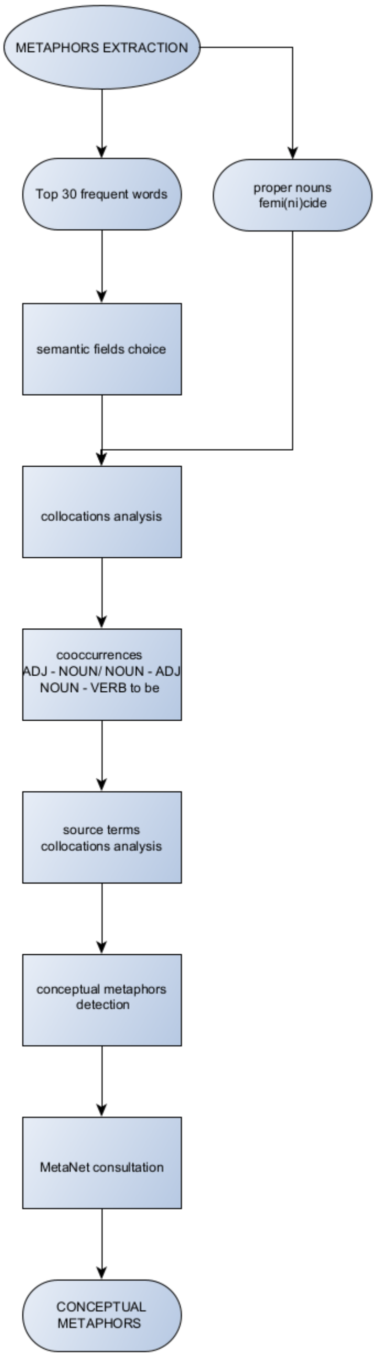


Figure 5.1: The procedure followed to extract metaphors.

ITALIAN SEMANTIC FIELDS				
CRIME	LAW ENFORCEMENT	FEMALES	MALES	PEOPLE
Omicidio[17] Vittima Corpo Vita Crimine Cadavere	Polizia Carabinieri Indagine	Donna Moglie Figlia Ragazza	Uomo Marito Amico Figlio	Coppia Famiglia Persona

Table 5.1: Seeds for metaphors extraction.

5.2 Results – Italian subcorpus

The starting point of the method described in these pages is the grouping of words according to their semantic field. Of the five selected semantic fields – <crime>, <law enforcement>, <females>, <males>, <people> – the one with more metaphorical expressions is the field labelled as <males>. Most of the times, indeed, the male protagonist of the event is often described with terms belonging to the animal domain. 30 occurrences out of 33 of the type *caccia* (‘hunting’) are metaphorically used referring to the terms *uomo, marito/ex marito*, or, more generally, to the person who committed the crime. The expression *caccia all’uomo* (‘man hunting’) is not the only example in which the source domain of ANIMALS shapes metaphors concerning the criminal. The perpetrator is also presented as a wild beast – *belva* – as this word is metaphorically used six times out of eight to talk about the killer, while *bestia* is used only once in the whole corpus and describes the killer too. In other metaphors, the killer remains an animal but becomes a prey – *preda* – of some mental disease or emotional state five times out of eight, and he also is a predator once out of once. Also the violence employed is described as something inhuman, which is proper of wild animals and not of men. When men beat women, they beat them in a savage way – *selvaggiamente* – two times out of two, their violence is presented as brutal – *brutale/brutalmente* – 17 times out of the 18 times this word occurs in the corpus. Moreover, the killers are also introduced through their ferocity. 10 times out of 10 the term ferocity – *ferocia* – is used as a property of the killer or to describe the event – *la ferocia degli eventi*.

[17] Homicide, victim, life, crime, corpse, police, investigation, woman, wife, daughter, girl, man, husband, male friend, son, couple, family, person.

Metaphors formed with the linguistic material shown so far may be verbal representation of the following conceptual metaphors: PEOPLE ARE ANIMALS, UNCONTROLLED CRIME IS A RAMPAGING ANIMAL, ADDRESSING CRIME IS CONTROLLING ANIMAL.

These three conceptual metaphors are recorded in the MetaNet repository. The second and the third conceptual metaphors are complex metaphors and are an entailment of CRIME IS AN ANIMAL. All the three conceptual metaphors bound the idea of men to the idea of animals, working in our mind almost as a claim that who commits this crime is not human anymore.

Linguistic metaphors bounded to PEOPLE ARE ANIMALS are: *abbiamo conosciuto le belve, compagni predatori* (‘we knew the beasts’, ‘predator partners’).

Linguistic expressions of the conceptual metaphor UNCONTROLLED CRIME IS A RAMPAGING ANIMAL are: *lui mi picchiava selvaggiamente, violenze brutali* (‘he beated me savagely’, ‘brutal violence’).

Examples of linguistic metaphors of ADDRESSING CRIME IS CONTROLLING ANIMAL are those concerning hunting, such as *caccia all’uomo* and *far perdere le tracce* (‘man hunt’, ‘to lose tracks’).

Moreover, also linguistic metaphors conveying conceptual metaphors do not recorded in MetaNet are found. The linguistic metaphors of the killer committing the crime as a prey of some mental diseases or strong emotional states may be partially identified as part of the conceptual metaphor PEOPLE ARE ANIMALS. It seems a plausible metaphor as preys are usually animals, but it does not justify how emotions or diseases are conceived in this type of linguistic metaphors. The man is the patient of an action executed by another entity, he simply undergoes the action of the emotional state or the mental disease of taking him as in a hunting trip. Therefore, these linguistic metaphors should be the representation of a conceptual metaphor which gives more weight to the agents of the hunting action, which in this case are emotions and mental diseases. A possible conceptual metaphor may be MENTAL STATES ARE ANIMALS or MENTAL STATES ARE HUNTERS. Between these two conceptual metaphors, the one which seems to be more suitable is MENTAL STATES ARE HUNTERS, because a conceptual metaphor like the former - MENTAL STATES ARE ANIMALS – does not seem to consider the unbalanced relation between the hunters and the preys, but rather put them on the same level in the action, as they were both agents. The conceptual metaphor MENTAL STATES ARE HUNTERS makes explicit the agent of the hunting action – mental states – and implicitly makes men preys, an additional important information not contained in the first conceptual metaphor proposed.

Other linguistic metaphors linked to the figure of the killer and the violence beyond the crime are metaphors concerning the areas of monstruous and supernatural. The killers may be both presented as animals or as monsters. Even if with fewer occurrences than terms belonging to the semantic field of animal, words connected to irrational figures are used as metaphors to describe the killers and the violence of the crime. With four occurrences out of four, the noun monster – *mostro* – is used as a reference to the murderer. Out of a total of five occurrences in the whole Italian corpus, the adjective diabolic – *diabolico* – is used four times referring to the couple who committed a murder and once preceding the word deceit – *inganno*, while the four occurrences of the type ghost – *fantasma* – all refer to the murderer's escape. The remaining words referring to a supernatural world and metaphorically used to describe the criminal event and the people involved are hell – *inferno* – and giant – *gigante*. *Inferno* only has two occurrences in the whole subcorpus and both of the times is used metaphorically, once to indicate the hard life of the victim, and the other time to talk about the beginning of the criminal event – *si scatena l'inferno* ('hell begins'). The image of the giant is used to talk about the figure of the killer and is used in the phrases *gigante buono* and *gigante amico* ('good giant', 'friend giant'). All the occurrences of the type *giant* are used to identify the murderer, four occurrences out of six are of the phrase *gigante buono*, one of the phrase *gigante amico* and in the other one the noun has no modifier. As the conceptual metaphors already illustrated, which bound the idea of men to the idea of animals, also these types of metaphors dehumanise the murderer but, differently to the frame of animals, no conceptual metaphors at all concerning supernatural or monstruous is reported in MetaNet. Since all the metaphorical uses of the terms concerning evil irrational entities relate to the figure of the murderer, all the linguistic metaphors with this source domain may be embedded in a conceptual metaphor like MURDERERS ARE INHUMAN ENTITIES.

Partially bounded to the idea of a crime committed by non-human entities are the linguistic metaphors which report emotional states, violence or the event of femi(ni)cide itself as playing the role of the perpetrator of the crime. In fact, 25 occurrences out of 35 of the phrase *victim of* are completed by the terms femi(ni)cide, violence, anger, madness and so on. Only ten occurrences are completed by the name of the murderer or the type of relation of the murderer with the victim. These linguistic metaphors may be a different way of conveying the conceptual metaphor MURDERERS ARE INHUMAN ENTITIES. When the agent of a certain action is perceived as less human, it is then also possible to use the mental states of the agent, or the event caused by the agent as the agent itself. Therefore, the victims are described as victims of femi(ni)cide or victim of violence, even if the femi(ni)cide is the phenomenon occurring when a woman is killed because she is a woman, and the violence is caused by the actions of the killers.

