Esra Abdelzaher & Ágoston Tóth:
Defining Crime: A multifaceted approach based on Lexicographic Relevance and Distributional Semantics
Argumentum 16 (2020), 44-63
Debreceni Egyetemi Kiadó
DOI: 10.34103/ARGUMENTUM/2020/4

Tanulmány

Esra Abdelzaher & Ágoston Tóth

Defining Crime: A multifaceted approach based on Lexicographic Relevance and Distributional Semantics

Abstract

This paper demonstrates how the parallel examination of distributional data and frame semantic information can expose word senses that are not documented in FrameNet. In our case study, we compare the distributional features of the word *crime* to its properties stored in the FrameNet database also considering dictionary data that we find in three online monolingual dictionaries. Our analysis indicates that *crime* has senses that are absent from FrameNet. The five senses that we identify can be separated on the basis of (a) frame hierarchies, (b) frame elements, (c) syntactic and semantic data extracted from corpora using lexicographical tools and (d) distributional similarity. Annotated examples are provided to demonstrate each sense.
Keywords: crime, FrameNet, distributional semantics, lexicographic relevance, Sketch Engine

1 Background

Corpus analysis is indispensable for delineating and documenting word senses in modern lexicography. Lexicographic relevance is the compendium of various types of information relevant to (a) the lexicographical analysis process, which collects "the facts essential to the discovery and recording of how a word behaves" (Atkins & Bouillon 2006: 37), and to (b) the synthesis process, which is carried out "when editors come to the task of formulating the actual dictionary entry" (Atkins & Bouillon 2006: 26). Atkins et al. (2003: 271) explain that lexicographically relevant information about a keyword includes paradigmatic, syntagmatic and statistical information.

This theory of lexicographic relevance is also connected to Frame Semantics, which emanates from Fillmore's early work on the interaction between semantics and syntax. Fillmore (1968) argues that "there are many semantically relevant syntactic relationships involving nouns and the structures that contain them, that these relationships […] are in large part covert but are nevertheless empirically discoverable" (Fillmore 1968: 5). He referred to "agent", "patient", "goal" and other semantic notions as different semantic *cases* based on the semantic relations between the given NP and the verb of the sentence in various versions of his Case Grammar.

The analysis of the semantic argument structure of a target word in terms of phrase type, grammatical function and semantic role led to the foundation of a comprehensive theory in cognitive semantics: Frame Semantics (FS) (Fillmore et al. 2001). FS proposes that a frame

Esra Abdelzaher & Ágoston Tóth:
Defining Crime: A multifaceted approach based on Lexicographic Relevance and Distributional Semantics
Argumentum 16 (2020), 44-63
Debreceni Egyetemi Kiadó
DOI: 10.34103/ARGUMENTUM/2020/4

"represent[s] story fragments, which serve to connect a group of words to a bundle of meanings" (Ruppenhofer et al. 2016: 7). In this relatively new theory, "frame elements" (FEs) replaced semantic cases, customizing the traditional concept of semantic roles. FS makes some general roles, such as the category of *agent*, more specific, e.g. *killer* and *terrorist* in the frames of KILLING and TERRORISM, respectively. In other words, the case categories of Case Grammar are now more frame-specific and less abstract. FEs are either core arguments, or non-core, adding peripheral information to the sentence, which is determined by linguists based on corpus data on a frame-by-frame basis.

Distributional methods of meaning representation originate from Firth's (1957) argument that the meaning of a word is distributed among the neighboring words or the company this word keeps. Distributional Semantics (DS) does not explore first-order co-occurrence phenomena (cases when words occur together in a sentence); instead, it documents cases where words co-occur with the same words (second-order co-occurrence) (Tóth 2014). Lists of distributionally similar words tend to include items that are similar in meaning to the target word. From an extremist perspective, Distributional Semantics is "the unique possible source of evidence for the exploration of meaning" (Lenci 2008: 6). Second-order co-occurrence data are always collected using automatized tools, since manual corpus analysis would be prohibitively slow in this case.

The present study relies on two information sources to improve our understanding of the meaning of the word *crime*: the FrameNet database (FN) (Baker et al. 1998) and the Sketch Engine system (Kilgarriff et al. 2004). While FrameNet is a large FS database and it serves us with manually edited lexicographical information about lexical units relevant in the context of the documented frames, Sketch Engine includes built-in tools to retrieve information about words that occur in its massive store of corpora and has extensive support for quantitative and qualitative work.

We attempt to answer the following questions:

(a) What are the similarities and differences between FN's list of lexicographically similar words (which are co-listed in the same frame or listed in related frames) to *crime* and the automatically retrieved list of distributionally similar words?
(b) Can distributional similarity reveal new lexicographical features of *crime*?

The rest of this study is structured as follows: section 2 provides the reader with more information about the theoretical background of the study, section 3 describes the data, methodology and procedure of analysis, section 4 displays results and section 5 discusses them. Concluding remarks are given in section 6.

2 Lexicographic Relevance in FrameNet and Distributional Semantics in Sketch Engine

Ruppenhofer et al. (2016: 11) established a similarity criterion based on lexicographic relevance. Word senses are judged similar if (a) they have the same number and type of arguments, (b) semantic relations between the target words and their arguments are the same, (c) they adopt the same perspective of experiential knowledge and (d) they have similar denotation. Formal differences, such as parts of speech and passive forms, are not accounted for in these similarity judgments. In this approach, statistical co-occurrence information is not lexicographically significant, either, but it gives more space for native intuition.

Esra Abdelzaher & Ágoston Tóth:
Defining Crime: A multifaceted approach based on Lexicographic Relevance and Distributional Semantics
Argumentum 16 (2020), 44-63
Debreceni Egyetemi Kiadó
DOI: 10.34103/ARGUMENTUM/2020/4

FrameNet is based on Frame Semantics. Its latest release, FN 1.7, hosts 1224 frames; nearly 89% of the frames are lexical (evoked by lexical units), the remaining frames are non-lexical (they are created to connect frames).[1] FrameNet provides a definition for each frame, defines the frame elements of the frames and lists the frame-evoking lexical units. FN can be seen as an ontological language resource as most frames inherit the structure of the 5 top-level frames: EVENT, RELATION, STATE, ENTITY, LOCALE and PROCESS. Besides the inheritance relation, FrameNet defines other relations among frames: "Inherits from", "Is Inherited by", "Perspective on", "Is Perspectivized in", "Uses", "Is Used by", "Subframe of", "Has Subframe(s)", "Precedes", "Is Preceded by", "Is Inchoative of" and "Is Causative of".

The idea of using FrameNet as a source of information for delineating senses is elaborated in the literature of lexicography by Atkins and Bouillon (2006) in their case study on the senses of the word *argue*. The authors point out that the appearance of the lexical unit *argue* in three FrameNet frames (REASONING, EVIDENCE and QUARRELING) has lexicographic relevance and it is an indicator of the existence of (at least) three senses worth documenting in dictionaries (Atkins & Bouillon 2006: 28–32).

Distributional semantics uses statistical models to represent meaning regardless of native or expert intuition. It assumes that similar words occur in similar contexts. As we have discussed it in section 1 of this paper, distributional meaning representation uses second-order co-occurrence information to detect similarity of meaning.

