CONCRETE

CONCEPT CREATION TECHNOLOGY
PROJECT NUMBER: 611733

SMALL OR MEDIUM-SCALE FOCUSED RESEARCH PROJECT
ICT - FUTURE AND EMERGING TECHNOLOGIES (FET)

Deliverable 6.3:
Final report on framework and data

Jožef Stefan Institute (JSI) Version: 1.0, final

Executive summary
In this document we present the data and data-related services that have been developed in ConCreTe since the initial report on these resources in Deliverable D6.2. We present the datasets that were created and made available by project partners, based on the needs of the research work of the project. These are the various datasets of concept associations and the datasets on the literature and vocabulary of the research field of computational creativity. Besides datasets, we also report on the newly developed data access and data service components, particularly on the triplet querying tool that provides access to triplets generated from relevant freely available resources and offers three different interfaces for its use.

Dissemination Level		
PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	-
RE	Restricted to a group specified by the Consortium (including the Commission Services)	-
CO	Confidential, only for members of the Consortium (including the Commission Services)	-

Revision history

Document administrative information	
Project acronym:	ConCreTe
Project number	611733
Deliverable number:	D6.3
Deliverable full title:	Final report on framework and data
Deliverable short title D6.3:	Final report on framework and data
Document identifier:	ConCreTe-del-D6.3-FinalReportData-final-v1.0
Lead partner short name:	Jožef Stefan Institute (JSI)
Report version:	1.0, final
Report preparation date:	31/03/2016
Dissemination level:	PU
Nature:	R = report
Lead author:	Martin Žnidaršič
Co-authors:	Senja Pollak, Janez Kranjc, Vid Podpečan, Borut Lesjak, Nada Lavrač
Status:	Final

Changes to this document are detailed in the change log table below.

Change log

Date	Editor	Summary of changes made
01/03/2016	Martin Žnidaršič, all	Sectioning, Datasets
08/03/2016	Senja Pollak	ICCC datasets
01/03/2016	Borut Lesjak, Vid Podpečan	Triplet querying service description
10/03/2016	Vid Podpečan	Web application and widget for the triplet querying service
21/03/2016	Martin Žnidaršič	Associations dataset
29/03/2016	Sue White	Proof reading and language corrections
29/03/2016	Nada Lavrač	Quality control
31/03/2016	Martin Žnidaršič	Final corrections

Contents

1 Introduction

In this document we report on the datasets and data-related services that were developed in scope of task T6.2 of the project ConCreTe. Preliminary work on this topic was described in the Deliverable D6.2 in month 12. The need for common datasets was not as high as expected at the start of the project. On the other hand, there was interest for data access and querying services, which were initially not explicitly planned.

The datasets that were prepared and made available since the last report are described in Section 2. These are the datasets of concept associations and the datasets on the literature and vocabulary of the research field of computational creativity.

The newly developed data access and querying services are described in Section 3 where we report on the associations access widget and the triplet querying tool that provides access to triplets generated from relevant freely available resources.

2 Datasets

In the first report on data and data services (D6.2) we presented a collection of exemplary datasets which were mostly prepared for demonstrational purposes. These were the (I) Cocktail recipes dataset, (II) Pancake recipes dataset, (III) Aesop's fables morals dataset, (IV) English proverbs dataset and (V) Single sentences dataset with sentences from three distinctly different domains.

The research work in the project, however, had to be supported by other datasets. These are the datasets on associations, which were studied in the scope of work-packages WP4 and WP8, and used in some applications of WP5. Moreover, the datasets of computational creativity bibliography and vocabulary that were used in activities of WP9.

2.1 EAT associations dataset

A dataset of associations among concepts was added to ConCreTeFlows and an accompanying software component (widget) was implemented, which enables its use in this platform (see Section 3.1). The dataset is a JSON format version[1] of the Edinburgh Associative Thesaurus[2]. The dataset provides a list of associations for 8400 words.

2.2 Card sorting study dataset

Inspired by the work in the project's workpackage on evaluation (WP8), a card sorting (CST) study was conducted in which human subjects grouped together associated concepts, which were represented with words printed on cards that were used for grouping. Experiments of this kind are very valuable, but also costly and time consuming.

This data was used to derive a metric of relations between the words and investigate some properties of conceptual blending (e.g., whether the blending terms violate the triangle inequality of the metric).

The dataset is in the form of a spreadsheet that represents a matrix in which we have a row and a column for every word that was used in the experiment. Each cell holds the number of times the two words (from its row and its column) were listed in the same group. The dataset, which is currently only available to project partners, is available at:
`http://concreteflows.ijs.si/static/creativity/CCdata/ConCreTe_CardSortingData.zip`

[1] As prepared by Darius Kazemi (`https://github.com/dariusk/ea-thesaurus`)

[2] `http://www.eat.rl.ac.uk/`

Table 1: Terms from EAT dataset that were removed, in addition to the stopwords and numbers.

abcess	aquaintance	aweful	batchelors	beguilingly	cadbury's	candlestickmaker
cannot	casterbridge	catastrophy	charlady	cornflake	cornflakes	couldn't
crawly	detatched	didn't	doctor's	doesn't	don't	dumbells
evington	flutterby	fry's	gazers	handsom	hasn't	humph
isn't	laundrette	lente	limy	matelot	mother's	negress
noisesome	pillowslip	shouldn't	smartie	sssh	stroganoff	uncomfy
unladen	what's					

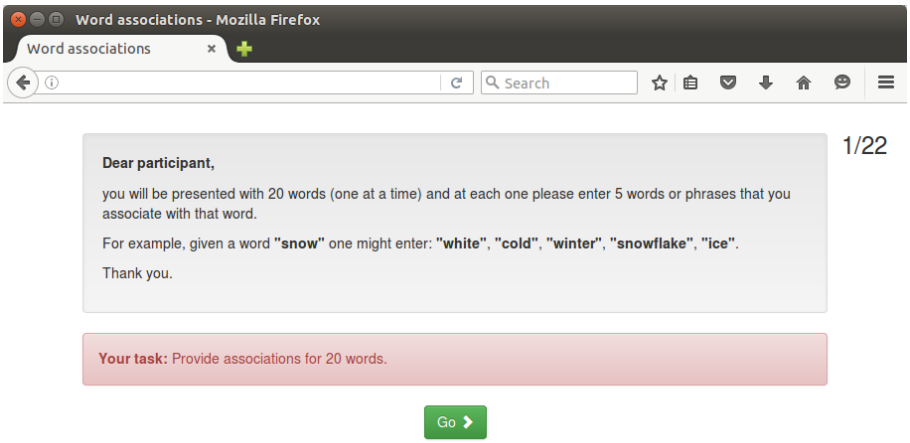

Figure 1: Introductory screen of the associations questionnaire.

2.3 ConCreTe associations dataset

As the dataset of associations from Section 2.1 is rather small and old (it was prepared in 1972), we have designed a crowdsourcing experiment to obtain an up-to-date dataset that would also cover more concepts.

We have selected 13,333 terms for our associations experiment, based on these resources:

1. 48 terms that were used in the card sorting experiment (see Section 2.2),
2. 18 trigger words (representing animals) that correspond to the input spaces in a study about perception of blends [2] which was reported in Section 5.1 of Deliverable D3.1,
3. 7,995 words from EAT dataset (see Section 2.1) from which we removed numbers, stopwords[3], the 66 terms from items 1 and 2 above, and the terms that appear neither in WordNet nor in the Google n-grams corpus (these are listed in Table 1),
4. 5,272 words with the most frequent appearance from the Google n-grams dataset (provided at: `http://norvig.com/ngrams/`), without duplicating the ones that had already been added from the resources in items 1-3.

[3]The *Default English stopwords list* was used from `http://www.ranks.nl/stopwords`

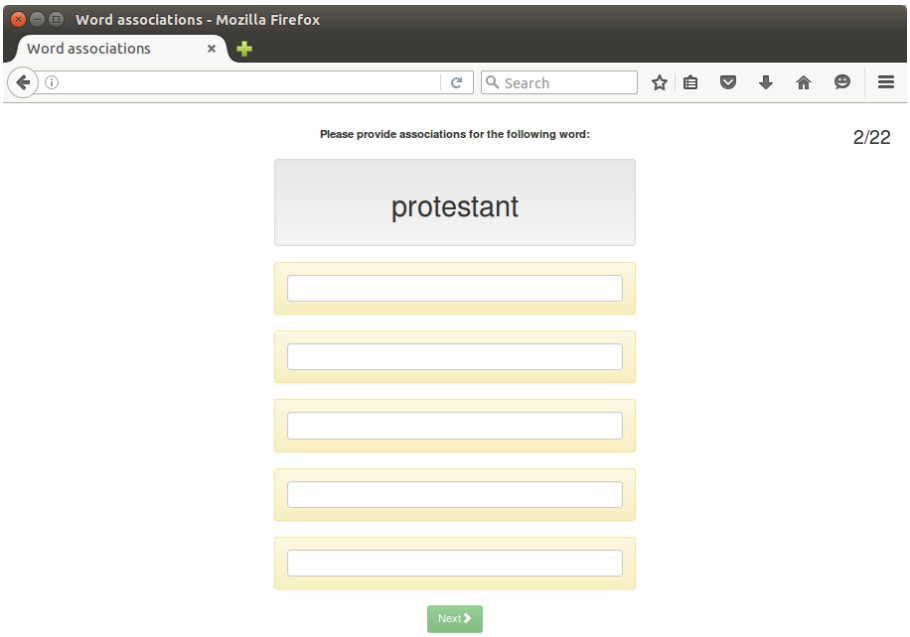

Figure 2: The data collection interface of the associations questionnaire.

