CONCRETE

CONCEPT CREATION TECHNOLOGY
PROJECT NUMBER: 611733

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

Deliverable 8.7:
D8.7: Report(s) on validation studies of concepts produced by the prototype

Universidad Complutense de Madrid (UCM), Queen Mary University of London (QMUL), Universidade de Coimbra (UCO), University of Helsinki (UH), Jožef Stefan Institute (JSI),

Version: 1.0, final

Executive summary

This document reports on validation studies of concepts produced by the prototypes developed in the project.

While this deliverable reports work carried out in the context of Task 8.4 of Work Package 8 of ConCreTe, the work has very close connections to many other tasks and deliverables, in particular Deliverables D5.1 and D5.2 where applications developed in the project are reported. Many of the papers that form the body of this deliverable also contribute to other tasks, and we refer to deliverables D5.1 and D5.2 for more detailed information on the aspects of the prototypes that are not concerned directly with validation.

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:	D8.7
Deliverable full title:	Report(s) on validation studies
Deliverable short title D8.7:	Report(s) on validation studies
Document identifier:	ConCreTe-del-D8.7-validations-draft-v1.0
Lead partner short name:	Universidad Complutense de Madrid (UCM)
Report version:	1.0, final
Report preparation date:	13/09/2016
Dissemination level:	PU
Nature:	R = report
Lead author:	Pablo Gervás
Co-authors:	Alberto Díaz, Antonio F. G. Sevilla, Alberto Fernández-Isabel, Anna Kantosalo, Gonzalo Méndez, Pedro Martins, Alexandre Miguel Pinto, Senja Pollak, Matthew Purver, Hannu Toivonen, Martin Žnidaršič, Hugo Gonçalo Oliveira
Status:	Final

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

Change log

Date	Editor	Summary of changes made
13/09/2016	Alberto Díaz	First draft
13/09/2016	Pablo Gervás	Outline, with space for various contributions
20/09/2016	Pablo Gervás	Minimal elaboration of partner contributions, in search for a structuring thread
20/09/2016	Hannu Toivonen	Update of University of Helsinki collaborations
27/09/2016	Hugo Gonçalo Oliveira	Update of University of Coimbra collaborations (evaluation meme generation and poetry)
28/09/2016	Antonio F. G. Sevilla	Update of UCM contribution
28/09/2016	Pedro Martins	Update of UC contribution
28/09/2016	Matthew Purver	Update of QMUL contributions in geometric concept modelling
29/09/2016	Alberto Fernández Isabel	Update of UCM contribution
29/09/2016	Alberto Díaz	Review of UCM contribution
29/09/2016	Gonzalo Méndez	Review and update of UCM contribution
29/09/2016	Pablo Gervás	Final re-structuring, elaboration of introduction and conclusions

Contents

1 Introduction

This document reports on the validation procedures employed for different prototype applications developed in Work Package 5.

The document is structured as follows:

- In each section, contributions in the form of papers (published, submitted, thesis, ...) are listed.
- The actual papers are included at the end of the deliverable in the order in which they are mentioned in the body.

2 Document Summaries

We have developed novel methods to create a short text as a summary of a given (set of) document(s) [11, 12]. The summaries were validated as follows.

We have used the ROUGE [15] evaluation method for evaluating summaries. The ROUGE method uses the overlap of n-grams between model summaries, written by humans, and generated summaries to measure the similarity. For instance, ROUGE-1 score just looks at unigrams, ROUGE-2 score looks at 2-grams and ROUGE-L looks for the longest common sequence between two texts. The ROUGE score breaks down into two components, precision and recall. For evaluation we used the combined score, F-measure, computed as the harmonic mean between precision and recall.

The developed methods are largely language-independent, so validation of the quality of the summaries also needs to be done in multiple languages.

We used the MultiLing-2013 [5] dataset to evaluate our methods. The dataset contains documents in 10 different languages – English, French, Chinese, Romanian, Spanish, Hindi, Arabic, Hebrew, Greek and Czech. Our method assumes that the text has been (or can trivially be) broken to words. Since this assumption does not hold for Chinese, we omitted it from our experiments. MultiLing contains 15 topics for each language except for French and Hindi, for which the number of topics is 10. On average each topic consists of 10 documents which need to be collectively summarized into a text of 250 words.

In the evaluation, the proposed method outperformed all methods that participated MultiLing: it ranked first in six languages out of nine, and was among the best ones in the remaining three. A statistical analysis shows that it is significantly better than the other methods. This is a striking result given that the method was applied to the nine different languages without any changes between languages.

However, the summarization problem will still require significant further work before it is solved. The coherence and fluency of generated summaries is an issue especially for methods based on sentence selection, such as ours. Further work is needed in making summaries better in these respects.