The Sketch Engine is a web application that implements distributional methods and syntactic analysis for processing corpora and for creating "word sketches", which sum up the grammatical and collocational behavior of words. Rychlý and Kilgarriff (2007) developed a distributional thesaurus function, which is available via the Sketch Engine. This distributional thesaurus starts by using a large lemmatized and parsed corpus to capture the context of each word. This context also includes the grammatical relations linking a target word to context words. Finally, the procedure selects words that share the same contexts. Large corpus size helps avoid noise and improves the accuracy of the thesaurus. To exemplify, the triplets "object, drink, beer" and "object, drink, wine" provide context-based information that places *beer* and *wine* in the same thesaurus category. If a lexicographer is interested in further exploration for similar words, the Sketch Engine also offers a "sketch difference" function based on distributional semantics. Sketch differences use lexical collocates and grammatical relations in the contexts of words to show how (dis)similar two words are (Kilgarriff et al. 2014).

3 Methodology

The current study uses the FrameNet database to gather lexicographic information about the lexical unit *crime.* This includes the different senses of the word, denotation, predicate-argument structure, the frames that the word has the potential to evoke and the frame elements of these frames. In addition, the study searches for lexical units evoking any relevant frame so that the final list of words can embrace items directly similar or indirectly relevant to *crime*, as well as their frame elements.

[1] Project status information has been retrieved from https://framenet.icsi.berkeley.edu/fndrupal/current_status on December 1, 2019.

Esra Abdelzaher & Ágoston Tóth:
Defining Crime: A multifaceted approach based on Lexicographic Relevance and Distributional Semantics
Argumentum 16 (2020), 44-63
Debreceni Egyetemi Kiadó
DOI: 10.34103/ARGUMENTUM/2020/4

Besides, the study makes use of three corpora to retrieve distributionally similar words. First, Sketch Engine's thesaurus retrieves words similar to *crime* from the *British National Corpus* (BNC), which is a 100-million-word collection of genre-diversified British English texts (Leech 1992). The BNC corpus is neither the largest nor the most up to date, but it has also been the primary corpus for creating FrameNet's frames, defining word senses and detecting frame elements. *EnTenTen2015* is the second general reference corpus analyzed in the study. It is a web-based corpus of 15 billion words from contemporary English texts (Jakubíček & Kilgarriff 2013). The *Timestamped JSI web corpus 2014-2019 English* (*Timestamped* for short) is the third large corpus selected to represent the concept of *crime*; it is a web corpus of news articles obtained from RSS feeds automatically (Trampuš & Novak 2012). All these corpora are available through the Sketch Engine.

The study uses the 20 most similar words to *crime* in each corpus, which gives us a unified wordlist of 33 word types. For every word, we retrieve (a) the sense that relates to crime, (b) evocative frame, (c) frame elements and (d) relevance to the frame of COMMITTING_CRIME. This initiates the comparison of words that are lexicographically and distributionally similar to *crime* in terms of denotation, frame and frame elements.

The primary concern of this study is the set of new words that are absent in FN and are flagged by distributional semantics as similar to *crime*. We explore this similarity through the sketch differences function in Sketch Engine to realize any lexicographic significance behind this similarity. Figure 1 illustrates the methodology of analysis.

Esra Abdelzaher & Ágoston Tóth:
Defining Crime: A multifaceted approach based on Lexicographic Relevance and Distributional Semantics
Argumentum 16 (2020), 44-63
Debreceni Egyetemi Kiadó
DOI: 10.34103/ARGUMENTUM/2020/4

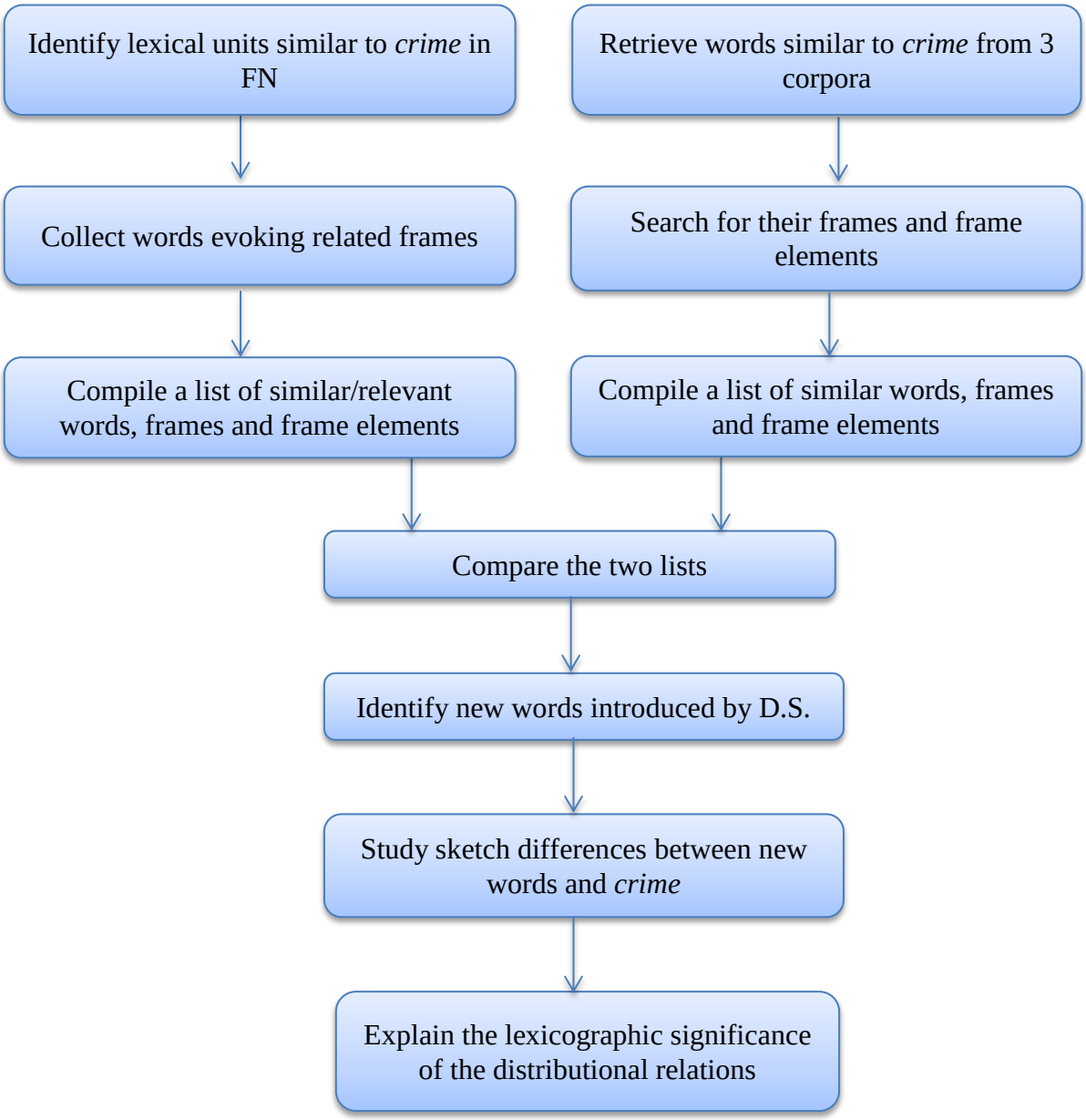

Figure 1. Methodology of Analysis

Esra Abdelzaher & Ágoston Tóth:
Defining Crime: A multifaceted approach based on Lexicographic Relevance and Distributional Semantics
Argumentum 16 (2020), 44-63
Debreceni Egyetemi Kiadó
DOI: 10.34103/ARGUMENTUM/2020/4