The resulting 13,333 terms were used in a crowdsourcing experiment, in which a questionnaire with 20 terms was presented to each participant. The questionnaire consisted of the introductory page with instructions (see Figure 1) and a page for gathering 5 associations for every of the 20 terms, as shown in Figure 2.

There were 2,000 questionnaires deployed, each containing 20 terms. This yields exactly 3 sets of annotations for each of the 13,333 terms. Each questionnaire was answered by a different user of the crowdsourcing platform[4]. We opted for only the most reliable crowdworkers (highest reliability level that is offered on the platform) and used a simple input verification: entries shorter than two characters were not accepted.

Raw unfiltered data from this experiment is available to project partners here:
`http://concreteflows.ijs.si/static/creativity/CCdata/associations2016.zip`

2.4 ICCC datasets

Here we report on the dataset of the ICCC proceedings as well as its terminology and ontologies, which were extracted in our work that was submitted to the ICCC2016 (see the paper in Appendix A).

2.4.1 The ICCC dataset

The proceedings of International Conferences of Computational Creativity (ICCC) are available in PDF format. We transformed the proceedings from 2010 to 2015 to raw text format. In total, the dataset contains 247 articles from 6 consecutive proceedings, containing the following numbers of articles: ICCC-2010: 43, ICCC-2011: 30, ICCC-2012: 44, ICCC-2013: 40, ICCC-2014: 49, ICCC-2015: 41.

In addition, we also created a version without the references section. This dataset is available in named line document format that we used for creating the topic ontologies described in this section. Each article (without the reference section) represents one line, starting with article ID,

[4]We used CrowdFlower: `http://www.crowdflower.com/`

followed by the exclamation mark and the year of conference edition (2010, 2011, 2012, 2013, 2014 or 2015).

2.4.2 Computational creativity terminology

The goal of terminology extraction is to automatically extract relevant terms for a given domain, represented by a given corpus. We used terminology extraction method LUIZ-CF [3], a modified version of the LUIZ term recognition tool [4].

Term extraction consists of two steps: extracting the noun phrase candidates based on morphosyntactic patterns; followed by weighting and ranking of the candidates based on their 'termhood' value, for single word and multi-word terms. The termhood value was computed based on the comparison of relative frequencies of lemmas of a term in the domain corpus (here, the ICCC proceedings) compared to a reference corpus: for English, the frequencies of the British National Corpus were used. The extracted terms were ranked by termhood value on a scale between 1 and 0. In addition to default stopwords, we eliminated also the names of ICCC PC members, leading to the exclusion of some of the paper authors from the term list. The top ranked candidates are available here:

- top 1,500 terms (single and multiword terms):
  `http://kt.ijs.si/senja_pollak/CC_resources/terminology/terminology1500.txt`
- top 1,500 multiword terms -the first 15 are presented in Table 2:
  `http://kt.ijs.si/senja_pollak/CC_resources/terminology/terminology1500MULTI.txt`

Table 2: Top 15 multi-word terms from the ICCC corpus.

Score	Term
1.000000	[computational creativity]
0.247012	[creative system]
0.102306	[creative process]
0.099844	[conceptual space]
0.030894	[computational model]
0.021593	[computational system]
0.018712	[fitness function]
0.018064	[jaguar knight]
0.012638	[genetic algorithm]
0.012078	[human creativity]
0.011300	[poetry generation]
0.011171	[story generation]
0.010011	[neural network]
0.009657	[creative domain]
0.009299	[transformational creativity]

2.4.3 The created CC topic ontologies

A tool named OntoGen[5] was used to build a topic ontology for CC domain conceptualization. OntoGen is a semi-automatic and data-driven ontology editor. Semi-automatic means that the system is an interactive tool that aids the user during the topic ontology construction process. Data-driven means that most of the aid provided by the system is based on the underlying text data (document corpus) provided by the user. The system combines text mining techniques with an efficient user interface. The OntoGen platform is illustrated in Figure 3.

We constructed three topic ontologies:

[5] `http://ontogen.ijs.si/`[1]

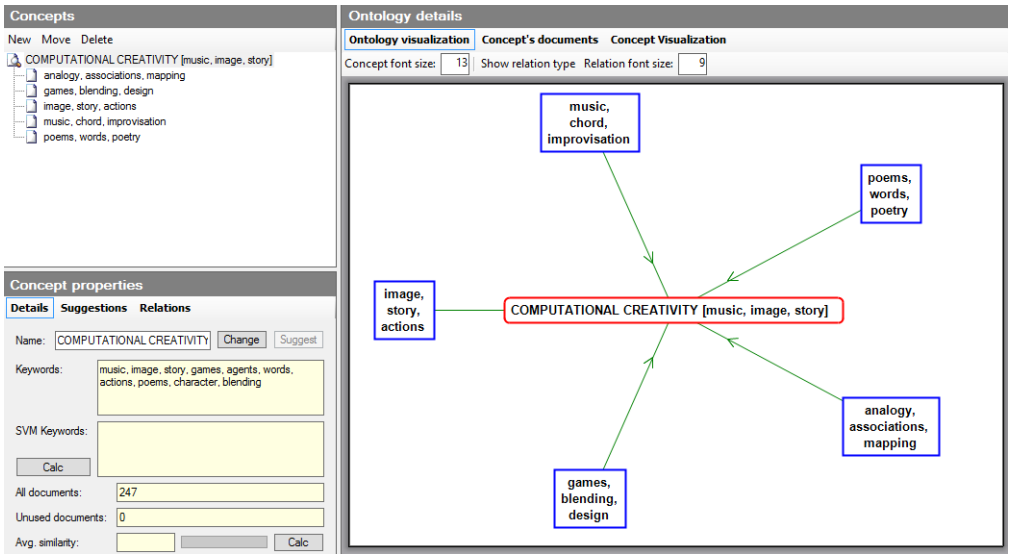

Figure 3: The OntoGen interface, showing the automatically generated conceptualization of the CC domain and the keywords describing the top-level concept.

- Automatically constructed topic ontology of CC domain (cf. Figure 3)

  The proposed topic ontology was obtained by k-means clustering (for k set to 5). The first three automatically extracted keywords are extracted from document centroid vectors.

- Semi-automatically constructed ontology, obtained by minor human effort invested: automated clustering, manual cluster names, manual selection of topic descriptors out of list of automatically proposed keywords (cf. Figure 4).

  By inspecting the keywords of each cluster, we manually named the concepts of the automatically generated topic ontology as follows: “Musical creativity”, “Visual creativity”, “Linguistic creativity”, “Creativity in Games” and “Conceptual creativity”. Out of a set of ten automatically generated keywords characterizing each of the five document clusters, we manually selected three keywords that we believe best describe the cluster of papers belonging to the given concept. These keywords were added to the concept labels.

- Semi-automatically constructed ontology, using various OntoGen functionalities for ontology manipulation (cf. Figure 5)

  This improved computational creativity conceptualization was created by manipulating the initial ontology using different OntoGen functionalities. The main concepts obtained were further divided and when forming meaningful concepts, the categories were added as sub-concepts (see e.g. the sub-concepts of Linguistic creativity). In addition, some (sub)concepts were moved, e.g., the “narrative” category was moved from Visual to Lexical creativity. Some concepts were added by query and active learning functionality of OntoGen (e.g., concept Evaluation). On the first level we distinguish between Musical, Visual, Linguistic creativity, Games and creativity, Conceptual creativity as well as the newly created category of Evaluation. On lower levels, we added e.g. Narratives, Poetry, Recipes and Lexical creativity for Linguistic creativity, where the latter comprises e.g., humour, neologisms, etc. Each concept is represented by descriptive keywords (cf. keywords for six first level concepts shown in Table 3) out of which we selected three keywords (in italics) to represent the concept in the visually presented topic ontology (Figure 5).

The generated topic ontologies are available in .png and .rdf formats. All the resources are available at:
`http://kt.ijs.si/senja_pollak/CC_resources/`.

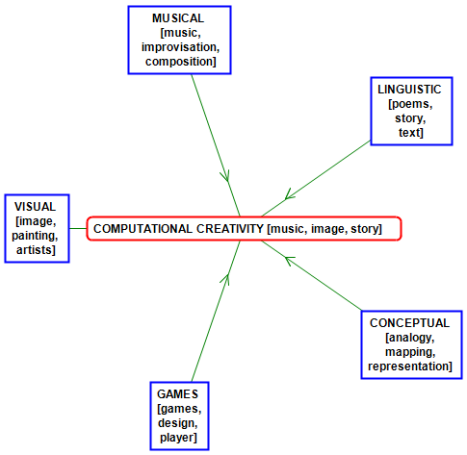

Figure 4: CC conceptualization through automated clustering, concept naming and manual keywords selection.