According to the source domain of the extracted linguistic metaphors, other groups of conceptual metaphors are found.

The source frame of war is present when journalists talk about femi(ni)cide, about arguments and also about relationships. People fight to solve the phenomenon of femi(ni)cide, which is described as a battle (3 times out of 3). Also relationships and arguments cooccur with words belonging to the semantic field of war. The adjective *conflittuale* ('conflicting') appears only three times in the subcorpus, and all these three times are used to modify the noun *relationship*, which is also introduced as full of riots – *tumultuosa*. For what concerns the argument sphere, arguments explode – *scoppiare* – 20 times out of its 30 occurrences in the corpus. Linguistic metaphors already reported may be the linguistic expression of the conceptual metaphor ARGUMENT IS WAR. In fact, arguments explode like bombs, weapon used during conflicts, and relationships with recurrent arguments are described as conflicting and with may riots.

Also a set of vegetable metaphors is extracted from the corpus. Five times out of six, the verb *troncare* occurs in the sense of closing a relationship. With the similar meaning of ending something, *spezzare* describes the death of the victims. On the contrary, the verb *maturare* ('ripen') occurs 8 times out of 9 describing the process which led to commit homicide. The first two linguistic metaphors may be the lexical representation of the conceptual metaphors PEOPLE ARE PLANTS. The latter linguistic metaphor has no conceptual counterpart in MetaNet and a possible conceptual metaphor may be CRIMES ARE PLANTS. This conceptual metaphor may be also justified by the linguistic metaphors *architettare un delitto* – ('to plan a crime'). The Italian verb *architettare,* literally used in the semantic field of house building as *to design* has three occurrences in the corpus and all three of them are used referring to the crime. However, in this case a conceptual metaphor more like CRIMES ARE COMPLEX STRUCTURES seems to be more suitable. In both cases, the crime is perceived as something complex, which needs to grow up or to be planned before it occurs. However, in the former conceptual metaphor it has more agentivity than in the second, where it is the patient of the planning action.

Belonging to similar source frames are the linguistic metaphors which describe relationships, arguments, and people as inanimate objects. The adjective fragile – *fragile* – occurs six times in the corpus, four times it is used to modify nouns referring to the victims and twice as a modifier with nouns referring to the murderer. However, it modifies the noun bringing different readings. In fact, when it is used to talk about women, this adjective conveys the meaning of someone who is weak and needs protection while, when it is used referring to the men, it is used to describe the man as someone who is not able to act and think by himself because there is a more powerful entity – a mental disease – acting for him. Relationships and arguments are described as something to be put back together – *ricomporre* (metaphorically used three times out of its three occurrences in the corpus) - and that can deteriorate – *incrinare* (metaphorically used twice out of its two occurrences in the corpus). In these linguistic metaphors, relationships are objects with a stable size which can be broken by the fights and need to be repaired. Relationships are considered as something to be recomposed, to be fixed after a fight or a tough period. Therefore, the possible conceptual metaphor beyond the verbal metaphors *relazione incrinata* and *ricomporre la crisi* is RELATIONSHIPS ARE OBJECTS, a metaphor also reported in MetaNet.

Of the metaphors reported in this paragraph, only two conceptual metaphors are extracted from all the newspapers in analysis, namely ADDRESSING CRIME IS CONTROLLING ANIMAL and MURDERERS ARE INHUMANE ENTITIES, with the linguistic expressions *caccia all'uomo ('man hunting')* and *vittima di* followed by terms referring to the criminal event, the mental states of the perpetrator or *violence*. Another commonly used metaphor is UNCONTROLLED CRIME IS RAMPAGING ANIMAL, contained in all the newspapers in analysis except *La Stampa*. Conceptual metaphors concerning people (PEOPLE ARE PLANTS and PEOPLE ARE ANIMALS) are both reported in *Il Giornale*, but PEOPLE ARE PLANTS also occurs in *La Repubblica*, while PEOPLE ARE ANIMALS occurs in *Il Fatto Quotidiano* and *La Stampa*. A further conceptual metaphor appearing only in *Il Giornale* and *La Repubblica* is RELATIONSHIPS ARE OBJECTS. *Il Resto del Carlino* has only one conceptual metaphor in addition to the two most common, namely CRIME ARE PLANTS, which is shared with *Il Fatto Quotidiano*, *La Repubblica* and *La Stampa*. The remaining conceptual metaphors seem to be less common and are noticed only in two newspapers – ARGUMENT IS WAR in *Il Fatto Quotidiano* and in *La Stampa*, MENTAL STATES ARE HUNTERS in *La Repubblica* and in *La Stampa* – or one newspaper – CRIMES ARE COMPLEX STRUCTURES is only reported in *il Giornale*. The newspaper presenting the higher number of conceptual metaphors (8) is *La Repubblica*, a left-wing newspaper, which has all conceptual metaphors except ARGUMENT IS WAR and CRIMES ARE COMPLEX STRUCTURES, while the newspaper with less conceptual metaphors is *Il Resto del Carlino*, a right-wing newspaper, which only reports the most common conceptual metaphors, namely ADDRESSING CRIME IS CONTROLLING ANIMAL, MURDERERS ARE INHUMAN ENTITIES – reported in all the newspapers – and UNCONTROLLED CRIME IS A RAMPAGING ANIMAL, CRIME ARE PLANTS – reported in four newspapers out of five. The other newspapers present a similar number of conceptual metaphors, seven conceptual metaphors for *Il Giornale*, and five conceptual metaphors for *La Stampa* and *Il Fatto Quotidiano*.

The following two table**s** provide a list of the conceptual metaphors extracted from the Italian subcorpus, dividing the metaphors reported in MetaNet to the other conceptual metaphors, and a visual schema clarifying which conceptual metaphors appear in which newspaper.

	CONCEPTUAL METAPHORS
in MetaNet	PEOPLE ARE ANIMALS (*abbiamo conosciuto le belve*)
	UNCONTROLLED CRIME IS A RAMPAGING ANIMAL (*lui mi picchiava selvaggiamente*)
	ADDRESSING CRIME IS CONTROLLING ANIMAL (*caccia all'uomo)*
	RELATRIONSHIPS ARE OBJECTS (*relazione incrinata*)
	PEOPLE ARE PLANTS (*una vita spezzata*)
	ARGUMENT IS WAR (*una lite violenta scoppiata in piena notte*)
not in MetaNet	MURDERERS ARE INHUMAN ENTITIES (*abbiamo accolto un mostro in casa*)
	CRIMES ARE PLANTS (*omicidio maturato in un contesto familiare*)
	CRIMES ARE COMPLEX STRUCTURES (*ha architettato l'omicidio*)
	MENTAL STATES ARE HUNTERS (*uomo fosse preda di una qualche forma di depressione*)

Table 5.2: a list of the conceptual metaphors extracted from the Italian subcorpus. A linguistic metaphor is reported as example for each conceptual metaphor.

	La Repubblica	La Stampa	Il Giornale	Il Fatto	Il Resto del Carlino
PEOPLE ARE ANIMALS	X		X		
UNCONTROLLED CRIME IS A RAMPAGING ANIMAL	X		X	X	X
ADDRESSING CRIME IS CONTROLLING ANIMAL	X	X	X	X	X
RELATIONSHIPS ARE OBJECTS	X		X		
PEOPLE ARE PLANTS	X		X		
ARGUMENT IS WAR		X		X	
MURDERERS ARE INHUMAN ENTITIES	X	X	X	X	X
CRIMES ARE PLANTS	X	X		X	X
CRIMES ARE COMPLEX STRUCTURES			X		
MENTAL STATES ARE HUNTERS	X	X			

Table 5.3: Distribution of conceptual metaphors among the five Italian newspapers.

5.3 Metaphors extraction on the German subcorpus

Metaphors extraction on the German subcorpus follows the method already used for the Italian subcorpus, with a few language specific adjustments.