4 Results

By listing it in one frame only, FN acknowledges one sense of *crime*, which denotes an illegal act, "an action constituting a serious offence against an individual or the state and is punishable by law".[2] *Crime* activates the frame of COMMITING_CRIME in which "perpetrator" and "crime" are core FEs. The "perpetrator" can be lexically instantiated (Example A) or can receive indefinite null instantiations (INI) (Example B). However, if "crime" is lexically present, it plays a dual role. As a lexical unit, *crime* activates the frame of COMMITTING_CRIME, and – as a frame element – it fills the slot of "crime". The frame element "crime" can also be lexically instantiated through several other words, such as *treason*, *murder*, or *offence* (Example C). *Commission*, *commit*, *perpetrate* and *crime* are co-lexical units in the same frame. The following annotated examples are cited from FrameNet:[3]

A. The crimes$_{[crime]}$ of the Iraqi regime$_{[Perpetrator]}$
B. Organized crime$_{[crime]}$ was soon to have a formidable adversary $_{[perpetrator\ INI]}$
C. And how can he$_{[perpetrator]}$ commit treason against the King of England$_{[crime]}$

COMMITTING_CRIME is the first sub-frame in the complex frame CRIME_SCENARIO, where COMMITTING_CRIME precedes CRIMINAL_INVESTIGATION and CRIMINAL_PROCESS. These frames also have sub-frames. Overall, COMMITTING_CRIME is linked to 17 frames containing 164 lexical units (Appendix 1). The hierarchy that embraces *crime* in FN is the following:

EVENT → MISDEED → COMMITTING_CRIME

As far as the word similarity lists generated by the thesaurus function of Sketch Engine are concerned, the three selected corpora produced rather different results. While six words were common among the three corpora, 8 words were similar to *crime* only in the *EnTenTen2015* and *Timestamped* corpora. The final, unified similarity list included 33 words ordered by their highest similarity scores according to the *Timestamped* corpus (Appendix 2). Similarity scores exceeded 0.5 in *EnTenTen2015* and *Timestamped*, and it reached only 0.3 in *BNC* for the most similar word to *crime.*

The comparison of these similarity lists with FN's list of lexical units shows that 14% of distributionally-relevant words are already indexed in the FN database as related to *crime* while 5% of the words are missing; the remaining words are present but not linked to a relevant frame. Common words between distributional wordlists and FN belong to the frames of ABUSING, CRIMINAL_INVESTIGATION, TRIAL and OFFENSES. However, the focus of this study is placed on the 81% that is retrieved by distributional methods and is indexed in FN with no relevance to crime.

Table 1 lists these words along with their frames and FEs. It is worth noting that Sketch Engine's thesaurus retrieves (tagged) words rather than lexical units, so retrieved words may correspond to different frames. In such cases, POS and definitions were the criteria for choosing the most appropriate frame.

[2] This definition has been taken from the lexical entry for *crime.n* via https://framenet.icsi.berkeley.edu/fndrupal/luIndex on December 1, 2019.
[3] Retrieved from https://framenet.icsi.berkeley.edu/fndrupal/luIndex on December 1, 2019.

Esra Abdelzaher & Ágoston Tóth:
Defining Crime: A multifaceted approach based on Lexicographic Relevance and Distributional Semantics
Argumentum 16 (2020), 44-63
Debreceni Egyetemi Kiadó
DOI: 10.34103/ARGUMENTUM/2020/4

Similar Word	Frame	Core FEs
act, activity, action	INTENTIONALLY_ACT	Agent, Act
attack	ATTACK	Assailant, Victim
behaviour	CONDUCT	Agent, Manner
conflict	HOSTILE_ENCOUNTER	Issue, Purpose, Side1, Side2
corruption	MORALITY_EVALUATION	Behavior, Evaluee, Expressor
death	DEATH	Protagonist
disaster, crisis	CATASTROPHE	Patient, Undesirable Event
drug	INTOXICANTS	Intoxicant
health, disease	MEDICAL_CONDITIONS	Ailment, Patient
incident, accident	EVENT	Patient, Undesirable Event
incident, accident	CATASTROPHE	Place, Time, Event
killing	KILLING	Cause, Instrument, Killer, Means, Victim
offence	CAUSE_EMOTION	Agent, Event, Experiencer
violation	COMPLIANCE	Act, Norm, Protagonist, State of Affair
violence	VIOLENCE	Aggressor, Aggressors, Cause, Victim
practice	PRACTICE	Agent, Action, Occasion
problem	PREDICAMENT	Experiencer, Situation
threat	RISKY_SITUATION	Asset, Dangerous entity, situation
terrorism	TERRORISM	Act, Terrorist, Victim
poverty	WEALTHINESS	Institution, Person

Table 1. New words similar to crime *provided by Sketch Engine's thesaurus function*

Some of these frames are in contact with each other via higher-level nodes in the frame-hierarchy of FN. For example, figure 2 visualizes the hierarchy that descends from the EVENT frame based on the "Uses" and "Is Causative of" relations. Sketch Engines's thesaurus data suggest that *crime* is distributionally compatible with a kind of EVENT which can be an UNDESIRABLE_EVENT affecting a "patient", where the "act" is intentionally committed by the "agent" and it may affect a "victim" or an "experiencer". Regardless of the several viewpoints reflected by the FEs, all these frames indicate that *crime* is an event.

4.1 Crime in the EVENT top-level hierarchy

The distributional similarity between *crime, disaster* and *catastrophe* suggests that *crime* is related to a scene in which the "agent" is peripheral and the focus is on the "patient" element that undergoes an "undesirable event". CATASTROPHE in FN is illustrated by several examples, including "…various natural disasters[Undesirable_event] in central Asia[Patient]"[4] or "the humanitarian[Undesirable_event] crisis in Iraq[Patient]",[5] "human[Patient] and ecological[Patient]

[4] FN annotation data for *disaster.n* retrieved from https://framenet.icsi.berkeley.edu/fndrupal/luIndex on December 1, 2019.

[5] FN annotation data for *crisis.n* retrieved from https://framenet.icsi.berkeley.edu/fndrupal/luIndex on December 1, 2019.

Esra Abdelzaher & Ágoston Tóth:
Defining Crime: A multifaceted approach based on Lexicographic Relevance and Distributional Semantics
Argumentum 16 (2020), 44-63
Debreceni Egyetemi Kiadó
DOI: 10.34103/ARGUMENTUM/2020/4

catastrophe[Undesirable_event]".[6] The examinations of the sketch differences between *crime* and *disaster*, as well as between *crime* and *crisis* indicate that they can be used interchangeably in some cases. The following examples are extracted from *EnTenTen2015* and are manually annotated using the previous FEs of CATASTROPHE.

D. The Alberta tar sands are the world's most polluting source of transport fuel, and the one of the most devastating ecological *crimes*[Undesirable_event] of our times.
E. This small seaside village serves as a horrifying microcosm of massive ecological *crimes*[Undesirable_event] happening worldwide[Patient].
F. Tens of thousands of households are having their water service terminated for late payments. What is taking place in Detroit[Patient] is a social *crime*[Undesirable_event], which has the backing of the entire political establishment.
G. With smoked windows, it is a major ecological[Patient] crime[Undesirable_event].