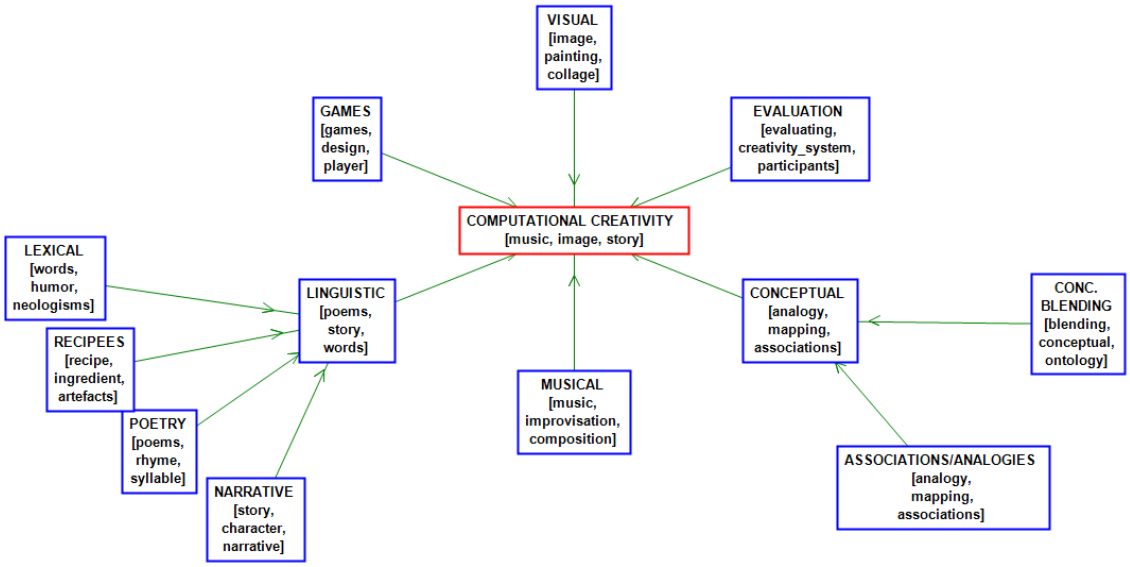

Figure 5: Semi-automatically generated conceptualization of the CC domain, with concept naming and subconcept creation.

Table 3: Categories and keywords of the first layer of the semi-automatically constructed CC topic ontology.

Category	Automatically extracted keywords
Musical	*music*, chord, *improvisation*, melodies, harmonize, *composition*, accompaniment, pitch, emotions, beat
Visual	*image*, *painting*, darci, artifacts, *collage*, adjectives, associations, rendered, colored, artists
Linguistic	*story*, *poems*, actions, character, *words*, agents, narrative, artefacts, poetry, evaluating
Games	*games*, *design*, *player*, games_design, angelina, agents, code, jam, filter, gameplay
Conceptual	*analogy*, *blending*, mapping, conceptual, objective, *associations*, team, graphs, concepts, domain
Evaluation	music, poems, improvisation, *evaluating*, interactive, poetry, *creativity system*, musician, *participants*, behavioural
Comp. creativity	*music*, *image*, *story*, games, agents, words, actions, poems, character, blending

3 Data acquisition and manipulation services

In this section we describe the software components that provide access to relevant data resources directly from the common software platform of the project.

3.1 Associations access widget

The access to the dataset from Section 2.1 is enabled through a ConCreTeFlows widget named Associations. In Figure 6, the widget is shown in a minimal workflow, illustrated with association results for the word *zebra*. The widget provides a ConCreTeFlows interface and some basic text transformations on this dataset.

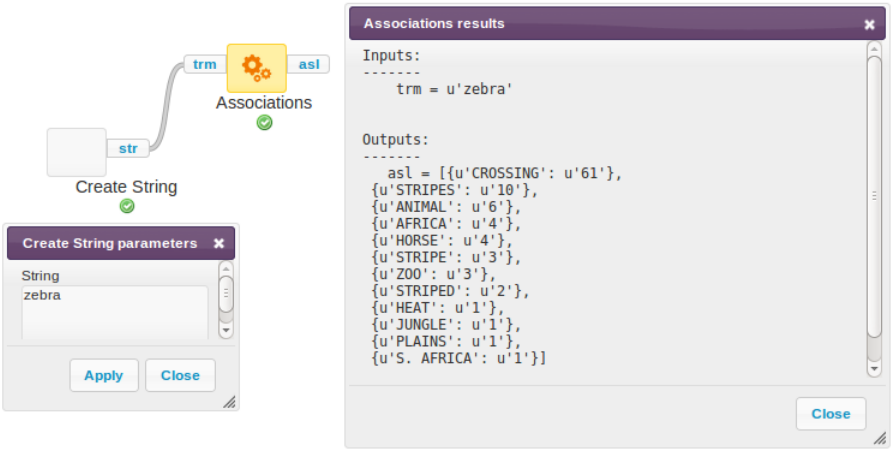

Figure 6: Associations access widget in ConCreTeFlows.

3.2 Triplet querying service

Textual triplets are used in many ConcreTeFlows components. To simplify and enrich the access to triplets we have gathered and unified data from four of the largest structured free online data sources: WordNet, YAGO, NELL, and ConceptNet.

We parsed each of the four data sources separately and cleaned the data appropriately, to be able to unify it in the form of relation-entity-value triplets. An example of a (YAGO) triplet is listed below:

```
<actedIn>, Robert De Niro, Raging Bull
```

We define and use a set of common relation IDs, to which we map all the retrospective data source relation IDs. The entity (the left, or first, concept) and the value (the right, or second, concept) are not unified, and are left intact, other than that they are thoroughly syntactically cleaned.

Using Google Books n-grams (specifically, 1-grams), we calculate a score of typicality T for each triplet, with the following simple formula:

```
T = sum( log( p_1gram( word ))) / #words + log( p_1gram( min_freq_word ))
```

Basically, we work with logs of word probabilities to prevent underflows when multiplying. The function p_1gram returns the distributional probability of a word (its frequency divided by the total number of words in the corpus). #words is the number of all the words together in the triplet's entity and value (only those words are considered), so when the sum of logs is divided by this number, we actually find a root of the probability product. To this we then add the log of probability

Table 4: Frequencies of relations, triplets and duplicates by data source.

data source	WordNet	YAGO	NELL	ConceptNet	TOTAL
#relations	12	72	427	56	567
#triplets	1273157	10424841	2550217	11386306	25634521
#duplicates	46818	10812	2507	400149	460286
%duplicates	3%	0%	0%	3%	1%

of the least frequent of the words, because we want to significantly reduce the weight of the less typical triplets.

So for the above example, the fully processed triplet is:

```
Robert De Niro    Acted    Raging Bull                -53.75    <actedIn>    YAGO
```

The processed triplet is stored in the following format, as a text String:

```
<entity> TAB <rel_ID> TAB <value> TAB TAB TAB TAB <T> TAB <original_ID> TAB <source>
```

During the process, all the duplicates are detected and discarded (only about 1% all together), and so are all the triplets where either entity or value is empty (an insignificant amount, just 52 out of millions of triplets).

WordNet 3.1 is a highly reliable source, because it is manually authored and curated, so all of its data is used here (117,791 senses), and the references (380,689) are used as relations (12 different).

YAGO 3 is automatically (with manual assistance) created from Wikipedia, WordNet, and GeoNames, and only the most reliable core data is used here, about 3M (out of 10M total) of its entities and 10M assertions (out of 120M total). We use only the files yagoFacts.tsv, yagoDateFacts.tsv, and yagoLiteralFacts.tsv.

NELL (iteration 965) is also automatically created from the Web, with manual assistance, and its core of 2M+ most reliable assertions (out of 50M total) is used here. We use only the file NELL.08m.965.esv.csv.txt.

Finally, we find that the automatically created ConceptNet 5 is by far the least reliable of the four, and we use around 12M of its core English assertions.

Table 4 provides some basic statistics about the used data resources.

Six of the most frequent relations in triplets are:

- IsInstance x 7,467,825
- IsHypernym x 2,441,806
- IsLocated x 2,199,041
- IsRelated x 1,879,061
- WasBornOn x 1,334,801
- HasGender x 1,108,862

Querying of this resource is enabled by Triplet Searcher, a Java program that can be run from a command line on a Mac or a Linux machine. It loads the preprocessed triplets and in the process indexes all the words that appear in the entity or value fields. For the 25M triplets it generates an index with 3M keywords and 95M triplet references. It uses around 10GB of heap and around 16GB of RAM in total. Due to its relatively high computational requirements, its functionality is available publicly as a Web application, which is presented in Section 3.2.1, and through a ConCreTeFlows widget that is described in Section 3.2.2.

3.2.1 Web application

The triplet querying functionality is available online as: (I) a Web application and (II) a Web-service. The application is intended for interactive use in a Web browser (see screenshot in

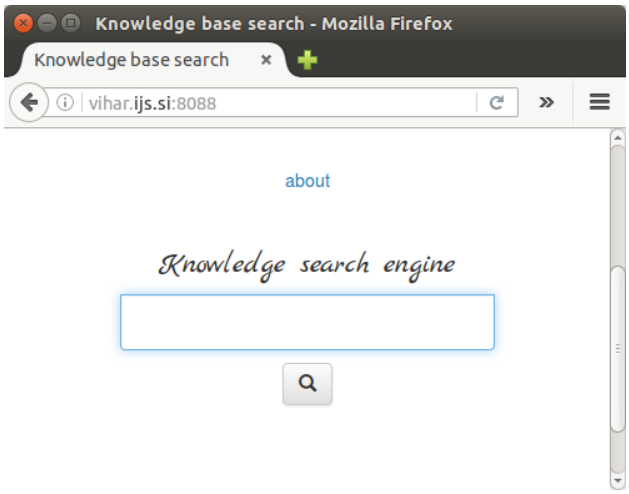

Figure 7: Interactive user interface to the triplet querying tool.