The first publication [11] contains a preliminary evaluation using ROUGE, the later publication [12] a full multi-lingual evaluation using ROUGE on the MultiLing-2013 dataset,

- Oskar Gross, Antoine Doucet, and Hannu Toivonen. Document summarization based on word associations. In *The 37th Annual ACM SIGIR Conference*, pages 1023–1026, Gold Coast, Australia, 2014. [11]
- Oskar Gross, Antoine Doucet, and Hannu Toivonen. Language-independent multi-document text summarization with document specific word associations, 2016. [12]

Additionally, similar summarization experiments and evaluation were done based on concept graphs, resulting in an additional publication:

- Antonio F. G. Sevilla, Alberto Fernández-Isabel, and Alberto Díaz. Enriched semantic graphs for extractive text summarization. In *Proceedings of Conference of the Spanish Association for Artificial Intelligence (CAEPIA)*, pages 217–226, Salamanca, Spain, September 2016. Springer. [23]

3 Co-creative Poetry Writing

We have developed an interactive tool, Poetry Machine, for writing poetry co-creatively between a human and a computer program. Since the goal of the system is not to produce poems autonomously, but rather to help users write poems, the target of evaluation is not the poems but the system and its support of human-computer co-creation. Little methodology exists for evaluation of human-computer co-creative systems, so we have developed methodology for the purpose and applied it to our Poetry Machine.

In the paper where we develop the evaluation methodology [14], we first briefly discuss the similarities and differences between human-computer co-creativity evaluation and computational creativity evaluation in more general. We then proceed to view Interaction Design in the context of computational creativity: We see how Interaction Design currently connects to computational creativity and view previous human-computer co-creation and creativity support system evaluation projects in the light of the DECIDE framework [22].

In the paper we then illustrate how the evaluation methodology can be applied to practical computational creativity development work by providing a list of gathered user ideas and presenting concrete ideas on how to use them for further development of the Poetry Engine.

We learned a number of lessons from the evaluation, not only about the Poetry Machine but also about human-computer co-creation in more general. For instance, it seems that some traditionally used interaction design evaluation measures, such as time, or facial gestures are not useful within a creative context, as some negative signs, such as frowning, may actually indicate positive aspects, such as concentration or immersion instead. Most of the issues related to human-computer co-creativity testing with interaction design evaluation methods still seem to be concerned with typical interaction design evaluation problems, such as selecting suitable users.

- Anna Kantosalo, Jukka M Toivanen, and Hannu Toivonen. Interaction evaluation for human-computer co-creativity: A case study. In *Proceedings of the Sixth International Conference on Computational Creativity*, pages 276–283, 2015. [14]

4 Lexical Replacement Humor

We have investigated lexical replacement humor, a minimal creative activity. Given a short text, such as an SMS message, the task is to substitute one word such that the created text is funny [25]. This is a form of mutational creativity where the new concept (humorous text) is on the surface very similar to an existing text (the given, non-humorous text) but gives quite a different expression. We next describe how we evaluated different techniques (constraints) for word substitution in this task. The details are in the respective publication [25].

When designing the evaluation we made two central decisions. The first one was to use real SMS messages as the texts to be modified using the proposed constraints. This choice reflects our desire to make the test setting as realistic as possible. Second, we crowdsourced the assessment of humor responses to a large number of independent subjects, for a large and unbiased evaluation. We are aware of only one previous use of crowdsourcing for assessment of humorous texts [13]. Unlike our work, it was performed on texts collected from the web and not produced automatically.

As input texts, we employed *NUS SMS Corpus*[1], a collection of 10116 real SMS messages [2].

[1] `http://wing.comp.nus.edu.sg/SMSCorpus`

The texts make heavy use of abbreviations and colloquial expressions, providing a challenging but realistic case for linguistic processing.

Based on preliminary tests, we identified a sample size of 75 texts per each of the 12 experimental conditions (combinations of constraints) as a suitable compromise between statistical power and crowdsourcing expense. We asked the subjects to assess individual messages for their funniness on a scale from 0 to 4.

Each subject was assigned a random set of 20 messages out of the total of 787 messages. Some subjects chose to evaluate several such sets; the total number of subjects evaluating the messages was 524. The crowdsourcing evaluation was running until each message was judged by at least 90 different subjects. After removing likely scammers (see the paper for details), we had a total of 70,848 assessments of messages.

The main findings are the following. (1) Each lexical constraint considered contributes to the humorous effect with a different weight. (2) The effects of the constraints are cumulative and provide first evidence of a compositional nature of the humorous effect in the studied context. (3) There are combinations of constraints that support each other and amplify their individual contribution.