INTENTIONALLY_ACT is a sub-hierarchy of EVENT gathering several words similar to *crime.* It includes, at the most general level, *act, activity* and *action*, which are general hypernyms of *crime*. According to FN, the lexical unit *crime* is "an action"[7] and the FE "crime" denotes an "act".[8] Similarly, sketch differences indicate general grammatical relations between *crime* and *act, activity* and *action*, such as "X is a …" and "… is X". Intentionality is a new dimension added by the distributional results and it allows linking *crime* to new frames. Words evoking TERRORISM, VIOLENCE, HOSTILE_ENCOUNTER, ATTACK, KILLING and DEATH share general coordination relations, such as "X and/or …", with *crime*. They also share hypernym and hyponym relations: "X is a …" and "… is X". Peripheral elements including most adverbial and prepositional phrases are also common among them. This may suggest that acts of terrorism, violence, killing and hostility should be related to COMMITTING_CRIME. They are relevant to the sense of *crime* mentioned in FN.

PRACTICE and CAUSE_EMOTION, however, propose a more general use of *crime*. Sketch differences between *crime, practice* and *offense* indicate that they can be used in similar contexts to refer to wrong acts that are not necessarily illegal. The following examples are extracted from *EnTenTen2015* and are manually annotated in accordance with the semantic argument structure of the PRACTICE frame.

H. I have always maintained that it is a *crime* to raise people's expectations to unattainable[Action].
I. People are starving. It is a *crime* to waste even a single grain[Action].

[6] FN annotation data for *catastrophe.n* retrieved from https://framenet.icsi.berkeley.edu/fndrupal/luIndex on December 1, 2019.
[7] As per the lexical entry for *crime.n* retrieved from https://framenet.icsi.berkeley.edu/fndrupal/luIndex on December 1, 2019.
[8] Based on the frame specification for COMMITTING_CRIME retrieved from https://framenet.icsi.berkeley.edu/fndrupal/frameIndex on December 1, 2019.

Esra Abdelzaher & Ágoston Tóth:
Defining Crime: A multifaceted approach based on Lexicographic Relevance and Distributional Semantics
Argumentum 16 (2020), 44-63
Debreceni Egyetemi Kiadó
DOI: 10.34103/ARGUMENTUM/2020/4

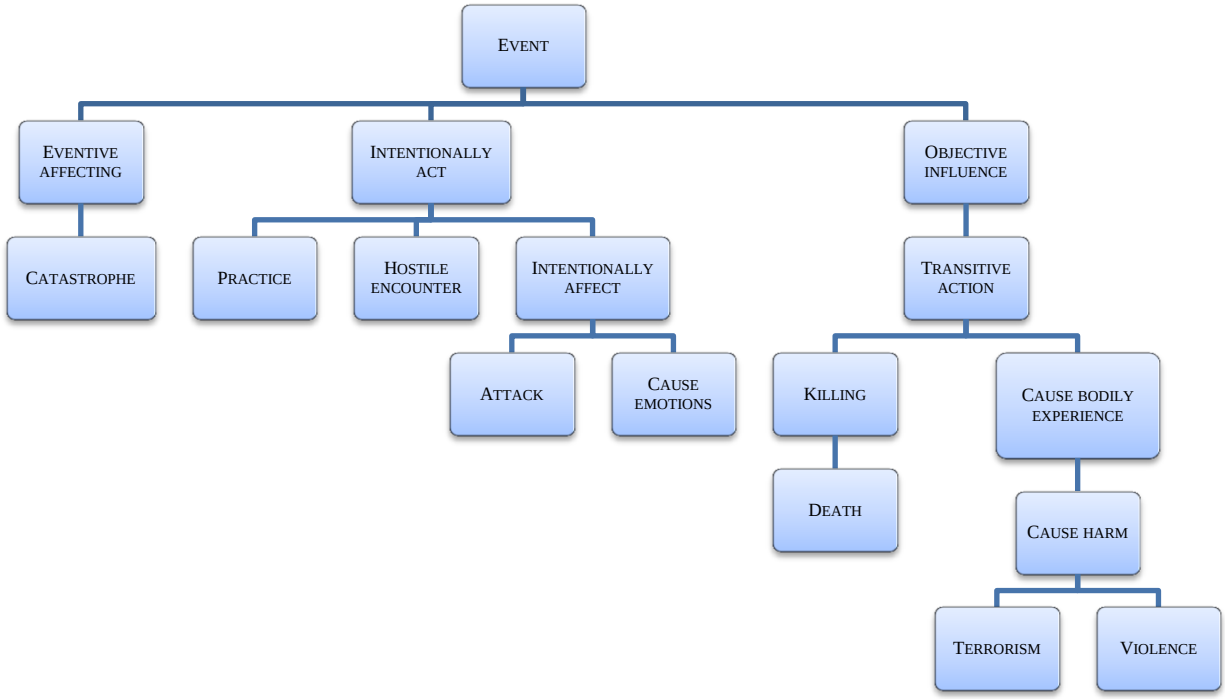

Figure 2. The EVENT*-hierarchy of new* crime*-related frames*

4.2 Crime in the ATTRIBUTES *hierarchy*

The second hierarchy is dominated by the ATTRIBUTES frame (Figure 3). In this hierarchy, words similar to *crime* (e.g. *poverty* in the WEALTHINESS frame or *disease* in MEDICAL_CONDITIONS) mostly refer to characteristics and qualities of persons and entities, not events or actions performed by agents. This new hierarchy signals a new sense of *crime* not connected to unlawful acts and undesirable events.

Esra Abdelzaher & Ágoston Tóth:
Defining Crime: A multifaceted approach based on Lexicographic Relevance and Distributional Semantics
Argumentum 16 (2020), 44-63
Debreceni Egyetemi Kiadó
DOI: 10.34103/ARGUMENTUM/2020/4

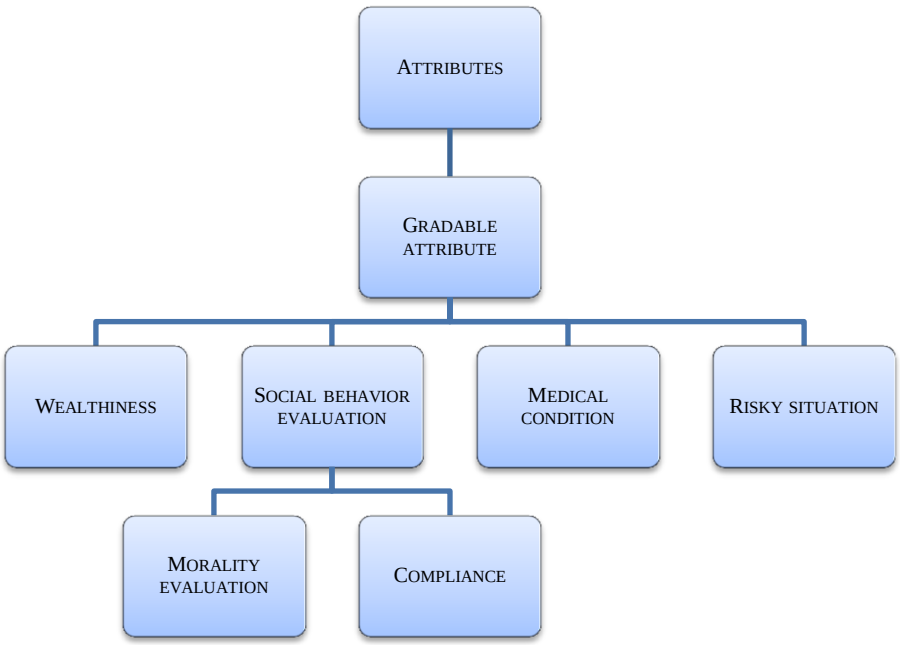

Figure 3. The ATTRIBUTES-hierarchy of crime*-related frames*

Checking the similarities between *crime*, *poverty* and *disease* in *EnTenTen2015* highlighted similar contexts in SketchEngine's sketch differences output, especially with the "subject, verb" and "verb, object" patterns. For instance, *prevent* and *combat* share *crime* and *disease* as objects. *Crime* and *disease* playing the grammatical role of the subject share the verbs *involve*, *occur*, *fight* and *affect*. *Combat*, *fight*, *reduce* and *tackle* are common verbs between *crime* and *poverty.*