Figure 7) and can be accessed at: `http://vihar.ijs.si:8088/`
The Web service provides an API for programmatic access at: `http://vihar.ijs.si:8088/api` which accepts GET and POST requests and two parameters: *query* and *maxhits*. The result is returned in a simple JSON format. Here is a simple example request:
`http://vihar.ijs.si:8088/api?query=spam&maxhits=5`

3.2.2 The ConCreTeFlows widget

Triplet querying in the merged resource is available also through a ConCreTeFlows widget named *Search KnowledgeBase* (in the *Triplet Extraction* folder), which extends the previously developed ConCeptNet and WordNet data service widgets.

The widget expects a query string as input and the maximum number of returned triplets as a parameter. In Figure 8, the widget is shown in an example workflow, where its resulting triplets go through some re-formatting and graph formation to finally enter an existing widget for graph visualization (the visualization itself is not shown in the figure).

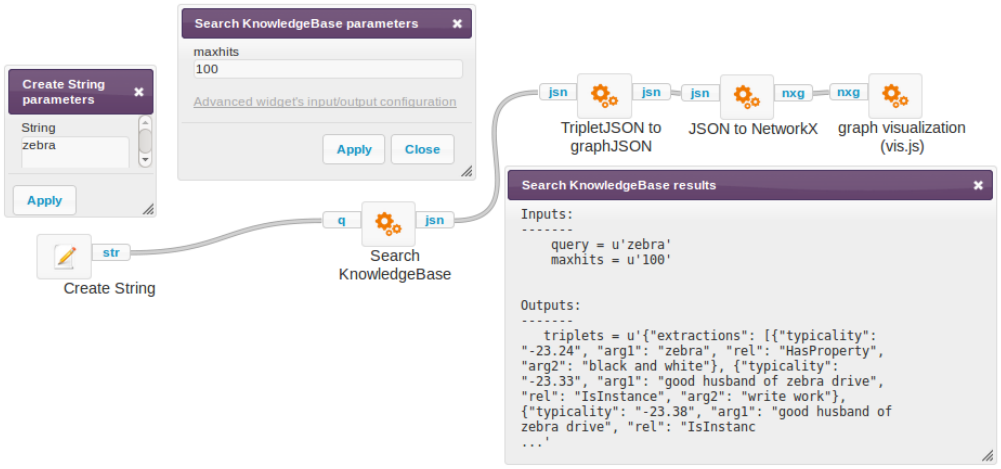

Figure 8: The *Search KnowledgeBase* widget in an example ConCreTeFlows workflow. In the *Search KnowledgeBase results* window we can see example inputs and first few lines of outputs.

4 Conclusions

In the scope of task T6.2 in WP6, we prepared a number of datasets and data-related services. Early and mostly exemplary items of this kind were reported on in Deliverable D6.2 at month 12 of the project, while the datasets and services that were needed and developed afterwards are presented in this document.

Demand for data was not as high and strict as envisaged at the start of the project. However, there was a need for common access to several data resources and for filtering and curation of such data. Namely, many of the data resources have deficiencies or unfavorable characteristics that require some data preprocessing before their use in the software components of the project. The data services that we describe in this document and some software tools from D6.5 are aimed at supporting these efforts.

The specific gathered datasets that are reported here were used in the research work in the project, but some of them are still to be employed in specific tasks and initiatives of WP3, WP5 and WP9. Most of the developed datasets and data-access tools are (or will be) made publicly available.

References

[1] Blaž Fortuna, Marko Grobelnik, and Dunja Mladenić. OntoGen: Semi-Automatic ontology editor. In *Human Computer Interface (Part II) (HCI 2007), LNCS 4558*, volume 4558, pages 309–318, 2007.

[2] Pedro Martins, Tanja Urbančič, Senja Pollak, Nada Lavrač, and Amílcar Cardoso. The good, the bad, and the AHA! Blends. In *Proceedings of the Sixth International Conference on Computational Creativity (ICCC 2015)*, page 166–173, Park City, Utah, June - July 2015. Brigham Young University, Brigham Young University.

[3] Senja Pollak, Anze Vavpetič, Janez Kranjc, Nada Lavrač, and Špela Vintar. NLP workflow for on-line definition extraction from English and Slovene text corpora. In Jeremy Jancsary, editor, *Proceedings of KONVENS 2012*, pages 53–60. ÖGAI, September 2012.

[4] Špela Vintar. Bilingual term recognition revisited the bag-of-equivalents term alignment approach and its evaluation. *Terminology*, 16:141–159, 2010.

A Semi-automated Computational Creativity Conceptualization Grounded on ICCC Proceedings Papers (paper submitted to the ICCC 2016 conference)

Semi-automated Computational Creativity Conceptualization Grounded on ICCC Proceedings Papers (Study paper)

Senja Pollak[1], Biljana Mileva Boshkoska[1,3], Dragana Miljkovic[1], Geraint A. Wiggins[2], Nada Lavrač[1,4]

[1]Dept. of Knowledge Technology, Jožef Stefan Institute, Ljubljana, Slovenia
[2]Computational Creativity Lab, Queen Mary University of London, London E1 4NS, UK
[3] Faculty of Information Studies, Novo Mesto, Slovenia
[4] University of Nova Gorica, Nova Gorica, Slovenia
{senja.pollak,biljana.mileva,dragana.miljkovic,nada.lavrac}@ijs.si; geraint.wiggins@qmul.ac.uk

Abstract

Concept Creation Technology is concerned with engineering software that exhibits creative behaviour of conceptualization. For a given domain of interest, whose conceptual space is closed, pre-defined and yet unexplored, it is interesting to study computational means for (semi-)automated domain conceptualization. In information science it is considered that domain conceptualization can be realized by (one or several) ontologies. This paper presents a method of semi-automated domain conceptualization, where the domain of interest is Computational Creativity (CC). Grounded on papers, which were published in six consecutive years since 2010 in the Proceedings of International Conferences on Computational Creativity (ICCC), this paper proposes a tentative conceptualization of the CC domain. Some additional properties of the CC domain are studied, analyzed by means of fully mechanical or semi-automated information extraction and dependency analysis techniques. This approach affords a unique opportunity for automated historiography of a research field.

Introduction

As a sub-field of Artificial Intelligence research, *Computational Creativity* (CC) (Boden, 2004; Colton, 2008) is concerned with engineering software that exhibits behaviours which would reasonably be deemed creative. A part of CC research addresses *Concept Creation Technology*, concerned with engineering software that exhibits creative conceptualization behaviour.

For a given domain of interest, whose conceptual space (Boden, 2004) is closed, pre-defined and yet unexplored, it is interesting to study computational means for automated (or semi-automated) domain conceptualization. In the current research we use the term *conceptualization* in alignment with its standard use in information science. In information science, a *conceptualization* is defined as "an abstract (simplified) view of some selected part of the world, containing the objects, concepts, and other entities that are presumed of interest for some particular purpose and the relationships between them." Domain conceptualization is, in information science, frequently realized by defining (one or several) ontologies formally describing the domain of interest (Gruber, 1993; Smith, 2003).

Manual construction of ontologies represents a significant investment of human resources when used for modeling a new domain. Therefore, methods for (semi-)automated extraction of domain knowledge from unstructured texts were developed, including automated taxonomy construction described by Velardi, Faralli, and Navigli (2013).

While an ontology is a "formal, explicit specification of a shared conceptualization" (Gruber, 1993), represented as a set of domain concepts and the relationships between them, a so-called *topic ontology* is a set of domain topics or concepts—formed of related documents—represented by the most characteristic topic keywords and related by the `subconcept-of` relationship (Fortuna, Grobelnik, and Mladenić, 2007). The task addressed in this paper is semi-automated construction of a topic ontology from documents in the area of computational creativity.

The problem of CC domain conceptualization has not been substantially addressed in the CC literature. Jordanous and Keller (2012) used automated natural language processing methods and a statistical measure of association to identify words related to creativity (in general, not specifically to CC). They clustered the words into semantically-related groups by using a lexical similarity measure (see the resulting ontology of creativity at `http://purl.org/creativity/ontology`). Other authors presented extraction of creativity concepts related to e.g., sub-fields of creativity (Agres et al., 2015) and creativity evaluation (van der Velde et al., 2015).

This approach to the study of collections of documents opens the prospect of an automated historiography of the field of computational creativity, an idea which constitutes a satisfyingly recursive application of the research outputs of that area of interest. In this paper, we illustrate, with real computational examples, the kinds of analysis (e.g., diachronic comparisons of conceptualisation) that would be used for such studies.

The current paper presents a method of semi-automated domain conceptualization, where the domain of interest is Computational Creativity (CC). The paper proposes a possible conceptualization of the CC domain grounded on papers, published in six consecutive years since 2010 in the Proceedings of International Conferences on Computational Creativity (ICCC). Some additional properties of the CC domain are studied, obtained by means of information extraction and dependency analysis techniques. The experimental data is presented, followed by the results of CC domain conceptualization and time dependency analysis.

The Computational Creativity Domain

This section describes the data used in the experiments, together with initial domain understanding achieved through automated terminology extraction.

ICCC proceedings data

The documents were taken from the ICCC proceedings published between 2010 and 2015, inclusive. In total, we considered 247 articles from 6 consecutive proceedings, containing the following numbers of articles: ICCC-2010: 43, ICCC-2011: 30, ICCC-2012: 44, ICCC-2013: 40, ICCC-2014: 49, ICCC-2015: 41.

ICCC papers, which were available in PDF, were first converted to text documents. We also omitted the references, but added the information about the conference year (note however, that for time dependency analysis, presented in the last section of this paper, the version of the corpus including references was used).