As for the Italian corpus, extraction of naturally occurring linguistic metaphors is pursued through Sketch Engine. The collected corpus is uploaded on this software and processed with version 4.2 of German RFTagger (Schmid and Laws, 2008), which tags the tokens according to their part of speech and some more specific grammatical categories – e.g., case, gender, number, tense and so on. When the corpus is ready, some words, which will work as the seed of the research, are extracted. The same parameters of the Italian corpus are followed, namely the most 30 common names belonging to specific semantic area are extracted and grouped together according to the semantic field. The chosen words belong to the semantic area of <females>, <males>, <crime> and <law enforcement> **a**nd are listed in table 5.4. Similarly to the methodology used for the Italian corpus, a further word is added to functions as seed, *Frauenmord*, but there has no occurrence in the subcorpus. The first step for extracting metaphors consist in analysing the concordance of the seeds, observing the contexts in which they are used and the possible metaphors. Then, they are analysed in more precise contexts, namely when preceded by an adjective – as the regular structure of German noun phrases is usually *adjective + noun* – and when they function as a subject of the verb *to be*. As in German the verb may be found both in second position and at the end of the clause, a different search than the search conducted for the Italian subcorpus is carried out. Instead of extracting a query like *subject + finite form of the verb*, concordances of the verb *to be – sein* – are extracted and the search is refined looking for the occurrences of the selected nouns in a span size of the dimension of ten tokens at the left and ten tokens at the right of each occurrence of the finite form of the verb. Cooccurrences where nouns are subject of a copular sentence are extracted also for the proper noun of victims and perpetrators, when available. As for the Italian corpus, proper nouns seem to be not included in metaphorical contexts.

All the linguistic metaphors are then gathered according to the source frame. Then, concordances of words belonging to the source frames are extracted, in order to quantify the metaphorical use of certain words and to detect further linguistic metaphors.

During the whole process, linguistic metaphors not bounded to the initial seed come out, so the different steps are repeated to extract also these new linguistic metaphors.

At the end of all the steps, the obtained metaphors are grouped together according to their source and target domains and the possible conceptual metaphors behind them are compared to the conceptual metaphors recorded on MetaNet.

GERMAN SEMANTIC FIELDS			
CRIME	LAW ENFORCEMENT	FEMALES	MALES
Tat[18] Täter Leiche Opfer	Polizei Staatsanwaltschaft Ermittler Ermittlung	Frau Freundin Mutter	Mann

Table 5.4: Seeds for metaphor extraction.

[18] Criminal act, perpetrator, corpse, victim, police, public prosecution, investigator, investigation, woman, female friend, mother, man.

5.4 Results – German subcorpus

As already reported in the previous paragraph, words working as seeds for this research belong to four different semantic area. The fist semantic area, <females>, is formed by words which do not seem to be used in metaphorical expressions, while words belonging to the semantic areas of <males> and <law enforcement> appears to be in common metaphorical contexts. Of the fourth semantic area, only one word is used in a metaphorical context, *victim*. Moreover, other two metaphors, not related to the semantic areas of the chosen seeds, merged from the analysis.

Words belonging to the semantic area of <males> and <law enforcement> join together in metaphors based on the source domain of animals. In fact, linguistic metaphors as *die Polizei jagt dort einen Mann* ('police are chasing a man') is a hunting metaphor formed by the agent police and the patient man. The verb *jagen* ('to hunt') occurs seven times in the corpus, of which, six times is used to describe the action of the police (or public prosecution) of looking for the accused person. These linguistic metaphors are not the only metaphors evoking hunting images. Another recurrent image is *Visier* ('gunsight'), namely the investigator which has the suspected on their sight. This linguistic metaphor may remind of war images, but it mainly seems to be a reference to a hunting scene, when the hunters stay still staring at the preys through the gunsight waiting for the perfect moment to shoot them. The source domain of ANIMALS also recurs in other linguistic metaphors, with the adjective *brutal* ('brutal') and *bestialisch* ('bestial'). *Brutal* occurs 35 times in the whole corpus, and all its occurrences modify something related to the crime, namely who committed the crime or the way the crime took place. Similarly, the adjective *bestialisch* occurs 16 times in the corpus, each occurrence works as a modifier of crime referents or to describe the victim states after the crime took place. The conceptual metaphors merging from these linguistic expressions may be PEOPLE ARE ANIMALS, ADDRESSING CRIME IS CONTROLLING ANIMALS and UNCONTROLLED CRIME IS A RAMPAGING ANIMAL. The two conceptual metaphors are mainly justified by the use of hunting terms to refer to human actions. Law enforcement looking for the possible killer becomes a hunting trip as the criminal is an animal committing acts of violence. Therefore, if the guilty part is an animal, then solving the case and arresting the criminal means to control the animal.

Another metaphor which tends to de-humanise the figure of the criminal is a linguistic metaphor which is a unicum in this subcorpus. It says *the world of the other criminals* as criminals are not from this world and there are criminals which are more human than other criminals. The idea of two worlds, one world for the human beings and the other for the criminals, finds no correspondence in the MetaNet database. The only source frame linked to the idea of crime are ANIMALS and DISEASE, but none of these two source frames may be connected to the thought of a different world for people committing crimes. A possible conceptual metaphor, already met in the Italian subcorpus, is MURDERERS ARE INHUMAN ENTITIES. It seems to be suitable as, in the common imaginary, entities from other worlds are aliens or monsters, both are nonhuman entities characterised by negative and frightening connotations.

A conceptual metaphor which is instead in MetaNet is RELATIONSHIPS ARE OBJECTS. This metaphor has only one evidence in the corpus, namely *on-off Beziehung* ('on-off relationship'), to refer to an unstable relationship. Even if there is only one linguistic expression of this conceptual metaphor, it seems important to be reported as it confirms the theory that metaphors are more used to talk about abstract concepts than more concrete ideas (Lakoff and Johnson, 1980/2003). The on-off option instantly reminds us to the tangible light switch, or to every other switch with two options, the option to turn it on or turn it off with one touch only. The on-off image rapidly gives the idea of something so unstable that no effort is needed to break the relationship and then to begin it again.

The last two metaphors detected from the German subcorpus are not reported on MetaNet. The first metaphor concerns social media, and Facebook in particular. In this subcorpus, Facebook is seen as a place of grieving. People mourns on Facebook and share their emotional upset as it were a personal diary or a close friend. For example, the linguistic metaphors *ihr Mann trauert auf Facebook* ('her man mourns on Facebook') means that he shared a post where he probably wrote something about his loss, but the choice of the verb *trauern*, which is intimate and deeply connected to the theme of grieving and suffering for the loss of a loved person, contributes to make the social media an intimate and safe space. There is no conceptual metaphor about social media recorded on MetaNet. A probable conceptual metaphor connected with this way of talking about the social network may be SOCIAL MEDIA ARE PLACES FOR MOURNING.

The last linguistic metaphors is *victims of* completed without a reference to the person who actually killed the victim, but to the fact itself. They are victim of a crime of power and violence – *Gewaltstraftat* – or of an act of murder – *Bluttat* – but not of the person who killed them. The conceptual metaphor MURDERERS ARE INHUMAN ENTITIES seems to be at the basis of these linguistic metaphors too. Therefore, if there is no precise human entity committing the criminal act, the fault is not of a precise human entity, but of the event itself. The human being who acts as the agent of the crime becomes a pale figure and the action itself is at the centre of the discourse.

As for the Italian subcorpus, also for the German subcorpus newspapers-based research is done, in order to see which newspapers contain which conceptual metaphors. Similarly to the Italian subcorpus, the solely conceptual metaphor extracted from all five newspapers is MURDERERS ARE INHUMAN ENTITIES, with the expression *Opfer* + genitive phrase. The other two most common conceptual metaphors are UNCONTROLLED CRIME IS RAMPAGING ANIMAL and ADDRESSING CRIME IS CONTROLLING ANIMAL, which are reported in all the newspapers except for *Die Zeit.* The other three remaining conceptual metaphors are extracted for one newspaper each, namely SOCIAL MEDIA ARE PLACES FOR MOURNING occurs in *Bild*, RELATIONSHIPS ARE OBJECT occurs in *Süddeutsche Zeitung*, PEOPLE ARE ANIMALS appears in *Die Welt*. All the newspapers present almost the same conceptual metaphors. The newspaper with less conceptual metaphor is Die Zeit, which also present a really low number of articles about femi(ni)cide (6) published in 2019.

In conclusion, a modest number of linguistic metaphors and, therefore, of conceptual metaphors merged from the analysis. The table**S** below list the conceptual metaphors extracted from this corpus, which mainly focus on the idea of criminal as non-human beings, and which conceptual metaphor occurs in which newspaper.