Word sketches do not reveal a new sense of the word. They rather indicate metaphoric similarity related to the conceptualization of crime as a disease, poverty as a crime and poverty as a disease. According to MetaNet (Dodge et al. 2015), three conceptual metaphors link *crime* and *disease:* CRIME IS A DISEASE, CAUSING DECREASE IN CRIME IS TREATING DISEASE and INCREASE IN CRIME IS SPREAD OF DISEASE. The three metaphors are reflected in the shared verbs highlighted in the sketch differences retrieved, from the BNC, by Sketch Engine. Similarly, *poverty* and *crime* and *poverty* and *disease* are conceptually linked through POVERTY IS A CRIME and POVERTY IS A DISEASE.

The distributional similarity between *crime*, *corruption* and *violation*, however, suggested a new sense of *crime*. SOCIAL_BEHAVIOR_EVALUATION is a parent frame for both MORALITY_EVALUATION and COMPLIANCE, which are evoked by *corruption* and *violation*, respectively. SOCIAL_BEHAVIOR_EVALUATION contains a "judge" FE checking the "behavior" of an "individual" against "pre-existing standards of a society".[9] "Behavior" is a core FE, but it denotes an action, not an attribute, although the frame inherits data from GRADABLE_ATTRIBUTES. To further complicate matters, it is a non-lexical frame with no lexical units or annotated examples to suggest a solution to this hierarchy–FE conflict. The

[9] This information has been taken from the definition of the SOCIAL_BEHAVIOR_EVALUATION frame retrieved from https://framenet.icsi.berkeley.edu/fndrupal/frameIndex on December 1, 2019.

Esra Abdelzaher & Ágoston Tóth:
Defining Crime: A multifaceted approach based on Lexicographic Relevance and Distributional Semantics
Argumentum 16 (2020), 44-63
Debreceni Egyetemi Kiadó
DOI: 10.34103/ARGUMENTUM/2020/4

challenge is transferred to the two inheritor frames: MORALITY_EVALUATION and COMPLIANCE.

Acts that are described as crimes, although they are not criminal in the legal sense, involve a moral evaluation of some implicit judge or a violation of a societal norm. Re-annotating (H) and (I), which had been previously annotated according to the FE structure of PRACTICE, as well as new extracted examples, highlight this intersection. The following examples are extracted from *EnTenTen2015* and are manually annotated according to the semantic roles of SOCIAL_BEHAVIOR_EVALUATION:

J. I[Judge] have always maintained that it is a *crime* to raise people's expectations to unattainable[Behavior].
K. People are starving. It is a *crime* to waste even a single grain[Behavior].
L. It is a *crime* if she[Individual] is not getting upset with abortion[Behavior].
M. I want it all; hunger[Behavior] is my[Individual] *crime*.

The sense of wrong acts can be further customized to explain the whole judging process which involves (a) "judge", "evaluee"/"individual" and "behaviour" if wrong emanates from immorality, (b) "protagonist", "norm" and "act" if wrong emanates from violating the norms of a society.

The exploration of sketch differences between *crime, corruption* and *violation* also inferred an argument structure of *crime* fairly different from those of previous frames. Concordance lines showed that *crime* could be used to describe a characteristic which is negatively evaluated in a community or at a certain time. The following examples demonstrate this sense. They are extracted from the *EnTenTen2015* corpus and are manually annotated. The FEs of the top-level frame ATTRIBUTES are used to annotate these sentences to avoid referring to the "attribute" FE as a "behaviour" or an "act".

N. that it sounds as if it were a *crime* to be a Mexican[Attribute]
O. were arrested and imprisoned simply for the *crime* of being Irish in Britain[Attribute] at a particular period

4.3 Crime in the STATE top-level hierarchy

STATE is the top frame of the third hierarchy (figure 4), which has been introduced on the basis of the distributional similarity between *crime* and *problem* (cf. table 1 and Appendix 2). According to FN, *problem* refers to "an unwelcome or harmful" state which needs actions to be dealt with,[10] and it assigns the "situation" and "experiencer" FEs. *Crime* and *problem* share several grammatical relations and they can be used interchangeably in some contexts. Sketch differences between *crime* and *problem* in *EnTenTen2015* displayed that both fill the subject slot for *involve, occur* and *affect*. Also, both are preceded by the adjectives *prevalent* and *serious*. *Crime* can denote a problem or an undesirable situation when it refers to a harmful state suffered by an experiencer. This sense is frequently associated with reference to social, economic or political problems. The following examples are annotated according to the FE categories of PREDICAMENT.

P. It was a *crime* to be unemployed[Situation]
Q. It is a *crime* to be in Mosul[Situation]

[10] As per the lexical entry for *problem.n* retrieved from https://framenet.icsi.berkeley.edu/fndrupal/luIndex on December 1, 2019.

Esra Abdelzaher & Ágoston Tóth:
Defining Crime: A multifaceted approach based on Lexicographic Relevance and Distributional Semantics
Argumentum 16 (2020), 44-63
Debreceni Egyetemi Kiadó
DOI: 10.34103/ARGUMENTUM/2020/4

R. It's not a *crime* to have an obsession[Situation]
S. A healthy, adoptable animal[Experiencer] killed for the sole *crime* of being homeless[Situation].

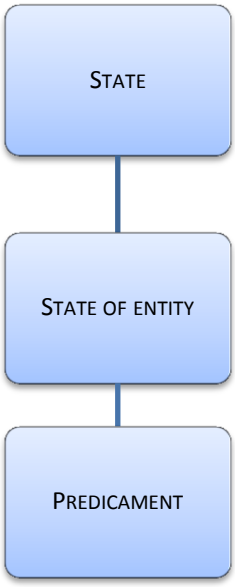

Figure 4. The STATE*-hierarchy of* crime*-related frames*

Challengingly, FN places several lexical units that do not indicate inherent features of entities – and are more related to social or economic states – within the ATTRIBUTES hierarchy, including *rich*, *active*, *ill* and *certain*. Furthermore, several potential FE fillers in frames related to the 'undesirable state' sense of crime are missing (e.g. *homeless*) or are not linked to the relevant frames (e.g. *unemployed*).

Figure 5 represents an analytical hierarchical process of selecting the right sense of *crime* according to four lexicographic and distributional criteria.