Automated CC terminology extraction

This section describes an attempt to improve ICCC domain understanding through CC terminology extraction. The goal of terminology extraction is to automatically extract relevant terms for a given domain, represented by a given corpus. We used terminology extraction method LUIZ-CF (Pollak et al., 2012), a modified version of the LUIZ term recognition tool (Vintar, 2010). LUIZ-CF is implemented as a workflow in the ClowdFlows environment.

Term extraction consists of two steps: extracting the noun phrase candidates based on morphosyntactic patterns; followed by weighting and ranking of the candidates based on their termhood value, for single word and multi-word terms. The termhood value is computed based on comparison of relative frequencies of lemmas of a term in the domain corpus (here, the ICCC proceedings) compared to a reference corpus: for English, the frequencies of the British National Corpus are used. The extracted terms are ranked by termhood value on a scale between 1 and 0. In addition to default stopwords, we eliminated also the names of ICCC PC members, leading to the exclusion of some of the paper authors from the term list. The top ranked candidates are listed in Table 1, followed by a list of top ranked multi-word terms from the same term list given in Table 2. The term extraction workflow is available in ClowdFlows at: `http://clowdflows.org/workflow/7219/`. The extracted terms may be considered as an initial computational creativity vocabulary for building a dictionary of computational creativity, which is planned in future work.

Table 1: Top 15 terms from the ICCC corpus.

Score	Term
1.000000	[creativity]
1.000000	[computational creativity]
0.862623	[system]
0.247012	[creative system]
0.182190	[process]
0.174810	[model]
0.141525	[image]
0.126607	[concept]
0.102306	[creative process]
0.101973	[word]
0.099952	[evaluation]
0.099844	[conceptual space]
0.081564	[domain]
0.080851	[generation]
0.073521	[story]

Table 2: Top 15 multi-word terms from the ICCC corpus.

Score	Term
1.000000	[computational creativity]
0.247012	[creative system]
0.102306	[creative process]
0.099844	[conceptual space]
0.030894	[computational model]
0.021593	[computational system]
0.018712	[fitness function]
0.018064	[jaguar knight]
0.012638	[genetic algorithm]
0.012078	[human creativity]
0.011300	[poetry generation]
0.011171	[story generation]
0.010011	[neural network]
0.009657	[creative domain]
0.009299	[transformational creativity]

CC domain conceptualization with OntoGen

A tool named OntoGen (`http://ontogen.ijs.si/`: Fortuna, Grobelnik, and Mladenić, 2007) was used to build a topic ontology for CC domain conceptualization. OntoGen is a semi-automatic and data-driven ontology editor. Semi-automatic means that the system is an interactive tool that aids the user during the topic ontology construction process. Data-driven means that most of the aid provided by the system is based on the underlying text data (document corpus) provided by the user. The system combines text mining techniques with an efficient user interface.

OntoGen accepts texts in various formats. We chose the named line document format, where each line represents one document, starting with the document ID and the conference edition (2010, 2011, 2012, 2013, 2014 or 2015) as a category. OntoGen performs basic lemmatization and stopword removal (and accepts additional user-defined stopword lists) and constructs Bag-of-Words (BoW) vector representations of documents, weighted by the TF-IDF weights (Salton and Buckley, 1988), where TF-IDF stands for Term Frequency-Inverse Document Frequency. For a given term w in document d from corpus D, the TF-IDF measure is defined as follows:

$$\text{tf–idf}(w,d) = \text{tf}(w,d) \times \log \frac{|D|}{|\{d \in D : w \in d\}|}, \qquad (1)$$

where $\text{tf}(w,d)$ represents the number of times term w appears in document d.

The OntoGen platform is illustrated in Figure 1. The "unsupervised concept suggestion" functionality is a central part of the system: for a given concept (e.g., the central concept "computational creativity" represented by all the documents of the ICCC domain), a list of sub-concepts is suggested by k-means clustering (Jain, Murty, and Flynn, 1999) and Latent Semantic Indexing (Deerwester et al., 1990) techniques. If the user does not want to affect the conceptualization outcome, only parameter k needs to be chosen by the user to determine the number of concepts, i.e., the number of categories in which the documents will be clustered. "Keywords" (automatically assigned names of clusters) are the words that are the most descriptive for the content of the concepts instances (articles), i.e. words with the highest weights in the document centroid vectors (Fortuna, Mladenič, and Grobelnik, 2006). The main OntoGen's window represents the ontology visualisation in which each concept is represented by top three key-

words unless manually edited, while the Concept hierarchy window (on the upper left corner) offers a quick overview of all the concepts with their position in the concept hierarchy that can be also directly manipulated.

An analternative view is over the Concepts' documents, where documents of each concept (document cluster) are visualized. Figure 2 shows documents of the selected concept, i.e. the one represented by keywords "music, chord, improvisation", which could be reasonably be called "Musical creativity". In the similarity graph (at the bottom of the figure), the red dots represent documents belonging to the selected concept, while blue dots the documents not belonging to the concept. The graph inspection functionality can be used for selecting documents to be manually inspected and eventually removed or added to the concept. SVM keywords (see left bottom corner of the figure) are composed from words most distinctive for the selected concept with regards to its sibling concepts in the hierarchy (obviously not available for the root concept). This functionality was particularly interesting for finding contrastive keywords for each year presented in Table 5.

An important additional functionality of OntoGen is a supervised method for adding concepts. We used it, for example, to create the concept "Evaluation" in Figure 4. It is based on SVM active learning method. The user supervision is provided first by a query describing the concept that the user has in mind and followed by a sequence of questions whether a particular instance (document) belongs to the concept and the user can select Yes or No. The questions are chosen from the instances on the border between being relevant to the query or not and are therefore most informative to the system. The system refines the suggested concept after each reply from the user and the user can decide when to stop the process based on how satisfied he is with the suggestions. After the concept is constructed it is added to the ontology as a sub-concept of the selected concept.

Automated CC conceptualization

First we performed k-means clustering for k set to 5. In the topic ontology, shown in Figure 1, the first three automatically extracted keywords are used by OntoGen as concept/topic descriptors.

By inspecting the keywords, we manually named the concepts of the automatically generated topic ontology as follows: "Musical creativity", "Visual creativity", "Linguistic creativity", "Creativity in Games" and "Conceptual creativity" (cf. Table 3). While some of the categories are quite uniform regarding the keywords (e.g., "music, chord, improvisation, melodies, harmonize, composition, accompaniment, pitch, emotions, beat" for the concept category that we named "Musical creativity"), other categories are more noisy, e.g.,"image, story, actions, painting, character, agents, narrative, artists, robot, darci" do not denote a uniform category. We decided to name this category "Visual creativity", but it obviously contains documents from other topics as well, such as narratives generation.

Out of a set of ten automatically generated keywords characterizing each of the five document clusters, presented in Table 3, we manually selected three keywords that we believe best describe the cluster of papers belonging to the given concept (in italics). These keywords were added to the concept labels of Figure 3.

In the next section, aiming at a more elaborated version of the CC ontology (cf. Figure 4), we use the concept moving facility of OntoGen, by which we moved e.g., the concept "Narrative" from "Visual" to "Lexical", together with other techniques for manipulating the initial ontology.

Semi-automated CC domain conceptualization

This section describes improved CC conceptualization, created by manipulating the initial ontology using different OntoGen functionalities. The main concepts were further divided and when forming meaningful concepts, the categories were added as subconcepts (see e.g. the sub-concepts of Linguistic creativity in Figure 4). As already mentioned, some (sub-)concepts were moved, e.g., the "narrative" category was moved from Visual to Lexical creativity. Some concepts were added by query and active learning. On the first level, this is the case for the category Evaluation, which was a recurrent topic in other categories and we used query and active learning to form an independent category. We also used it, e.g., for creating the category "Recipes" as a sub-concept of lexical creativity. We also used the OntoGen function to (de)select the documents being categorized to one concept category.

Figure 2 shows the documents belonging to a category. It is very interesting to inspect some outliers (documents similar to documents in the category not being classified to this category). In the concept document graph, we identified some of the outliers, represented by blue dots in the similarity line of red dots. An example is article 2014_44, entitled "Arts, News, and Poetry The Art of Framing", by Gross, Toivanen, Laane and Toivonen. This paper was not classified in the Linguistic creativity category, but was identified as similar to the documents of that category. The article indeed refers to linguistic creativity (poetry) but also to generated pictures as well as pictures painted by an artist. We manually added this document also to the category of linguistic creativity.

The result of this experiment is shown in Figure 4. On the first level we distinguish between Musical, Visual, Linguistic creativity, Games and creativity, Conceptual creativity as well as newly created category of Evaluation. On lower levels, we added e.g. Narratives, Poetry, Recipes and Lexical creativity for Linguistic creativity, where the latter comprises e.g., humour, neologisms, etc.