	CONCEPTUAL METAPHORS
in MetaNet	PEOPLE ARE ANIMALS (*‘brutal’ nannte die Anwältin […] das Verbrechen*)
	UNCONTROLLED CRIME IS A RAMPAGING ANIMAL (*bestialischen Messermord*)
	ADDRESSING CRIME IS CONTROLLING ANIMAL (*die Polizei jagt dort einen Mann*)
	RELATRIONSHIPS ARE OBJECTS (*on-off Beziehung*)
not in MetaNet	MURDERERS ARE INHUMAN ENTITIES (*aus der Welt der übrigen Täter herausfällt*)
	SOCIAL MEDIA ARE PLACES FOR MOURNING (*ihr Mann trauert auf Facebook*)

Table 5.5: a list of the conceptual metaphors extracted from the German subcorpus. A linguistic metaphor is reported as example for each conceptual metaphor

	Die Zeit	Süddeutsche Zeitung	Die Welt	Frankfurter Allgemeine Zeitung	Bild
PEOPLE ARE ANIMALS			X		
UNCONTROLLED CRIME IS A RAMPAGING ANIMAL		X	X	X	X
ADDRESSING CRIME IS CONTROLLING ANIMAL		X	X	X	X
RELATIONSHIPS ARE OBJECTS		X			
MURDERERS ARE INHUMAN ENTITIES	X	X	X	X	X
SOCIAL MEDIA ARE A PLACE FOR MOURNING					X

Table 5.6: Distribution of conceptual metaphors among the five German newspapers.

5.5 Discussion

An in-depth comparison of the analyses of the two subcorpora carried out in the previous section revels some commonalities and some striking similarities.

An important similarity is the absence of metaphors concerning the victims or, in general, the female part of the episode. No linguistic metaphors have been detected in concomitance to the proper nouns of the victims, nor to the common nouns used to refer to them – e.g., mother, girl, female friend, woman. On the contrary, in both subcorpora the most common target domain is that of the criminals, and they are treated similarly in both subcorpora. Killer references function as subjects of verbs or are modified by adjectives belonging to the animal semantic area. In both subcorpora, the conceptual metaphors PEOPLE ARE ANIMALS, ADDRESSING CRIME IS CONTROLLING ANIMALS, MURDERERS ARE INHUMAN ENTITIES emerge as the principal conceptual metaphors, because of their frequency of use compared to the reduced frequency of the other conceptual metaphors. In fact, in the Italian subcorpus, eight terms are used as source terms to create linguistic metaphors in these fields, while conceptual metaphors with other source domains are generally composed starting from four or six terms. It is, furthermore, the only group of conceptual metaphors with 60 verbal occurrences. For what concerns the German subcorpus, 75 linguistic metaphors are characterised by the semantic field of animals as source domain. In the German subcorpus too there are more terms characterising this group of metaphors – four – than for the other conceptual metaphors. The more specific conceptual metaphor ADDRESSING CRIME IS CONTROLLING ANIMALS and the broader conceptual metaphor PEOPLE ARE ANIMALS have similar linguistic representations in the two subcorpora. Verbal metaphors are centred on the figure of the animal to be captured to solve the crime, giving rise to recurrent hunting images. Moreover, the terms used to modify nouns referring to the doer of the crime and the crime itself are almost the same – *brutal* and *bestial* for the German subcorpus, *brutal* and *savagely* in the Italian corpus – which add animal features to the killer. However, two main differences between the two subcorpora occur, namely presenting the killer as a predator – which occurs only one in the Italian corpus – and as prey of mental disease**s** or strong emotional states. Also in the German subcorpus the discourse of mental disease is reported, but it is always characterised by a more objective and literal language. The mental disease is presented as a disease and journalists use medical lexicon to introduce the disease in the narration, as *ein psychisch kranker Mann* ('a mentally ill man').

A similar commonality between the two subcorpora is the use of conceptual metaphors which do not only contribute to create the idea of a man more similar to a wild animal than to a human being, it reinforces this idea causing an effect of complete dehumanisation of the killer. The conceptual metaphor, which finds no confirmation on MetaNet, is here reported as MURDERERS ARE INHUMAN ENTITIES. As for almost all the conceptual metaphors extracted in this analysis, more linguistic metaphors occur in the Italian subcorpus than in the German subcorpus. This conceptual metaphor is developed in two ways in both subcorpora. In the first way, the target domain is the criminal, embedded in a discourse of monsters and unnatural entities. In linguistic expressions of the second way of developing this conceptual metaphor, the target domain is still the criminal, but he vanishes in favour of the crime itself. The only occurrence of the first way of developing MURDERERS ARE INHUMAN ENTITIES in the German subcorpus is the linguistic metaphor *the world of the other criminals*, which gathers human beings on a planet, and criminals on a different planet. In this verbal metaphor, criminals are implicitly classified as non-humans. This strong dehumanisation appears in the Italian subcorpus through the domain of monsters: killers are defined as *diabolic* or as *monster*. The second way of linguistically translate this conceptual metaphor is through the construction *victim of* + *noun*. This is the only metaphorical context, in both subcorpora, in which the victim occurs. The figure of the killer is not directly implied in this metaphor, but rather omitted and the empty space is filled by the criminal act itself or by anger or madness, killer’s emotional states.

These three conceptual metaphors, which are detected in both subcorpora, even if with slightly different verbal representations, may be the result of a trend of thinking that some behaviours are not proper of human beings, and if you act in a cruel way, you lose your humanity. This may be a way of justifying the act and to remove responsibility from the killer, as observed by Gius and Lalli (2014) in their analysis of femi(ni)cide Italian articles. They theorise that the use of words such as ‘raptus’ or ‘jealousy’ has the aim of removing responsibility from the perpetrator. They also observe that some responsibility is given to the criminal only when bounded to animal terms. However, using words belonging to the semantic field of animal surely recognises the fault of the single person but, implicitly saying that the criminal is not a man anymore, a part of fault is automatically removed because animals follow instinct and do not have the same conscience required to live in a human society. Therefore, even if they seem to recognise the role of the man in the violent act, linguistic metaphors using ANIMALS as source domain and CRIMINALS as target domain produce a de-responsibility effect similar to the effect caused by metaphors where the fault of the crime is given to moments of emotional instability.

A further metaphor detected in both corpora is RELATIONSHIPS ARE OBJECTS. It has only one occurrence in the German subcorpus – *on-off relationship* – while in the Italian subcorpus is linguistically express by *relazione incrinata* and *ricomporre la crisi.*

The metaphors listed so far appear in both subcorpora, but there are metaphors extracted only from the Italian subcorpus or only from the German subcorpus. A conceptual metaphor which does not appear in the Italian subcorpus are metaphors concerning the field of social media. While in the Italian subcorpus the lexicon used to talk about social media is the proper lexicon of the field of internet and social media, the lexicon used in the German subcorpus is more emotional. In fact, in the German subcoprus, people mourn on Facebook. They do not only share a post, they directly share their feelings as it were a private diary or a safe and intimate place where they can vent their emotions and their suffering.

On the other hand, some conceptual metaphors are found only in the Italian subcorpus, as ARGUMENT IS WAR and PEOPLE ARE PLANTS. Although usually considered as common metaphors and already theorised in Lakoff and Johnson (1980/2003), no linguistic metaphors connected to these conceptual metaphors is detected from the German subcorpus. A possible explanation may be that the German language is a more synthetic language as compared to Italian and it differs from Italian in its inherent structure as its meanings become more precise through the use of affixes added to the lemma. This different way of shaping meaning causes German to be a more precise and less abstract language, while in the Italian language is more common to find lemmas with less precise meaning. In fact, the German words formation process strongly uses the derivational morphemes to create words with more precise and concrete meanings (Bianco, 2004, Nied Curcio, 2016). Therefore, finding less metaphorical patterns in the German subcorpus may be connected to the characteristic feature of the German language of having more lemmas with precise meaning to be used in particular contexts. A complete list of the conceptual metaphors extracted from the two subcorpora is provided in the Venn diagram reported below, which makes easily understandable which conceptual metaphors appear in both subcorpora, and which conceptual metaphors appear only in the Italian or in the German part of the corpus.