Esra Abdelzaher & Ágoston Tóth:
Defining Crime: A multifaceted approach based on Lexicographic Relevance and Distributional Semantics
Argumentum 16 (2020), 44-63
Debreceni Egyetemi Kiadó
DOI: 10.34103/ARGUMENTUM/2020/4

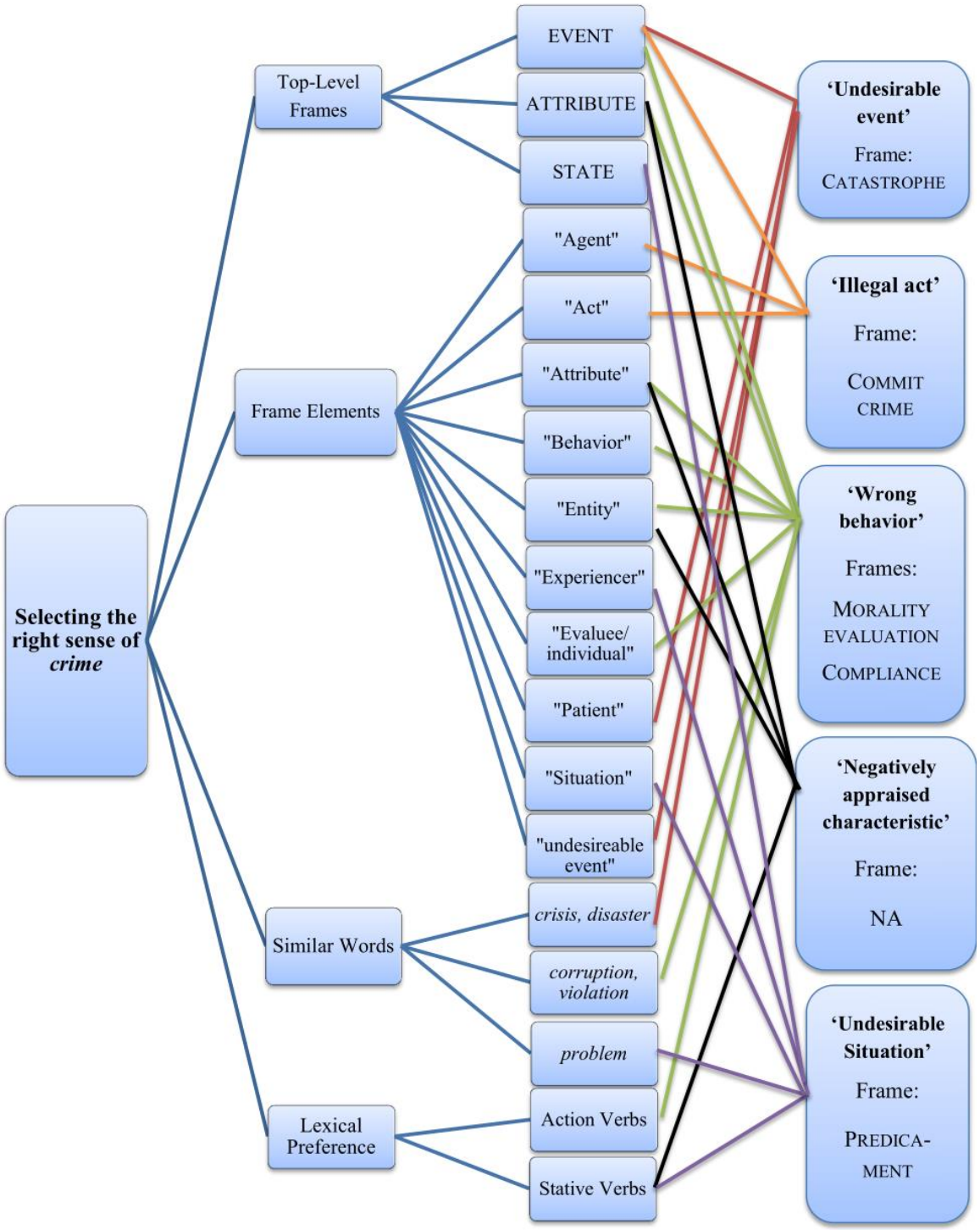

Figure 5. Decision tree for finding the right sense of crime

Esra Abdelzaher & Ágoston Tóth:
Defining Crime: A multifaceted approach based on Lexicographic Relevance and Distributional Semantics
Argumentum 16 (2020), 44-63
Debreceni Egyetemi Kiadó
DOI: 10.34103/ARGUMENTUM/2020/4

5 Discussion

Our study argues that *crime* has five senses: (a) 'illegal act', which is the only meaning explicitly mentioned in FN, (b) 'wrong behavior', which is partially reflected in the parent frame MISDEED, (c) 'undesirable event', which is not related to *crime* in FN, (d) 'undesirable situation' and (e) 'negatively appraised characteristic'. The multifaceted approach that we use defines four criteria to select the right sense of *crime* within the framework of lexicographic relevance and through the assistance of distributional methods, which is a novelty in the relevant literature.

The proposed approach can also be implemented in FN without affecting the precision of the manual lexicographical effort. We believe that it can speed up the process of enriching the database with more lexical units assigned to a given frame or set of frames. The Sketch Engine toolkit can suggest potential co-lexical units, portray the lexico-grammatical similarities and differences between them and give examples of their actual uses in several corpora. To date, FN's lexicographers have acknowledged the effectiveness of using word sketches (Atkins, Rundell & Sato 2003: 335), but they do not offer statistical information in the database about lexical units (Kwiatek 2013: 12), which is an issue that we cannot directly address in this paper, but it illustrates the paucity of interest in quantitative information on the side of FrameNet's editors.

A comparison of the uses of *crime* identified in this paper with the senses of this headword in online dictionaries displays some missing senses, too. Lexico.com by Oxford University Press defines *crime* as "[a]n action or omission which constitutes an offence and is punishable by law" (sense 1, OUP) which is consistent with FN's definition ("an action constituting a serious offence against an individual or the state and is punishable by law"[11]). Moreover, the provided example "shoplifting was a serious crime" is compatible with THEFT (the lexical unit *shoplifting.n* is listed in that frame), which is an inheritor frame of COMMITTING_CRIME. Also, *crime* in the same dictionary has the meaning of "illegal activities" collectively (sense 1.1, OUP). This dictionary also refers to the more general sense of *crime*, which is "[a]n action or activity considered to be evil, shameful, or wrong" (sense 1.2, OUP). Although this sense is missing from FN, it was detected by our proposed approach (examples D, E, F and G). The "illegal" use of *crime* is also included in the online Merriam-Webster Dictionary (sense 1, MW) and Collins Online Dictionary (sense 1, COLLINS). The other sense of "wrong activity" is also reflected, but it is defined as "something reprehensible, foolish or disgraceful" and "something […] is very wrong or a serious mistake" in Merriam-Webster (sense 4, MW) and Collins Online Dictionary (sense 2, COLLINS).

The three dictionaries conventionally use the words "act", "action" and "activity" to define *crime*, which places these senses in the EVENT hierarchy. According to the present study, EVENT includes three frames evoked by *crime*, and two of them ('illegal act' and 'wrong behavior') are reflected in the dictionaries. However, the sense of *crime* as 'undesirable event' (evoking the frame of CATASTROPHE and concentrating on the "patient" and "undesirable event" FEs) is absent from the OUP and MW dictionaries. COLLINS refers to the informal use of *crime* as "something to be regretted" (sense 4, *crime* in British English, COLLINS) and provides "it is a crime that he died young" (ibid.) as an illustrative example. This can be interpreted as consistent with the proposed sense of *crime* as an 'undesirable event'. However,

[11] From the lexical entry for *crime.n* via https://framenet.icsi.berkeley.edu/fndrupal/luIndex, which has been retrieved on December 1, 2019.