Each concept is represented by descriptive Keywords (cf. keywords for six first level concepts in Table 4) out of which we selected three keywords (in italics) to represent the concept in the visual ontology (Figure 4). The ontology can be considered as a draft to be improved by a collective effort by the CC community. In addition, since the concepts are grounded by the documents, the top ranked documents could be considered as interesting reading for incomers to the field of computational creativity. The bibliography can be created for concepts of different levels (as an example see three articles per selected topic):

Narratives:

- A System for Evaluating Novelty in Computer Generated Narratives (Pérez y Pérez et al., 2011)
- Kill the Dragon and Rescue the Princess: Designing a Plan-based Multiagent Story Generator (Laclaustra et al., 2014)
- Creativity in Story Generation From the Ground Up: Non-deterministic Simulation driven by Narrative (León and Gervás, 2014)

Games:

- Computational Game Creativity (Liapis, Yannakakis, and Togelius, 2014)
- Ludus Ex Machina: Building A 3D Game Designer That Competes Alongside Humans Michael (Cook and Colton, 2014)
- Knowledge-Level Creativity in Game Design (Smith and Mateas, 2011)

Music:

- Automatic Generation of Music for Inducing Emotive Response (Monteith, Martinez, and Ventura, 2010)

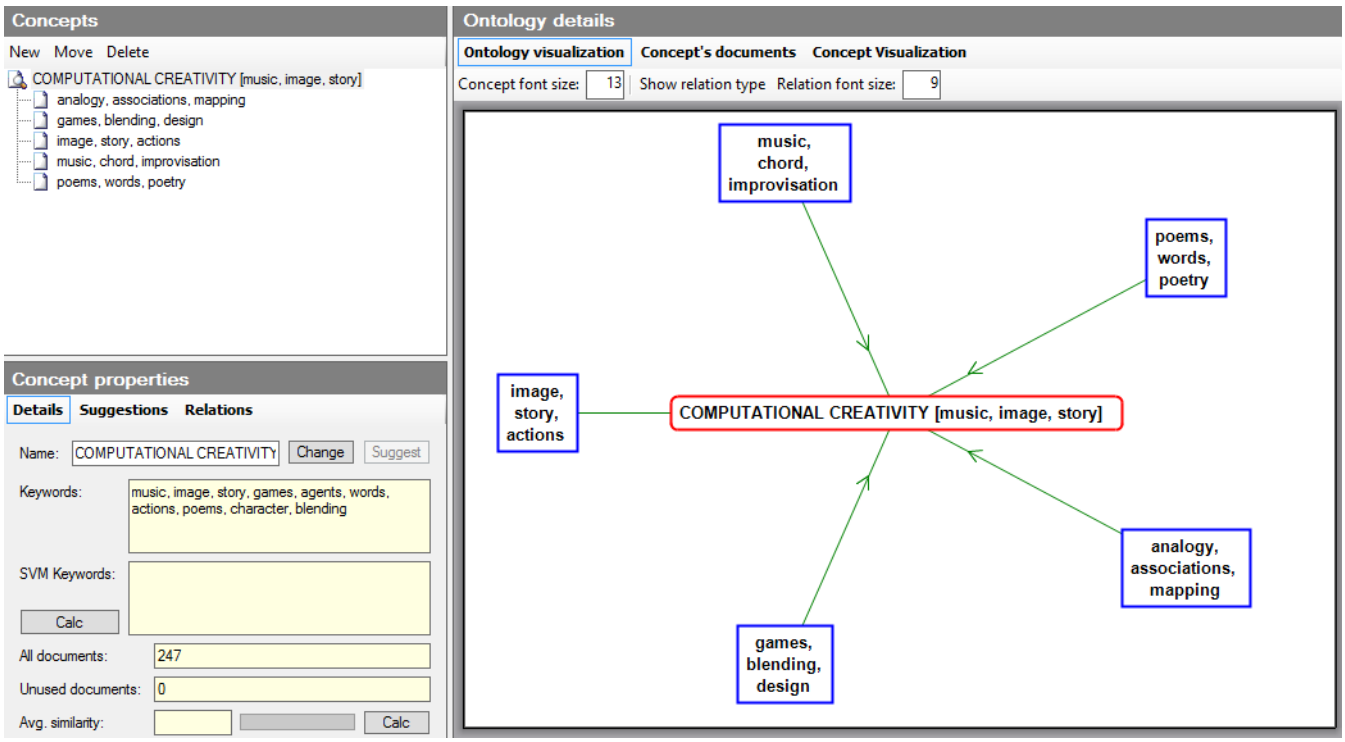

Figure 1: Automatically generated conceptualization of the CC domain.

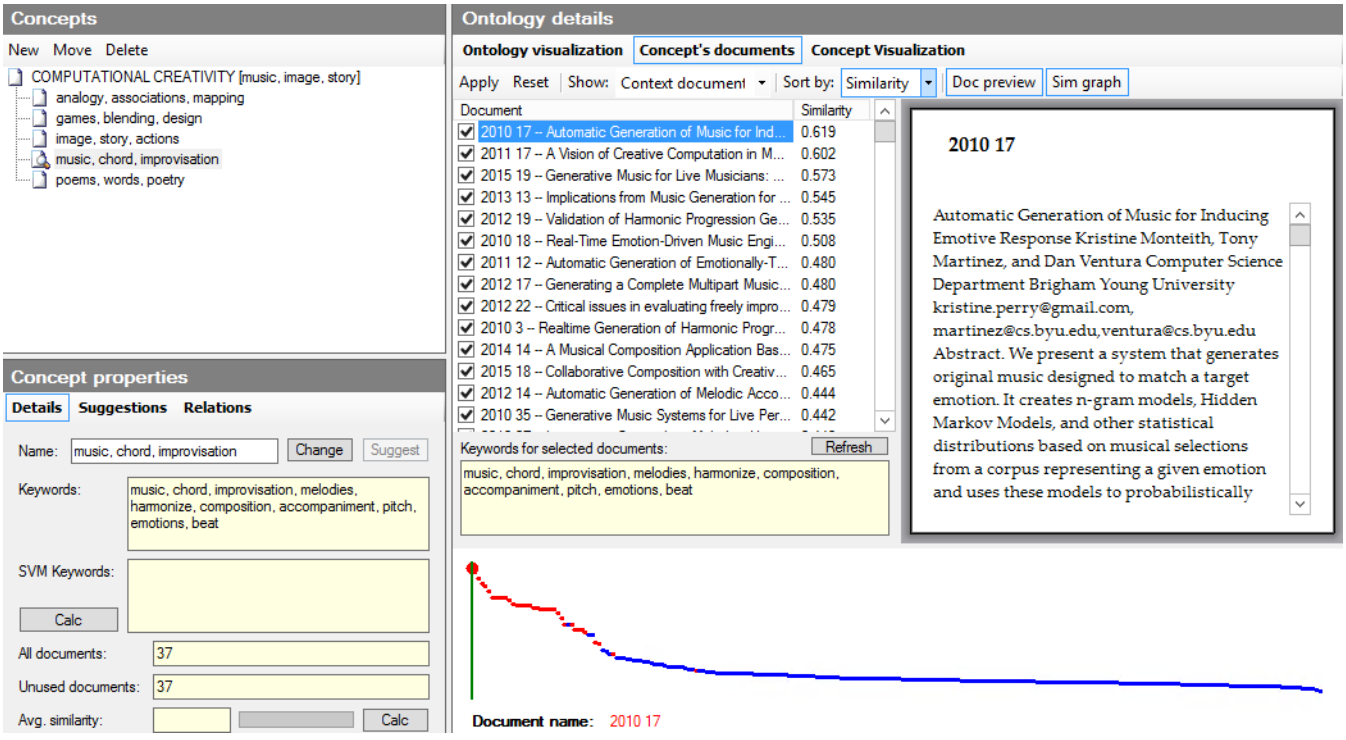

Figure 2: Document view of the automatically generated conceptualization of the CC domain.

Table 3: Automatically generated concepts (concept names were manually determined) and keywords.

Concept	Automatically extracted keywords
Music	*music*, chord, *improvisation*, melodies, harmonize, *composition*, accompaniment, pitch, emotions, beat
Visual	*image*, story, actions, *painting*, character, agents, narrative, *artists*, robot, darci
Linguistic	*poems*, words, poetry, artefacts, *story*, evaluating, creativity_system, predict, *text*, creativity
Games	*games*, blending, *design*, analogy, *player*, conceptual, games_design, angelina, ontology, agents
Conceptual	*analogy*, associations, *mapping*, graphs, objective, problem, fractal, domain, *representation*, relationship
Comp. creativity	*music*, *image*, *story*, games, agents, words, actions, poems, character, blending

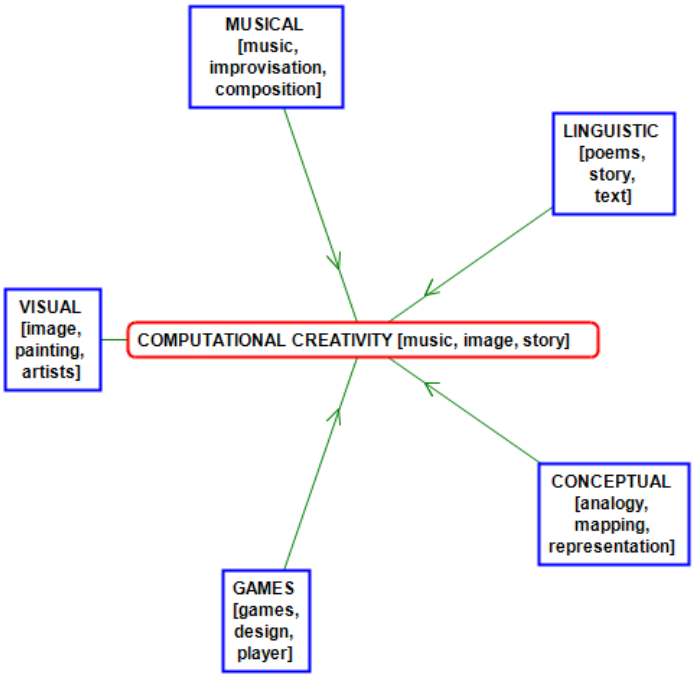

Figure 3: CC conceptualization through automated clustering, concept naming and manual keywords selection.