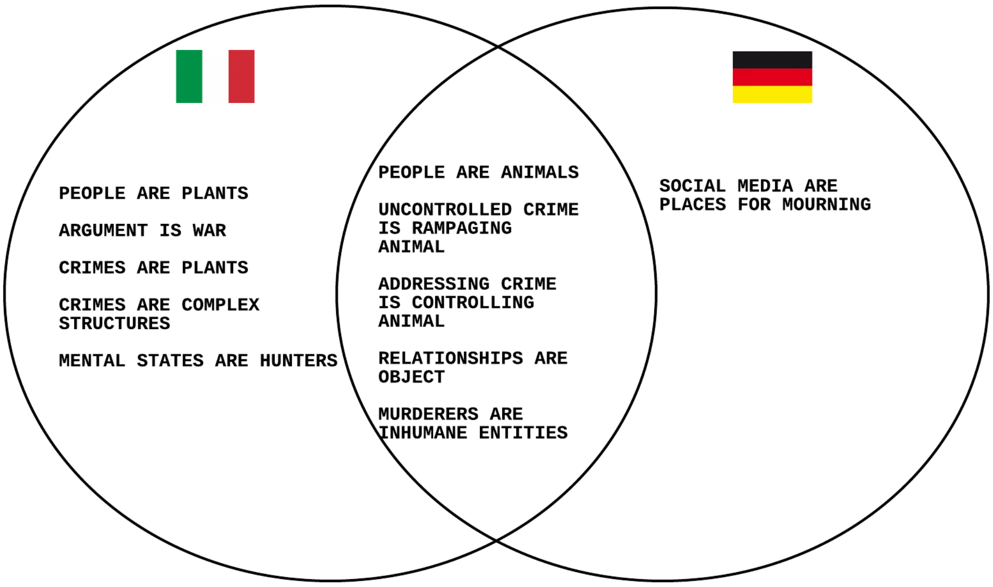

PEOPLE ARE PLANTS
ARGUMENT IS WAR
CRIMES ARE PLANTS
CRIMES ARE COMPLEX STRUCTURES
MENTAL STATES ARE HUNTERS
PEOPLE ARE ANIMALS
UNCONTROLLED CRIME IS RAMPAGING ANIMAL
ADDRESSING CRIME IS CONTROLLING ANIMAL
RELATIONSHIPS ARE OBJECT
MURDERERS ARE INHUMANE ENTITIES
SOCIAL MEDIA ARE PLACES FOR MOURNING

Figure 5.2: A Venn diagram reporting the conceptual metaphors extracted from the Italian subcorpus, from both subcorpora, and from the German subcorpus

Comparing the conceptual metaphors extracted from the Italian-German corpus to the results obtained by Busso et al. (2019), it is possible to noticed that some metaphors are used in both corpora.

Conceptual metaphors concerning the source frame of animals appear in both corpora, but Busso et al. (2019) detects metaphors concerning the figure of the woman which is a prey and an animal. No similar metaphor is found in the Italian-German corpus, even if the idea of the man as a beast and the hunting images characterise both corpora. On the other hand, metaphors with vegetables elements as source domain are detected in both Italian corpora, namely PEOPLE ARE PLANTS, but not in the German subcorpus. Also the conceptual metaphor ARGUMENT IS WAR is extracted from both Italian corpora but not form the German subcorpus. This similarity between the two Italian corpora and their dissimilarity with the German corpus is caused by the languages of the corpora. Conceptual metaphors are generally universal, but it is probable that some metaphors are more proper of some cultures instead of other cultures, or that some metaphors are used in some contexts in one culture and some different contexts in another culture. The time span from the publication date of the articles collected by Busso et al. (2019) and the newspapers articles collected for this Italian-German corpus is of two/three years, so they are written in the same cultural period. Although the cases reported in the articles are different, as they occurred in different ways and with different protagonists, there are some common traits among all femi(ni)cide cases, and the conceptual metaphors recurrent in Busso et al. (2019) and in this study deal with these aspects. Usually, femi(ni)cide cases involve a killer, a murdered victim, some or more arguments between the two people involved, the research of the killer by the police. Common conceptual metaphors are, in fact, those with ANIMALS as source domain, referring to the killer and to the search of the killer by the police, which occur in Busso et al. (2019), in the Italian subcorpus and in the German subcorpus. Conceptual metaphors as PEOPLE ARE PLANTS, usually concerning the people involved in the crime and conceptual metaphors concerning the fights between the killer and the victims only occur in Busso et al. (2019) and the Italian subcorpus. However, also metaphors shared between the Italian corpus of Busso et al. (2019) and the German subcorpus are detected. In the Italian subcorpurs, no metaphor concerning the area of social media or new technologies occurred, while in Busso (2019) many metaphors deal with this area. Also the German subcorpus presents a few metaphors centred on the target domains of social networks. In the two corpora, however, the target domain is treated differently, and different conceptual metaphors merged. The merging of different conceptual metaphors about the same target domain does not seem to be a culturally bounded phenomenon, but rather a phenomenon related to the type of femi(ni)cide cases reported in the corpus. As already introduced in the previous chapters, Busso et al. (2019) starts collecting their corpus after the death of a victim of revenge porn, a case which became quite famous in Italy, and which sees the web in general and social media in particular as the main characters of the event. In fact, conceptual metaphors reported by Busso et al. (2019) seem plausible to be extracted from particular cases where social media had a huge weight. On the contrary, the conceptual metaphor concerning web in the German subcorpus seems to be more general and suitable to any type of situations, even to personal loss which are not bounded at all to crime cases.

A further conceptual metaphor detected from all three corpora has the supernatural/inhumane as source domain and the murderer as target domain. However, we land on two different conceptual metaphors. According to the data extracted from the Italian-German corpus, a conceptual metaphor as MURDERERS ARE INHUMAN ENTITIES seem to be more suitable, while Busso et al. (2019) theorises a conceptual metaphor as ABUSE IS AN OVERWORDLY EXPERIENCE. A conceptual metaphor like ABUSE IS AN OVERWORDLY EXPERIENCE seems to be still part of the process of removing responsibility to the author of the crime as the author is not even mentioned in the conceptual metaphor. However, if a crime occurs, it occurs because people with their actions caused it. The author of the crime is still part of the metaphors even when not reported in the linguistic metaphor, it is the person causing the experience to be firstly identified as non-human, and, as a consequence, every action made by that person results to be *overwordly*. In fact, as people are the agents of the situation, even in linguistic metaphors as *si è scatenato l'inferno*, violence took place since one or more persons decide to become violent, not because of the violence itself. Moreover, taking into account the only German linguistic metaphor of this type, it does not take into consideration abuse in general, but people committing abuse, they are considered to be from another world, not the violence. Therefore, opting for a conceptual metaphor as MURDERERS ARE INHUMAN ENTITIES allows the criminals to keep the agentive force, underlying the fact that they are not considered human anymore, but still able to act. Instead, in the conceptual metaphor ABUSE IS AN OVERWORDLY EXPERIENCE where the target is the abuse, a linguistic metaphor as *abbiamo conosciuto le belve* is not to be read as referred to the person committing the violence but to the violence itself. With this type of reading, the linguistic metaphor result to be quite difficult to understand. However, the definitions seem to be nothing but a formal particular in this case, where the considerable thing is the presence of the same metaphorical mapping in both parts of the Italian-German corpus and in Busso et al. (2019) analysis.

In conclusion, common conceptual metaphors are those concerning the criminals as target and animals as the source domain, while metaphors including the victim as target domain are omitted in both subcorpora. The presence of different conceptual metaphors seems to be caused by the different peculiarities of each event – e.g., whether social media plays an important role in the crime – or to language specific peculiarities – as the general characteristic of German language of having more words with more specific meanings than Italian – and not to strong cultural differences not to different way of narrating the phenomenon on behalf of the media.

These results confirm the theory that femi(ni)cide is an extreme form of gender-based violence, which is perpetrated and accepted/justified every day in our society (Rodriguez. 2010). The violent actions of the perpetrators are almost lessened by the connection between the target domains MAN/CRIME and the source domains ANIMALS and INHUMAN ENTITIES. This implicit justification is strengthened in Italian newspapers by the conceptual metaphor MENTAL STATES ARE HUNTERS, which works recognising part of the fault, or probably all the fault, to the mental diseases or mental states, not to the man himself committing the crime. The idea of a man acting because he is possessed by some more powerful force cancel the systematicity of the violence, influencing its perception as an episode of isolated violence, and not as something occurring in a larger picture of common violence and patriarchal stereotypes. Moreover, this type of implicit justification occurs in all the newspapers in analysis through the linguistic metaphors 'victims of + noun of the crime event/mental states/violence', a phrase which causes a decriminalisation of the perpetrator as the real perpetrator of the event were the violent act itself.

According to the obtained results, the language used in the corpus seems to be a neutral and objective language at a first sight, mainly when considering the results obtained from the topic extractions, where only terms strictly related to the crime are found – e.g., terms related to the role of the victim and the perpetrator, the relatives of the people involved in the crime, law enforcement and the consequences of the crime, as the death of the victim and the subsequent legal process. On the other hand, the type of language emerging from the metaphor analysis seems a language aiming to diminish the phenomenon, as the perpetrator is exculpate trough the use of animal images, while the absence of metaphors concerning women seems a hint of the perception of the woman as a non-protagonist of the criminal event. The different worth given to the male and female characters of the event through the presence/absence of metaphors may be a clue of the patriarchal values behind femi(ni)cide and behind the whole culture which tends, even if unconsciously, to perceive men as more powerful and more important than women, and therefore more attention is paid on them also in the narration of criminal events and in the type of conceptual metaphors used in this narration.