Esra Abdelzaher & Ágoston Tóth:
Defining Crime: A multifaceted approach based on Lexicographic Relevance and Distributional Semantics
Argumentum 16 (2020), 44-63
Debreceni Egyetemi Kiadó
DOI: 10.34103/ARGUMENTUM/2020/4

the parallel example "it's a crime you didn't finish school" (sense 4, *crime* in American English, COLLINS) which is used for elaborating the same sense "something regrettable" (ibid.) illustrates that the distinction between the 'wrong behavior' and the 'undesirable event' senses is blurred in the Collins Dictionary.

The sense of *crime* that can be described as a 'negatively appraised characteristic', which belongs to the ATTRIBUTES hierarchy, is not reflected in the dictionaries or in FN. The corpus-based examples (N) and (O) are not covered by the definitions of *crime* provided by the abovementioned dictionaries. In terms of top-level FN frames, dictionaries refer to negative acts, not attributes. At the denotative level, the negativity of the appraisal in the dictionaries is attributed to the act itself regardless of the context of evaluation. However, the negatively appraised characteristics represented by examples (N) and (O) would be inherently neutral, they acquire the negative evaluation because of a certain unconventional context.

Similarly, the sense of *crime* related to the STATE top-level FN frame (*crime* as an 'undesirable situation') is missing from FN and dictionaries. In section 4.3, we annotated references to being "unemployed", "in Mosul" and "homeless" as crimes in this sense ('undesirable situation'). These instances are not covered by any of the definitions included in dictionaries. They do not refer to a mistake or a foolish behavior committed by an "agent". They do not denote illegal acts committed by a "perpetrator" against a "victim". Instead, they describe a state in which the "patient" suffers from an "undesirable situation".

6 Conclusion

The comparison of word similarity lists compiled from FN data and lists extracted from corpora (using the distributional thesaurus function), as discussed in section 4, has reflected a considerable gap between judgments made in FN and corpus-based statistical findings. To identify new senses of the word *crime*, we have followed a qualitative, lexicographic approach to explore the automatically identified words that were similar to our target word based on second-order co-occurrence information.

Through this case study, we have tried to demonstrate that distributional similarity can point to areas worthy of lexicographic investigation. The distributional methods can help lexicographers reveal significant features of words, which may lead to the discovery of new senses.

The multifaceted approach that we suggest has been effective in the identification of new senses of *crime* that are not present in the FN database or in dictionaries. This has enabled us to propose four criteria, based on FN's existing data and distributional similarity, to differentiate between five senses of *crime* systematically.

Dictionaries

OUP = Oxford University Press (2019): *Lexico.com*. Available at http://www.lexico.com

MW = Merriam Webster (2019): *Merriam-Webster Dictionary*. Available at http://www.m-w.com

COLLINS = Collins (2019): *Collins Online Dictionary*. Available at https://www.collinsdictionary.com

References

Atkins, S. & Bouillon, P. (2006): Relevance in dictionary making: Sense indicators in the bilingual entry. In: Bowker, L. (ed.): *Lexicography, terminology, and translation: Text-based studies in honour of Ingrid Meyer*. Ottawa: University of Ottawa Press, 25–43, https://doi.org/10.2307/j.ctt1ckpgs3.6.

Atkins, S., Fillmore, C.J. & Johnson, C.R. (2003): Lexicographic relevance: Selecting information from corpus evidence. *International Journal of Lexicography* 16.3, 251–280, https://doi.org/10.1093/ijl/16.3.251.

Atkins, S., Rundell, M. & Sato, H. (2003): The contribution of FrameNet to practical lexicography. *International Journal of Lexicography* 16.3, 333–357, https://doi.org/10.1093/ijl/16.3.333.

Baker, C. F., & Sato, H. (2003): The FrameNet data and software. In Matsumoto, Y. (ed.): *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics – Volume 2*. Stroudsburg, PA: Association for Computational Linguistics, 161–164, https://doi.org/10.3115/1075178.1075206.

Dodge, E., Hong, J. & Stickles, E. (2015): MetaNet: Deep semantic automatic metaphor analysis. In: Shutova, E., Klebanov, B.B. & Lichtenstein, P. (eds.): *Proceedings of the Third Workshop on Metaphor in NLP*. Denver, CA: Associaton for Computational Linguistics, 40–49, https://doi.org/10.3115/v1/w15-1405.

Fillmore, C.J. (1968): The Case for Case. In: Bach, E. & Harms, R.T. (eds.): *Universals in Linguistic Theory*. New York, NY: Holt, Rinehart and Winston, 1–88.

Fillmore, C.J., Wooters, C. & Baker, C.F. (2001): Building a large lexical databank which provides deep semantics. In: T'sou, B.K., Kwong, O. & Lai, T. (eds.): *Proceedings of the 15th Pacific Asia Conference on Language, Information and Computation*. Hong Kong: City University of Hong Kong, 3–26.

Firth, J.R. (1957): A synopsis of linguistic theory, 1930-1955. In: Firth, J.R. (ed.): *Studies in linguistic analysis*. Oxford: Basil Blackwell, 1–32.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P. & Suchomel, V. (2013): The TenTen Corpus Family. In: Hardie, A. & Love, R. (eds.): *Corpus Linguistics 2013 Abstract Book*. Lancaster: UCREL, Lancaster University, 125–127.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014): The Sketch Engine: ten years on. *Lexicography* 1.1, 7–36, http://doi.org/10.1007/s40607-014-0009-9.

Kwiatek, E. (2013): *Contrastive analysis of English and Polish surveying terminology*. Newcastle: Cambridge Scholars Publishing.

Leech, G.N. (1992): 100 million words of English: the British National Corpus (BNC). *Language Research* 28.1, 1–13.

Lenci, A. (2008): Distributional semantics in linguistic and cognitive research. *Italian journal of linguistics* 20.1, 1–31.

Ruppenhofer, J., Ellsworth, M., Petruck, M.R., Johnson, C.R. & Scheffczyk, J. (2016): FrameNet II: Extended theory and practice. Retrieved from https://framenet2.icsi.berkeley.edu/docs/r1.7/book.pdf on December 1, 2019.

Rychlý, P. & Kilgarriff, A. (2007): An efficient algorithm for building a distributional thesaurus (and other Sketch Engine developments). In: Ananiadou, S. (ed.): *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*. Stroudsburg, PA: Association for Computational Linguistics, 41–44, http://doi.org/10.3115/1557769.1557783.

Tóth Á. (2014): *The Company that Words Keep: Distributional Semantics*. Debrecen: Debrecen University Press.

Trampuš, M. & Novak, B. (2012): Internals of an aggregated web news feed. In: Bellatreche, L. & Mohania, M. (eds.): *Proceedings of the Conference on Data Mining and Data Warehouses (SiKDD 2012) co-located with the 15th International Multiconference on Information Society*, 221–224. Retrieved from https://ailab.ijs.si/dunja/SiKDD2012/Papers/Trampus_Newsfeed.pdf on December 1, 2019.