- A Vision of Creative Computation in Music Performance Roger (Dannenberg, 2011)
- Generative Music for Live Musicians: An Unnatural Selection (Eigenfeldt, 2015)

Analysis of CC domain development in time

In this section we investigate temporal aspect of ICCC proceedings. First, we use OntoGen by which we first categorize the articles into editions and then observe the words distinguishing these categories. Second, the frequency analysis of terms was used to identify terms that are characteristic only for first or last editions. Last but not least, we use the copulas for investing the time dependency between the content of different conference editions.

Distinctive keywords by years using OntoGen

One of the functionalities of the tool OntoGen (Fortuna, Grobelnik, and Mladenić, 2007) is the supervised categorization of documents to predefined categories. For this experiment we used conference editions as categories and extracted for each edition its characteristic keywords. The first set of descriptive keywords (KeyW in Table 5) is extracted using document centroid vectors, while the second set of distinctive keywords (SVM in Table 5) is extracted from the SVM classification model dividing documents in the topic from the neighbouring documents (Fortuna, Mladenič, and Grobelnik, 2006). In Table 5 we present both sets of words for each year and in Figure 5 the Concept view in which each year is represented by a selection of three keywords out of SVM keywords denoting this year in the list of Table 5.

As one would expect, the descriptive keywords overlap in different editions. For example, the most recurring words across years are “creativity, design, modelling, system”. More interesting are the distinctive keywords. In this respect, ICCC-2015 might be characterized for example by “bots”, ICCC-2014 and ICCC-2011 by “games”, 2014 by “metaphors” and “stereotypes”, ICCC-2012 by “melodies” and “associations”, and ICCC-2010 by “analogies” among others.

The categorization by years can be used also for specific topics. For example, in the ontology presented in Figure 4, we can split a selected topic to year categories and observe the distinctive (SVM) keywords. In this case, the papers representing the concept of Musical creativity in 2015 contain words such as “musebot, pc, unnatural...”, while in 2010 the words are “chord, improvisation, jazz ...”, etc.

Terminology distribution by years

The frequency distributions of the top 1,000 terms obtained by the terminology extraction process described earlier in the paper are an indicator to detect terms that recently occurred or were present only in the early editions. Examples of terms that appear in 2014

Table 4: Categories and keywords of the first layer of the semi-automatically constructed CC ontology.

Category	Automatically extracted keywords
Musical	*music*, chord, *improvisation*, melodies, harmonize, *composition*, accompaniment, pitch, emotions, beat
Visual	*image*, *painting*, darci, artifacts, *collage*, adjectives, associations, rendered, colored, artists
Linguistic	*story*, *poems*, actions, character, *words*, agents, narrative, artefacts, poetry, evaluating
Games	*games*, *design*, *player*, games_design, angelina, agents, code, jam, filter, gameplay
Conceptual	*analogy*, *blending*, mapping, conceptual, objective, *associations*, team, graphs, concepts, domain
Evaluation	music, poems, improvisation, *evaluating*, interactive, poetry, *creativity system*, musician, *participants*, behavioural
Comp. creativity	*music*, *image*, *story*, games, agents, words, actions, poems, character, blending

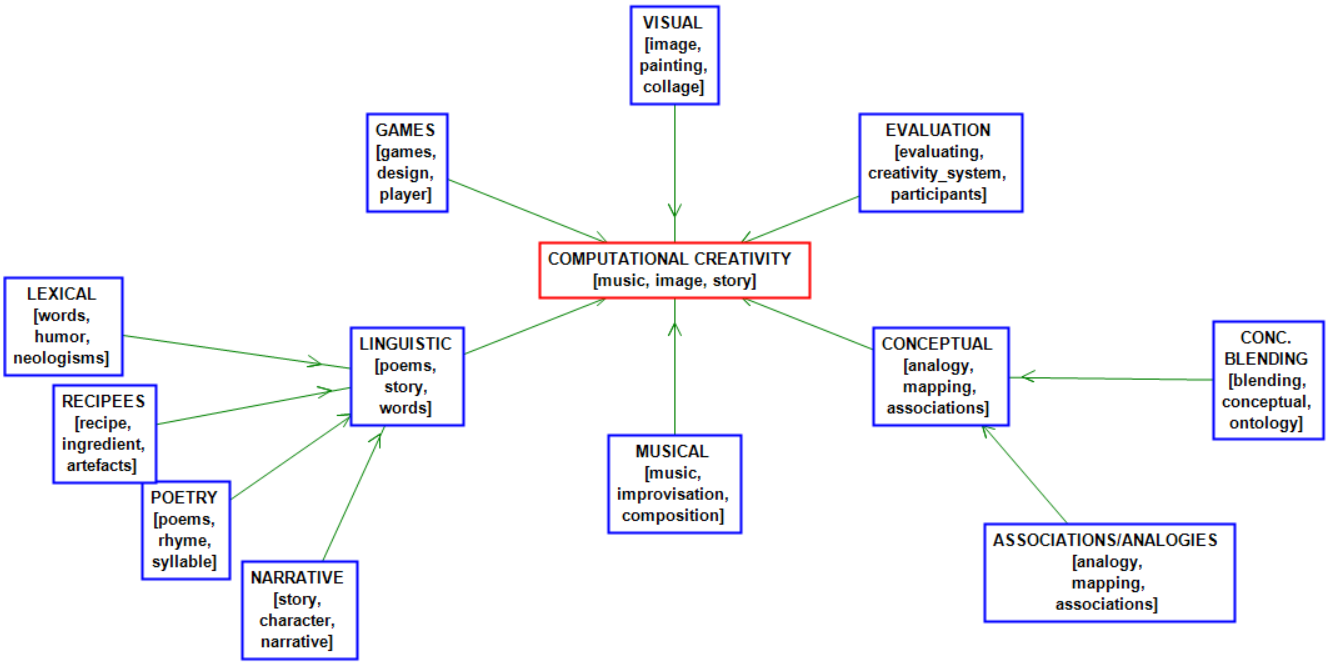

Figure 4: Semi-automatically generated conceptualization of the CC domain, with concept naming and subconcept creation.

Table 5: Keywords and distinctive (SVM) words by year.

Year	Category	Words
2015	KeyW:	creativity, generated, image, work, human, blending, design, based, music, words
	SVM:	*blending, humor, bots, choice, vectors, musician, jam, conceptual, colour, participants*
2014	KeyW:	creativity, computer, process, modelling, evaluating, words, agents, story, domain, based
	SVM:	*games, agents, story, artists, adjectives, ontology, domain, motifs, poems, actions*
2013	KeyW:	modelling, process, figure, image, design, performs, based, levels, interactive, concepts
	SVM:	*robot, metaphor, motion, surprising, evolved, image, composition, mechanism, stereotypes, fictional*
2012	KeyW:	creativity, system, computer, music, evaluating, user, human, figure, work, set
	SVM:	*melodies, associations, accompaniment, character, template, shape, player, monotone, text, cluster*
2011	KeyW:	creativity, system, story, design, modelling, results, games, music, set, problem
	SVM:	*story, games, movements, playing, graphs, games_design, darci, actions, identical, strategies*
2010	KeyW:	generated, system, user, set, idea, design, emotions, analogy, developments, based
	SVM:	*analogy, chord, emotions, improvisation, genes, filter, lives, team, jazz, songs*

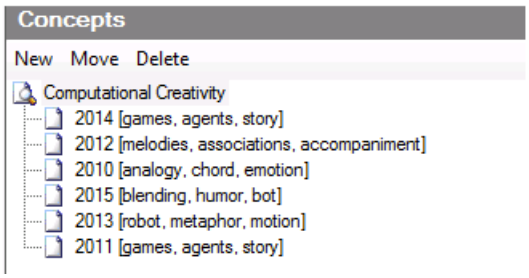

Figure 5: CC distinctive SVM keywords per year.

and 2015 and not before are: "game jam, co-creative system, concept invention, generative software, curation coefficient, procedural generation, player goal, network analysis, simulation model" (30 terms in total). In contrast, the terms that were used in 2010 and 2011 are: "fractal representation, fractal feature, basel problem, sensory system, fractal algorithm".

Table 6: Results from bi-variate copulas for all terms.

No.	Copula type	Coupling order of domains in MVC	θ
1	Best Clayton	2010-2012	3.1941
2	Best Frank	2010-2012	9.1507
3	Worst Clayton	2010-2015	2.4010
4	Worst Frank	2010-2015	7.3955

Copula-based analysis of dependencies between ICCC proceedings

This section describes measuring the detected dependencies between different years of ICCC proceedings. As in the previous section, we counted the frequencies of automatically extracted terms in the ICCC proceedings of each year from 2010 until 2015. This information was used as input for the copula-based dependency analysis between six ICCC proceedings, described below.

The correlation between two variables (e.g., the ICCC proceedings of two distinctive years) can be measured by means of the Pearsons correlation coefficient. It is a dimensionless quantitative measure of statistical relationships between two (or more) variables. It measures the degree of linear correlation, however the two variables may have different functional dependency. For this reason we apply the copula functions as a tool for studding and measuring the dependences of random variables.