Conclusions

This study focused on the narration of a universal phenomenon, which is considered to be increasing, but which was officially recognised only ten years ago from the Istanbul Convention. Of the two States whose media narration is taken in this study, Italy ratifies the Convention the year after it, while Germany ratifies the Convention only after five years (in 2017). These may be connected to the peculiarity of the phenomenon of femi(ni)cide of being generally more difficult to be recognised than the more general gender-based violence, even if it is considered an expression of this type of violence (Rodriguez, 2010). This type of violence is rooted in almost every society, as it may be seen in United Nations Office on Drugs and Crime (2018)'s report of femi(ni)cide cases all over the world in 2017. Moreover, other than being a world spread phenomenon, femi(ni)cide is seen as a primary cause of premature death among women and girls (Kouta et al., 2018). This type of violence is impregnated by the values at the basis of patriarchal society, which recognises a lower value to women than men, and occurs when these values are not observed by women and after other violent acts influenced by gender stereotypes rooted in society. Definition reported in Grzyb, Naudi and Marcuello-Servós (2018) explicitly recognises the role of gender inequality at the origin of the phenomenon:

> 'Feminicide: the killing of a woman because some man or men, although occasionally also some women who accept men's values, has or have sentenced her to death adducing whatever reasons, motives or causes, but nonetheless actually and ultimately because he or they believe she has defied (the words they often use are 'offended' or 'insulted') patriarchal order (in their words 'honourable' societies) beyond what her judge (often but not always the same person who kills her) is prepared to tolerate without retaliating in that way.' (Grzyb et al., 2018, p. 29).

This phenomenon in general, and the media narration on behalf of online newspapers in Italy and Germany in particular, are at the heart of this study. A bilingual corpus with newspaper articles about femi(ni)cide has been created and analysed with text mining techniques, aiming to extract topics and metaphors from documents. The two experiments have been conducted separately for each subcorpus, in order to respect the language peculiarities and to obtain clearer results, even if the same methodologies are applied to both subcorpora.

The algorithm used to extract topics is Latent Dirichlet Allocation algorithm with the number of topics set to four. Topics extracted from both subcorpora are strictly related to the criminal event narrated in the newspaper articles, as they belong to the semantic field of crime, justice, death, and private relationships. Words appearing in more than one topic and with a high probability rate are woman – *donna, Frau* – and police – *carabinieri, Polizei* – in both set of topics, while a term missing in the Italian set of topics and appearing only one time in the German set of topics is perpetrator – *colpevole, Täter*. Taking into account this analysis, the two subcorpora show similar results.

The method used to extract metaphors is more analytical. It is based on collocations of the target terms and, then, of the source terms. Firstly, the collocations of some common words belonging to the semantic field of crime and people involved in the crime are analysed. From their contexts of use, some metaphorical expressions are detected. Collocations of the source terms of these metaphorical expressions are analysed in order to see if they are used with other target terms and to quantify how often they are used metaphorically. All the linguistic metaphors extracted from the documents are then grouped together to reach a broader metaphor, the conceptual metaphor behind them. Conceptual metaphors extracted from the two subcorpora have as target domain the perpetrator or the crime in general – PEOPLE ARE ANIMALS, ADDRESSING CRIME IS CONTROLLING ANIMAL – but also relationships – RELATIONSHIPS ARE OBJECTS. Metaphors appearing only in the Italian corpus concerns mental diseases and mental states – MENTAL STATES ARE HUNTERS – and also the more classical conceptual metaphor ARGUMENT IS WAR. This difference may be mainly due not by different way of treating the criminal phenomenon, but rather to differences of the language themselves. A further conceptual metaphor detected only in one subcorpus, the German subcorpus, concerns social media intended as a place to grieve. This metaphor is absent in the Italian subcorpus not because social media are not intended in a similar way, but because the language used to talk about social media is always its proper language in the Italian subcorpus.

From the results obtained from topics and metaphors extraction, the narration of the phenomenon seems to be similar across the two countries. Some differences arose, both in the topic extracted and in the metaphors detected from the two subcorpora, but they do not seem to justify a different way of perceiving the phenomenon across the two cultures. The set of topics merged from the two subcorpora do not display many differences, as they both concern the same semantic fields. In addition, all the terms extracted by the LDA algorithm are term strictly bound to the criminal event in object, as law enforcement, private relationships, and words concerning the consequences on the victim and the perpetrator.

In the same way, conceptual metaphors mainly concern the criminal event narrated in the newspaper articles and some of them are shared between the two subcorpora. Conceptual metaphors characterising only one subcorpus – as MENTAL STATES ARE HUNTERS, CRIMES ARE COMPLEX STRUCTURES and ARGUMENT IS WAR for the Italian subcorpus, and SOCIAL MEDIA ARE PLACES FOR MOURNING for the German subcorpus – does not seem to be bound to a different perception of the event narrated, but rather to linguistic differences.

From the first analysis, the topic modelling analysis, an objective and impartial language seems to merge, which does not have any connection to any type of patriarchal society, but rather represent a society where the crime of femi(ni)cide is described as a crime in which both the man and the woman involved have the same value, where the perpetrator and the victim are at the same level. The only type of consideration that may be inferred from the topic analysis is that the victim of the crime is a female and that the main characters of the documents of the corpus are people bound to the broad are of crime and justice.

Form the metaphors analysis, however, a society narrating the event of femi(ni)cide through a language bound to a gender-unbalanced society arises. These considerations concern both part of the corpus, as the only linguistic metaphors occurring in all the ten newspapers is 'victim of' followed by *femi(ni)cide* – this term is used only in Italian articles –, *violence*, and words referred to the crime, or mental diseases or mental states. This type of linguistic metaphor, conceptually recognised as MURDERERS ARE INHUMAN ENTITIES, plays an important role in the decriminalisation of the perpetrator, as in these metaphors the murderer is not the person committing the crime, but the mental states is acting on behalf of the person, or the criminal act itself. A similar result is obtained using conceptual metaphors employing ANIMALS as source domain and MAN/CRIME as target domain. The mapping created between the regular simultaneous activation of the two nodes justifies a frequent use of the association MAN/CRIME – ANIMAL, probably also outside the discourse of femi(ni)cide, and it strongly contributes to partially remove the fault from the perpetrator. This idea of the criminal as an animal implicitly justifies the violent act, because animals follow the instincts and are not expected to act according to reason, as human being should. This idea is strengthened further by the conceptual metaphor MENTAL STATES ARE HUNTERS. This group of metaphors (PEOPLE ARE ANIMALS, UNCONTROLLED CRIME IS A RAMPAGING ANIMAL, ADDRESSING CRIME IS CONTROLLING

ANIMAL, MURDERERS ARE INHUMAN ENTITIES, MENTAL STATES ARE HUNTERS), is detected in all the newspapers in analysis and seem to be a clue of the implicit justification of the phenomenon by the hand of the journalist. This justification is probably unconscious as it roots in the intrinsic values of our societies (Johnson, 1995, Rodriguez, 2010). This attitude towards femi(ni)cide is not only mirrored in the conceptual metaphors used to describe the event, but also in the absence of conceptual metaphors implying the woman as target domain. The lack of conceptual metaphors about the female protagonist of the event seems to highlight the lower value given to the woman in societies still characterised by gender inequality, as the woman is not considered as one of the two protagonists of the event, but the main character is only the man, who is included as target domain in metaphors aiming to lowering his fault.

This study is only a little work, it is surely not enough to fill the lack of linguistic studies about femi(ni)cide narration and femi(ni)cide in general. As already introduced in the first chapters of the thesis, it was quite hard to find corpus studies investigating femi(ni)cide narration. This lack of literature and the information provided by more sociological studies (Weil, 2018, UNODC, 2018) suggest that research on this field should be increased as it may help us improve our understanding of a large-scale social phenomenon which causes a huge number of deaths among women and girls all over the world.

Ideas for possible future research could involve the type of texts collected to build the corpus and the general methodology used to conduct the experiments. The employment of types of texts different than news articles would probably reveal different results for what concerns the figure of the woman, which results to be absent in the metaphorical expressions extracted from both subcorpora. It would also probably reveal a more prominent presence of the term perpetrator for the topics extracted, which is absent in all the Italian topics. It would be therefore interesting to analyse types of texts with different features, where the amount of subjectivity of the writer is required not to be at a minimum level.

In general, the methodology to analyse the corpus resulted to be a sort of compromise, between what seemed to be more suitable for the Italian subcorpus and what seemed to work better for the German subcorpus. Although a minimum of independence is kept for what concerns language specificity – e.g., the different types of collocations looked for extracting metaphors in German and in Italian –, the research line is that of trying to analyse both subcorpora in the most similar way possible, for both tasks. A differentiated analysis would probably provide with different results, respecting the specificity of each subcorpurs.