Esra Abdelzaher
University of Debrecen, Doctoral School of Linguistics
University of Debrecen, Institute of English and American Studies
Pf. 400
H-4002 Debrecen
lingcorpus2@gmail.com

Dr. Ágoston Tóth
University of Debrecen
Institute of English and American Studies
Department of English Linguistics
Pf. 400
H-4002 Debrecen
toth.agoston@arts.unideb.hu

Esra Abdelzaher & Ágoston Tóth:
Defining Crime: A multifaceted approach based on Lexicographic Relevance and Distributional Semantics
Argumentum 16 (2020), 44-63
Debreceni Egyetemi Kiadó
DOI: 10.34103/ARGUMENTUM/2020/4

Appendix 1: Lexical units and frames linked to the COMMITTING_CRIME frame

Lexical Unit	Frame	Lexical Unit	Frame	Lexical Unit	Frame
abduct	KIDNAPPING	*hold up*	ROBBERY	*purloin*	THEFT
abducted	KIDNAPPING	*hold-up*	ROBBERY	*ransack*	ROBBERY
abduction	KIDNAPPING	*homicide*	OFFENSES	*rape*	RAPE
abductor	KIDNAPPING	*illegal*	LEGALITY	*rape*	OFFENSES
abscond (with)	THEFT	*illicit*	LEGALITY	*raped*	RAPE
abstract	THEFT	*indecent assault*	OFFENSES	*rapist*	RAPE
abstraction	THEFT	*inquire*	CRIMINAL INVESTIGATION	*rifle*	ROBBERY
abuse	ABUSING	*inquiry*	CRIMINAL INVESTIGATION	*rob*	ROBBERY
abusive	ABUSING	*investigate*	CRIMINAL INVESTIGATION	*rob blind*	ROBBERY
apprehend	ARREST	*investigation*	CRIMINAL INVESTIGATION	*robber*	ROBBERY
apprehension	ARREST	*kidnap*	KIDNAPPING	*robbery*	ROBBERY
arraign	ARRAIGNMENT	*kidnapped*	KIDNAPPING	*robbery*	OFFENSES
arraignment	ARRAIGNMENT	*kidnapper*	KIDNAPPING	*rustle*	THEFT
arrest	ARREST	*kidnapping*	KIDNAPPING	*sabotage*	OFFENSES
arson	OFFENSES	*kidnapping*	OFFENSES	*send up*	SENTENCING
assault	OFFENSES	*larceny*	THEFT	*sentence*	SENTENCING
bag	THEFT	*larceny*	OFFENSES	*sexual assault*	OFFENSES
batter	ABUSING	*lawful*	LEGALITY	*sexual harassment*	OFFENSES
battery	OFFENSES	*lead*	CRIMINAL INVESTIGATION	*sexually assault*	RAPE
book	ARREST	*legal*	LEGALITY	*shanghai*	KIDNAPPING
burglary	OFFENSES	*legitimate*	LEGALITY	*shoplift*	THEFT
bust	ARREST	*licit*	LEGALITY	*shoplifter*	THEFT
carjack	PIRACY	*lift*	THEFT	*shoplifting*	THEFT
carjacking	PIRACY	*light-fingered*	THEFT	*sin*	MISDEED
case	CRIMINAL INVESTIGATION	*make off (with)*	THEFT	*smuggle*	SMUGGLING
case	TRIAL	*maltreat*	ABUSING	*smuggler*	SMUGGLING
child abuse	OFFENSES	*maltreatment*	ABUSING	*smuggling*	SMUGGLING
clue	CRIMINAL INVESTIGATION	*manslaughter*	OFFENSES	*snatch*	KIDNAPPING
collar	ARREST	*misappropriate*	THEFT	*snatch*	THEFT
commission	COMMITTING CRIME	*misappropriation*	THEFT	*snatcher*	KIDNAPPING

Esra Abdelzaher & Ágoston Tóth:
Defining Crime: A multifaceted approach based on Lexicographic Relevance and Distributional Semantics
Argumentum 16 (2020), 44-63
Debreceni Egyetemi Kiadó
DOI: 10.34103/ARGUMENTUM/2020/4

commit	COMMITTING CRIME	*misdeed*	MISDEED	*snatcher*	THEFT
condemn	SENTENCING	*mug*	ROBBERY	*snitch*	THEFT
conspiracy	OFFENSES	*mugger*	ROBBERY	*statutory rape*	OFFENSES
contraband	SMUGGLING	*mugging*	ROBBERY	*steal*	THEFT
cop	THEFT	*murder*	OFFENSES	*stealer*	THEFT
cop	ARREST	*nab*	KIDNAPPING	*stealing*	THEFT
copyright infringement	OFFENSES	*nab*	ARREST	*stick up*	ROBBERY
crime	COMMITTING CRIME	*negligence*	OFFENSES	*stick-up*	ROBBERY
criminal	LEGALITY	*nick*	THEFT	*stolen*	THEFT
cutpurse	THEFT	*order*	SENTENCING	*summons*	ARREST
domestic violence	ABUSING	*peccadillo*	MISDEED	*swipe*	THEFT
embezzle	THEFT	*peculation*	THEFT	*theft*	THEFT
embezzlement	THEFT	*permissible*	LEGALITY	*theft*	OFFENSES
embezzler	THEFT	*perpetrate*	COMMITTING CRIME	*thief*	THEFT
fair	LEGALITY	*pickpocket*	THEFT	*thieve*	THEFT
felony	OFFENSES	*pilfer*	THEFT	*thieving*	THEFT
filch	THEFT	*pilferage*	THEFT	*transgress*	MISDEED
flog	THEFT	*pilferer*	THEFT	*transgres-sion*	MISDEED
fraud	OFFENSES	*pilfering*	THEFT	*treason*	OFFENSES
heist	THEFT	*pinch*	THEFT	*trial*	TRIAL
hijack	PIRACY	*piracy*	PIRACY	*unlawful*	LEGALITY
hijacked	PIRACY	*pirate*	PIRACY	*wrong*	LEGALITY
hijacker	PIRACY	*possession*	OFFENSES	*wrongful*	LEGALITY
hijacking	PIRACY	*probe*	CRIMINAL INVESTIGATION	*wrongly*	LEGALITY
hijacking	OFFENSES	*prohibited*	LEGALITY		

Esra Abdelzaher & Ágoston Tóth:
Defining Crime: A multifaceted approach based on Lexicographic Relevance and Distributional Semantics
Argumentum 16 (2020), 44-63
Debreceni Egyetemi Kiadó
DOI: 10.34103/ARGUMENTUM/2020/4

Appendix 2: Words similar to *crime* in selected corpora of the Sketch Engine

Similar Words	Similarity Score			Similar Words	Similarity Score		
	Timestamped	*TenTen*	*BNC*		*Timestamped*	*TenTen*	*BNC*
abuse	0.521	0.504	0.247	*politics*	0.354	NA	NA
violence	0.516	0.516	0.270	*disaster*	0.352	NA	NA
violation	0.458	0.432	NA	*fraud*	0.350	0.378	NA
corruption	0.429	0.413	NA	*killing*	0.337	NA	NA
terrorism	0.416	0.399	NA	*activity*	NA	0.398	0.205
murder	0.405	0.448	0.238	*offence*	NA	0.394	0.304
incident	0.403	0.434	0.206	*death*	NA	0.379	NA
threat	0.398	0.402	NA	*accident*	NA	NA	0.206
poverty	0.393	NA	0.188	*disease*	NA	NA	0.203
attack	0.380	0.403	0.183	*drug*	NA	NA	0.194
act	0.378	0.409	NA	*action*	NA	NA	0.187
situation	0.376	0.386	0.206	*health*	NA	NA	0.195
conflict	0.375	NA	0.192	*problem*	NA	0.378	0.186
failure	0.373	NA	0.186	*issue*	NA	NA	0.185
behavior	0.372	0.384	NA	*practice*	NA	NA	0.195
crisis	0.367	NA	0.184	*case*	NA	0.388	NA
assault	0.359	0.444	NA				