In probability theory, the dependence between random variables is completely defined by their joint distribution function. The joint distribution function $H(x, y)$ for two random variables (r.v.) X and Y, specified on the same probability space, defines the probability of a random event in terms of both X and Y. It is given by:

$$H(x, y) = P[0 \leq X \leq x, 0 \leq Y \leq y] \quad (2)$$

where P is a probability distribution function. To find the joint distribution function in analytical form, we use the Sklar's theorem (Sklar, 1959) which proves that the joint distribution function of two r.v. is equal to the copula of their uniform distributions on the unit interval [0,1].

Theorem 1 (Sklar's theorem): Let H be a bivariate distribution function with marginal distribution functions $u_1 = F(x)$ and $u_2 = G(y)$. Then copula C exists such that for all $x, y \in \mathbb{R}$:

$$H(x, y) = C(F(x), G(y)) = C(u_1, u_2) \quad (3)$$

If $F(x)$ and $G(y)$ are continuous, then C is unique; otherwise C is uniquely determined on $Range(F) \times Range(G)$. Conversely, if C is a copula and $F(x)$ and $G(y)$ are distribution functions, then the function H defined by (3) is a joint distribution function.

Copulas are functions that manage to formulate the multivariate distribution in such a way that various general types of dependences including the non-linear one may be captured. We focus on two families of bi-variate Archimedean copulas: Clayton and Frank. Their usage is mainly motivated by their convenient properties, such as symmetry and associativity. Their mathematical forms are presented in Table 7. The notation $\varphi_\theta(t)$ represents a so called generator function that is responsible for constructing the copula function. The parameter θ is estimated from the data. Higher values of θ mean higher dependence between the two variables.

We explored the dependencies between pairs of proceedings. For this purpose we built Clayton and Frank bi-variate copulas. The results of best copulas (the most dependent pairs of proceedings) and the worst copulas (the least dependent pairs) are provided in Table 6. It can be observed that the ICCC-2010 and ICCC-2015 proceedings are contents-wise the least connected, while the most dependent proceedings are those from 2010 and 2012, where both conferences were organized in Europe.

Conclusions

In the paper, we have presented the conceptualization of the computational creativity domain by semi-automated topic ontology construction based on the corpus of ICCC proceedings. We analysed automatically extracted keywords and subconcepts for CC domains (Visual, Musical, Linguistic, Conceptual creativity, Creativity and games and Evaluation). In addition we analysed characteristics of different editions of CC conferences and used copulas to measure the dependency between proceedings of different editions. As result of this research, we make available for further research a) the ICCC proceedings corpus in .txt format with and without reference sections, b) automatically extracted ICCC terminology that can be used for future efforts in creating a CC glossary, c) fully automated topic ontology with automatic keywords extraction, as well the semi-automated CC ontology, which is the result of manual manipulation of the automatic ontology. Ontologies are available in .png and .rdf formats. All the resources are available at: `http://kt.ijs.si/senja_pollak/CC_resources/`.

In future, as these techniques develop to full automation, and the amount of data increases with successive conferences, it will be possible to construct a timeline of the development of the field of computational creativity, using the objective analysis of the literature. It will be possible to trace the rise and fall of trends, and their success or failure, and to identify the development of core CC science, as proposed by (Lakatos, 1970). This activity will be unique in science, and will support an unprecedented level of unbiased philosophical reflection on the field of computational creativity.

Acknowledgements

We acknowledge the support of the Slovenian Research Agencyand European projects Prosecco (g.n.611560) and ConCreTe (g.n.611733) that acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET grant number 611733.

Table 7: Various Archimedean copulas, their generator functions φ and borders of θ parameter.

Copula type	$C_\theta(u,v)$	$\varphi_\theta(t)$	θ
Clayton	$\left[\max\left(u^{-\theta}+v^{-\theta}-1,0\right)\right]^{-1/\theta}$	$\frac{1}{\theta}\left(t^{-\theta}-1\right)$	$[-1,\infty)\setminus\{0\}$
Frank	$-\frac{1}{\theta}\ln\left(1+\frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{e^{-\theta}-1}\right)$	$-\ln\frac{e^{-\theta t}-1}{e^{-\theta}-1}$	$(-\infty,\infty)\setminus\{0\}$

References

Agres, K.; McGregor, S.; Purver, M.; and Wiggins, G. 2015. Conceptualizing creativity: From distributional semantics to conceptual spaces. In *In the Sixth International Conference on Computational Creativity, ICCC 2015*.

Boden, M. A. 2004. *The Creative Mind: Myths and Mechanisms*. Routledge.

Colton, S. 2008. Creativity versus the perception of creativity in computational systems. In *In Proceedings of the AAAI Spring Symp. on Creative Intelligent Systems*.

Cook, M., and Colton, S. 2014. Ludus ex machina: Building a 3d game designer that competes alongside humans. In *Proceedings of the Fifth International Conference on Computational Creativity*, ICCC2014, 54–62. International Association for Computational Creativity.

Dannenberg, R. B. 2011. A vision of creative computation in music performance. In *Proceedings of the Second International Conference on Computational Creativity*, 84–89.

Deerwester, S.; Dumais, S. T.; Furnas, G. W.; Landauer, T. K.; and Harshman, R. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 41(6):391–407.

Eigenfeldt, A. 2015. Generative music for live musicians: An unnatural selection. In *Proceedings of the Sixth International Conference on Computational Creativity (ICCC 2015)*, 142–149. Park City, Utah: Brigham Young University.

Fortuna, B.; Grobelnik, M.; and Mladenić, D. 2007. Ontogen: Semi-automatic ontology editor. In *Human Computer Interface (Part II) (HCI 2007), LNCS 4558*, volume 4558, 309–318.

Fortuna, B.; Mladenič, D.; and Grobelnik, M. 2006. *Semantics, Web and Mining: Joint International Workshops, EWMF 2005 and KDO 2005, Porto, Portugal, October 3-7, 2005, Revised Selected Papers*. Berlin, Heidelberg: Springer Berlin Heidelberg. chapter Semi-automatic Construction of Topic Ontologies, 121–131.

Gruber, T. R. 1993. A translation approach to portable ontology specifications. *Knowledge Acquisition* 5(2):199–220.

Jain, A. K.; Murty, M. N.; and Flynn, P. J. 1999. Data clustering: a review. *ACM Computing Surveys* 31(3):264–323.

Jordanous, A., and Keller, B. 2012. Weaving creativity into the semantic web: a language-processing approach. *Proceeding of the Third International Conference on Computational Creativity* 216 –220.

Laclaustra, I. M.; Ledesma, J. L.; Mendez, G.; and Gervás, P. 2014. Kill the dragon and rescue the princess: Designing a plan-based multi-agent story generator. In *Proceedings of the Fifth International Conference on Computational Creativity*. Ljubljana, Slovenia: Jožef Stefan Institute, Ljubljana, Slovenia.

Lakatos, I. 1970. Falsification and the methodology of scientific research programmes. In Lakatos, I., and Musgrave, A., eds., *Criticism and the Growth of Knowledge*. Cambridge, UK: Cambridge University Press. 91–196.

León, C., and Gervás, P. 2014. Creativity in story generation from the ground up: Non-deterministic simulation driven by narrative. In *5th International Conference on Computational Creativity, ICCC 2014*.

Liapis, A.; Yannakakis, G. N.; and Togelius, J. 2014. Computational game creativity. In *Proceedings of the Fifth International Conference on Computational Creativity*. Ljubljana, Slovenia: Josef Stefan Institute, Ljubljana, Slovenia.

Monteith, K.; Martinez, T.; and Ventura, D. 2010. Automatic generation of music for inducing emotive response. In *Proceedings of the International Conference on Computational Creativity*, 140–149. Lisbon, Portugal: Department of Informatics Engineering, University of Coimbra.

Pérez y Pérez, R.; Ortiz, O.; Luna, W.; Negrete, S.; Castellanos, V.; Alosa, E. P.; and Ávila, R. 2011. A System for Evaluating Novelty in Computer Generated Narratives. In Ventura, D.; Gervás, P.; Harrell, D. F.; Maher, M. L.; Pease, A.; and Wiggins, G., eds., *Proceedings of the Second International Conference on Computational Creativity*, 63–68.

Pollak, S.; Vavpetič, A.; Kranjc, J.; Lavrač, N.; and Špela Vintar. 2012. NLP workflow for on-line definition extraction from English and Slovene text corpora. In Jancsary, J., ed., *Proceedings of KONVENS 2012*, 53–60. ÖGAI. Main track: oral presentations.

Salton, G., and Buckley, C. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management* 24(5):513–523.

Sklar, A. 1959. Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris* 8:229231.

Smith, A. M., and Mateas, M. 2011. Knowledge-level creativity in game design. In *In Proc. of the 2nd International Conference in Computational Creativity, (ICCC 2011)*.

Smith, B. 2003. Chapter 11: Ontology. In Floridi, L., ed., *Blackwell Guide to the Philosophy of Computing and Information*, volume 7250. Blackwell. 155–166.

van der Velde, F.; Wolf, R. A.; Schmettow, M.; and Nazareth, D. S. 2015. A semantic map for evaluating creativity. In *Sixth Interantional Conference on Computational Creativity (ICCC 2015)*. Park City, Utah, USA: Brigham Young University.

Velardi, P.; Faralli, S.; and Navigli, R. 2013. Ontolearn reloaded: A graph-based algorithm for taxonomy induction. *Computational Linguistics* 39(3):665–707.

Vintar, Š. 2010. Bilingual term recognition revisited the bag-of-equivalents term alignment approach and its evaluation. *Terminology* 16:141–159.