Analyses implying different types of texts or also more differentiated methodologies may bring to other interesting and unexpected results, contributing to the awareness raising goal set by the CEDAW (Convention on the Elimination of All Forms of Discrimination against Women) and by the Council of Europe Convention on Combating and Eliminating Violence against Women and Domestic Violence.

References

Baldry, A. C., & Magalhães, M. J. (2018). Prevention of femicide. In W. Shalva, C. Corradi & M. Naudi (Eds.), *Femicide across Europe: Theory, research and prevention* (pp. 71-92). Bristol, UK: Bristol University Press.

Baker, P. (2006). Beyond Collocation. *Using Corpora in Discourse Analysis,* (pp. 153-176). London, UK/New York, NY: Continuum.

Baker, P. (2014). *Using Corpora to Analyze Gender*. London, UK/New York, NY: Bloomsbury Academic.

Baroni, M., & Bernardini, S. (2004). BootCaT: Bootstrapping Corpora and Terms from the Web. *Proceedings of LREC 2004*, 1313-1316.

Baroni, M., Bernardini, S., Ferraresi, A., & Zanchetta, E. (2009). The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation*, *43*(3), 209-226.

Bianco M. T. (2008). *Introduzione al lessico del tedesco* (3rd ed.). Bari, IT: Edizioni B.A. Graphis

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84. doi: 10.1145/2133806.2133826

Blei, D. M., Ng, A. Y., & Michael, I. J. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research 3*, 993-1022.

BNC Consortium, *The British National Corpus, XML Edition*, 2007, Oxford Text Archive, http://hdl.handle.net/20.500.12024/2554.

The BNC Baby, version 2. 2005. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/

Boleda, G. (2020). Distributional Semantics and Linguistic Theory. *Annual Review of Linguistics*, *6*, 213-234. doi: 10.1146/annurev-linguistics-011619-030303

Brookes, G., & McEnery, T. (2019). The utility of topic modelling for discourse studies: A critical evaluation. *Discourse Studies 21*(1), 3-21. doi: 10.1177/1461445618814032

Bundeskriminalamt. (2020). Partnerschaftsgewalt. Kriminalstatistische Auswertung – Berichtsjahr 2019

Busso, L., Combei, R. C., & Tordini O. (2019). La rappresentazione lessicale della violenza di genere: "Donne come vittime" nei media italiani. In B. Aldinucci, V. Carbonara, G. Caruso, M. La Grassa, C. Nadal, & E. Salvatore. *Parola. Una nozione unica per una ricerca interdisciplinare* (pp.261-279). Siena, IT: Edizioni Università per Stranieri di Siena.

Collins Cobuild English grammar. (2005). Glasgow, UK: HarperCollins.

Commission on Crime Prevention and Criminal Justice. (2014, May 8). Statement by Germany on the investigation and prosecution of gender-related killings of women and girls. Annex.

Corradi, C., Baldry, A. C., Buran, S., Kouta, C., Schröttle, M. & Stevkovic, L. (2018). Exploring the data on femicide across Europe. In W. Shalva, C. Corradi & M. Naudi (Eds.), *Femicide across Europe: Theory, research and prevention* (pp. 93-166). Bristol, UK: Bristol University Press.

Cserép, A. (2014). Conceptual Metaphor Theory: in defence or on the fence? *Argumentum 10,* 261-288.

Dancygier, B. (2017). Figurativeness, conceptual metaphor, and blending. In E. Semino, Z. Demjén (Eds.), *The Routledge Handbook of Metaphor and Language* (pp. 28-41). Abingdon, UK/New York, NY: Routledge.

Davies, M. (2010-). *The Corpus of Historical American English: 400 million words, 1810-2009.*

Deignan, A. (2005) *Metaphor and Corpus Linguistics*. Amsterdam, NL/Philadelpha, PA: John Benjamins Publishing Company.

Deignan, A. (2006). The grammar of linguistic metaphors. In A. Stefanowitsch, & S. T. Gries (Eds.), *Corpus-based approaches to metaphor and metonymy* (pp. 106-122). Berlin, DE: Mouton de Gruyter.

Deignan, A. (2017). From Linguistic to Conceptual Metaphors. In E. Semino, Z. Demjén (Eds.), *The Routledge Handbook of Metaphor and Language* (pp. 102-116). Abingdon, UK/New York, NY: Routledge.

Deignan, A., Gabryś, D., & Solska, A. (1997). Teaching English metaphors using cross-linguistic awareness-raising activities. *ELT Journal, 51*(4), 352-360. doi: 10.1093/elt/51.4.352

Dorst, A. G. (2017). Textual patterning of metaphor. In E. Semino, Z. Demjén (Eds.), *The Routledge Handbook of Metaphor and Language* (pp. 178-192). Abingdon, UK/New York, NY: Routledge.

Foucault, M. (2013). *History of Madness*. London, UK: Routledge. doi: 10.4324/9780203642603

Francis, N. W., & Kucera, H. (1979). *Brown Corpus Manual*/Department of Linguistics, Brown University, Providence, Rhode Island, US.

Gatto, M. (2014). *The web as corpus: theory and practice*. London, UK/New York, NY: Bloomsbury Academic.

Gibbs, R. W. (2011). Evaluating Conceptual Metaphor Theory. *Discourse Processes, 48*(8), 529-562. doi: 10.1080/0163853X.2011.606103

Gius, C., & Lalli, P. (2014). "I loved her so much, but I killed her". Romantic love as a representational frame for intimate partner femicide in three Italian newspapers. *ESSACHESS. Journal for Communication Studies, 7*(2), 53-75.

Greenbaum, S. (1991). ICE: The International Corpus of English. *English Today, 7*(4), 3-7. doi:10.1017/S0266078400005836

Gries, S. T., & Newmann, J. (2013). Creating and using corpora. In R. J. Podesva, D. Sharma (Eds). *Research Methods in Linguistics* (pp. 257-287). Cambridge, UK: Cambridge University Press.

Grzyb, M., Naudi, M. & Marcuello-Servós, C. (2018). Femicide definitions. In W. Shalva, C. Corradi & M. Naudi (Eds.), *Femicide across Europe: Theory, research and prevention* (pp. 17-32). Bristol, UK: Bristol University Press.

Hamks, P. (2006). Metaphoricity is gradable. In A. Stefanowitsch, & S. T. Gries (Eds.), *Corpus-based approaches to metaphor and metonymy* (pp. 17-35). Berlin, DE: Mouton de Gruyter.

Hofmann, T. (1999). Probabilistic Latent Semantic Analysis. *Proc. of the Fifteenth Conference on Uncertainty in Artificial Intelligence, UAI'99*, 289–296. San Francisco, CA: Morgan Kaufmann Publishers Inc. doi: 10.1145/312624.312649

Hunter, J. D. (2007) Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering, 9*(3), 90-95. doi: 10.5281/zenodo.592536

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., & Suchomel, V. (2013). The TenTen Corpus Family. *7th International Corpus Linguistics Conference CL 2013*.

Jo, W. (2019). Possibility of Discourse Analysis using Topic Modeling. *Journal of Asian Sociology, 48*(3), 321-342. doi: 10.21588/jas/2019.48.3.002

Johansson, S. (1978). *Lancaster-Oslo-Bergen corpus of modern English (LOB): [tagged, horizontal format]*. University of Oxford, Oxford Text Archive. http://hdl.handle.net/20.500.12024/0167

Johnsons, M. P. (1995). Patriarchal Terrorism and Common Couple Violence: Two Forms of Violence against Women. *Journal of Marriage and Family, 57*(2), 283-294.

Jurafsky, D., & Martin, J. H. (2020). Naive Bayes and Sentiment Classification. *Speech and Language Processing* (2nd ed. draft).

Kilgarriff, A. (2007). Googleology is bad science. *Computational Linguistics 33*(1), 147-151. doi: 10.1162/coli.2007.33.1.147

Kilgarriff, A., Grefenstette, G. (2001). Web as corpus. *Proceedings of Corpus Linguistics 2001,* 342-344.

Kilgarriff, A., Grefenstette, G. (2003). Introduction to the Special Issue on the Web as Corpus. In *Computational Linguistics 29*(3), 333-347. doi: 10.1162/089120103322711569

Kilgarriff, A., Rychlý, P., Smrž, P., & Tugwell, D. (2004) ITRI-04-08 the sketch engine. *Information Technology, 105.*

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel. V. (2014). The Sketch Engine: ten years on. *Lexicography, 1*, 7-36.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. *Proceedings of machine translation summit x: papers*,79-86.

The Kolhapur corpus of Indian English. (1986). Department of Indian English Studies, Shivaji University.

Koller, V. (2006). Of critical importance: Using electronic text corpora to study metaphor in business media discourse. In A. Stefanowitsch, & S. T. Gries (Eds.), *Corpus-based approaches to metaphor and metonymy* (pp. 237-266). Berlin, DE: Mouton de Gruyter.

Kouta, C., Boira, S., Nudelman, A. & Gill., A. K. (2018). Understanding and preventing femicide using a cultural and ecological approach. In W. Shalva, C. Corradi & M. Naudi (Eds.), *Femicide across Europe: Theory, research and prevention* (pp. 53-70). Bristol, UK: Bristol University Press.

Kövecses, Z. (1991). A linguist’s quest for love. *Journal of Social and Personal Relationships, 8,* 77-97.

Kövecses, Z. (2017). Conceptual metaphor theory. In E. Semino, Z. Demjén (Eds.), *The Routledge Handbook of Metaphor and Language* (pp. 13-27). Abingdon, UK/New York, NY: Routledge.

Lakoff, G. (2014). Mapping the brain’s metaphor circuitry: metaphorical thought in everyday reason. *Frontiers in Human Neuroscience, 8.* doi: 10.3389/fnhum.2014.00958

Lakoff, G., & Johnson M. (2003) *Metaphors we live by* (2nd ed.). Chicago, IL: The University of Chicago Press.

Lenci, A., Montemagni, S., & Pirrelli, V. (2005). *Testo e computer. Introduzione alla linguistica computazionale.* Roma, IT: Carocci editore.

Lenci, A. (2018). Distributional Models of Word Meaning. *Annual Review of Linguistics, 4*, 151-171. doi: 10.1146/annurev-linguistics-030514-125254

Levin, S. R. (1982). Aristotle’s Theory of Metaphor. *Philosophy & Rhetoric, 15*(1), 24-46.

Mácha, J. (2016). Conceptual Metaphor Theory and Classical Theory: Affinities Rather than Divergences. In P. Stalmaszczyk (Ed). *Philosophy of Fiction to Cognitive Poetics* (pp. 93-115). Frankfurt am Main, DE: Peter Lang. doi:10.3726/978-3-653-06564-0

Martin, J. H. (2006). A corpus-based analysis of context effects on metaphor comprehension. In A. Stefanowitsch, & S. T. Gries (Eds.), *Corpus-based approaches to metaphor and metonymy* (pp. 214-236). Berlin, DE: Mouton de Gruyter.

Maslen, R. (2017). Finding systematic metaphors. In E. Semino, Z. Demjén (Eds.), *The Routledge Handbook of Metaphor and Language* (pp. 88-101). Abingdon, UK/New York, NY: Routledge.

McEnery, T., McGlashan, M., & Love, R. (2015). Press and social media reaction to ideologically inspired murder: The case of Lee Rigby. *Discourse and Communication*, *9*(2), 237-259. doi: 10.1177/1750481314568545

Navarro-Colorado, B. (2018). On Poetic Topic Modeling: Extracting Themes and Motifs From a Corpus of Spanish Poetry. *Frontiers in Digital Humanities, 8.* doi: 10.3389/fdigh.2018.00015

Nied Curcio, M. (2016), *La lingua tedesca. Aspetti linguistici tra contrastività e interculturalità*. Roma, IT: UniversItalia di Onorati.

Petruck, M. (2018) *MetaNet*, Amsterdam, NE: John Benjamins.

Pinelo, A. L. (2017). En Alemania "no hay feminicidios" según su gobierno pero en el 2015 se registraron 311 feminicidios íntimos. *Feminicidio.net.*

Pinelo, A. L. (2018). A Theoretical Approach to the Concept of Femi(ni)cide. *The Philosophical Journal of Conflict and Violence, 2*(1), 41-63. doi: 10.22618/TP.PJCV.20182.1.171003

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In A. Celikyilmaz, T. H. Wen (Eds). *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstration*, 101-108. Association fro Computational Linguistics. doi: 10.18653/v1/2020.acl-demos.14

Řehůřek, R & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks,* 45-50. Malta: University of Malta.

Röder, M., Both, A, & Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures. WSDM'15. doi: 10.1145/2684822.2685324

Rodriguez, G. (2010). From Misogyny to Murder: Everyday Sexism and Femicide in a Cross-Cultural Context. *UCLA: Center for the Study of Women.*

Santa Ana, O. (1999). 'Like an animal I was treated': anti-immigrant metaphor in US public discourse. *Discourse & Society, 10*(2). 191-224. doi: 10.1177/0957926599010002004

Schmid, H., & Laws, F. (2008). Estimation of Conditional Probabilities with Decision Trees and an Application to Fine-Grained POS Tagging, *COLING 2008,* Manchester, Great Britain.

Schröttle, M. & Meshkova, K. (2018). Data collection: challenges and opportunities. In W. Shalva, C. Corradi & M. Naudi (Eds.), *Femicide across Europe: Theory, research and prevention* (pp. 33-52). Bristol, UK: Bristol University Press.

Sievert, C., & Shirley, K. E. (2014). LDAvis: A method for visualizing and interpreting topics. *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces,* 63-70.

Simpson, R. C., Briggs, S. L., Ovens J., & Swales, J. M. (2002). *The Michigan Corpus of Academic Spoken English*. Ann Arbor, MI: The Regents of the University of Michigan.

Slater, M. D., Long, M., Bettinghaus, E. P., & Reineke, J. B. (2008). News coverage of cancer in the United States: a national sample of newspapers, television, and magazines. *Journal of health communication, 13*(6), 523-537. doi: 10.1080/10810730802279571

Smith, R. (2007). An Overview of the Tesseract OCR Engine. *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)(*2), 629-633. doi: 10.1109/ICDAR.2007.4376991

Somov, A., & Voinov, V. (2017). "Abraham's Bosom" (Luke 16:22-23) as a Key Metaphor in the Overall Composition of the Parable of the Parable of the Rich Man and Lazarus. *The Catholic Biblical Quarterly, 79,* 615-633. doi: 10.1353/cbq.2017.0081

Steen, G. (2017). Identifying metaphors in language. In E. Semino, Z. Demjén (Eds.), *The Routledge Handbook of Metaphor and Language* (pp. 73 – 87). Abingdon, UK/New York, NY: Routledge.

Stefanowitsch, A. (2006). Corpus-based approaches to metaphor and metonymy. In A. Stefanowitsch, & S. T. Gries (Eds.), *Corpus-based approaches to metaphor and metonymy* (pp. 1-16). Berlin, DE: Mouton de Gruyter.

Stefanowitsch, A. (2006). Words and their metaphors: A corpus-based approach. In A. Stefanowitsch, & S. T. Gries (Eds.), *Corpus-based approaches to metaphor and metonymy* (pp. 63-122). Berlin, DE: Mouton de Gruyter.

Stefanowitsch, A. (2021). Metaphor. *Corpus linguistics: A guide to the methodology*. (pp. 397-436). Berlin, DE: Language Science Press. doi: 10.5281/zenodo.3735822

Schmidt, H. & Laws, F. (2008). Estimation of Conditional Probabilities With Decision Trees and an Application to Fine-Grained POS Tagging. *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008),* 777–784.

Tissari, H. (2017). Corpus-linguistic approaches to metaphor analysis. In E. Semino, Z. Demjén (Eds.), *The Routledge Handbook of Metaphor and Language* (pp. 117-130). Abingdon, UK/New York, NY: Routledge.

United Nations Office on Drugs and Crime. (2018). Global study on homicide. Gender-related killing of women and girls.

Weil, S. (2018). Research and prevention of femicide across Europe. In W. Shalva, C. Corradi & M. Naudi (Eds.), *Femicide across Europe: Theory, research and prevention* (pp. 1-16). Bristol, UK: Bristol University Press.

Weil, S. & Naudi, M. (2018). Towards a European Observatory on Femicide. In W. Shalva, C. Corradi & M. Naudi (Eds.), *Femicide across Europe: Theory, research and prevention* (pp. 167-174). Bristol, UK: Bristol University Press.

Zhai, C. X., & Massung, S. (2016). Topic Analysis. *Text data management and analysis: a practical introduction to information retrieval and text mining,* (pp. 330-387). Association for Computing Machinery and Morgan & Claypool Publishers. doi: 0.1145/2915031.2915049

Sitography

La Repubblica: https://www.repubblica.it/

La Stampa: https://www.lastampa.it/

Il Giornale: https://www.ilgiornale.it/

Il Fatto Quotidiano: https://www.ilfattoquotidiano.it/

Il Resto del Carlino: https://www.ilrestodelcarlino.it/

Zeit: https://www.zeit.de/

Süddeutsche Zeitung: https://www.sueddeutsche.de/

Welt: https://www.welt.de/

Frankfurter Allgemeine Zeitung: https://www.faz.net/

Bild: https://www.bild.de/