The results suggest that fully automated production of humorous text can be feasible. Even if the practical value of SMS modification is questionable, we consider the empirical results significant for proving the potential of automated humor generation.

To the best of our knowledge, this is the first time that an evaluation of this scale and detail has been performed on machine-generated humorous texts, and with positive results.

- Alessandro Valitutti, Antoine Doucet, Jukka M. Toivanen, and Hannu Toivonen. Computational generation and dissection of lexical replacement humor. *Natural Language Engineering*, 22(5):727–749, 2016. [25]

5 Meme Generation

Internet memes were automatically produced for nine news headlines. For the same headlines, we asked humans to produce their own memes. Human judges were then presented with a headline and a set of memes that they would have to score according to four aspects – syntactic coherence, text suitability to the macro, surprise, humor value – using a 5-point Likert scale. In the end, we got 312 scores for human-created memes and 156 for the automatically-created. The median scores for the automatically-created memes were 4, 3, 3, 3, respectively for syntactic coherence, text suitability to the macro, surprise, humor value. This was 1 point below the humans in the coherence and humor aspect, the same as all the humans in the surprise, and the same as one human but 1 point below the others in the suitability aspect.

This evaluation is further described in the following paper, together with the meme generation procedure:

- Hugo Gonçalo Oliveira, Diogo Costa, and Alexandre Miguel Pinto. One does not simply produce funny memes! – explorations on the automatic generation of internet humor. In *Proceedings of 7th International Conference on Computational Creativity*, ICCC 2016, Paris, France, 2016. [8]

6 Rhetorical Figure Generation

We have carried out an evaluation to measure the quality of the rhetorical figure generator using word associations. The aim of this evaluation has been twofold. On the one hand, we intended to test the appropriateness of the analogies, similes and metaphors generated by our system, in order for us to be able to refine the process followed to generate them. On the other hand, we

also expected to find out what kind of rhetorical figure is more enlightening for the evaluators and which one is closer to a rhetorical figure generated by humans.

The evaluation was carried out as an online survey using Google Forms, where each evaluator received a link to one of the three surveys and was asked to score each of the figures using a Likert scale. Evaluators were asked to rate how appropriate or natural sounding each trope was, giving them a score from 1 to 7 (where 1 symbolizes a completely inappropriate trope and 7 represents a completely natural sounding trope).

In order to have two different baselines in our experiment to measure the quality of the figures generated by our system, we have used a set of commonly accepted rhetorical figures, together with a set of random manually generated ones, to compare them against the ones generated by our system.

In all cases the randomly generated metaphors are rated as meaningless by the evaluators. In contrast, commonly accepted metaphors get the highest results, with a slight preference for the metaphors created using abstract concepts over the ones that are based on the use of concrete concepts. The automatically generated metaphors using concepts of different categories are also poorly rated, which points out that sharing only one property is not enough to generate a good metaphor. For the generated metaphors using concepts that belong to the same category, the difference that exists between the modes of the metaphors that use concrete and abstract concepts is remarkable. This suggests that abstract metaphors are more evocative and offer a wider range of interpretations than concrete ones. Finally, the overall median for the metaphors also suggests that more aspects need to be taken into consideration to increase the perceived quality of these rhetorical figures.

This evaluation is the result of an ongoing masters thesis, expected to be finished by the end of 2016, and is further described in a conference paper:

- Paloma Galvan, Virginia Francisco, Raquel Hervas, Gonzalo Mendez, and Pablo Gervas. Exploring the role of word associations in the construction of rhetorical figures. In *7th International Conference on Computational Creativity (ICCC 2016)*, Paris, France, 06/2016 2016. [4]

7 Riddle Generation

We have carried out an evaluation to test whether word associations obtained by our system provided useful information for riddle generation, and to assess the quality of the resulting riddles. In order to do that, human evaluators were asked to guess the initial concepts which were used to create the riddles. Then, we studied the rate of success obtained by the evaluators, while at the same time analyzing how many comparisons were required to obtain the correct answers in different riddles. Some issues related to ambiguity and contradiction appeared when creating the riddles, so we decided to create two different sets of riddles to perform the evaluation.

Ten riddles were presented to human evaluators to see if they were able to find the initial target concepts. Riddles were presented in four phases, in order to know how many comparisons were needed to solve the riddle. In the first phase a single comparison was presented, in the second phase two comparisons were presented, three comparisons in the third phase and, finally, four comparisons in the fourth phase. The evaluation was carried out using Google Forms and some personal information was collected for statistical purposes (age, gender and riddle ability).

In order to carry out a more detailed evaluation, we decided to create two different riddle sets. Using the same ten concepts, but with some differences in the provided comparisons, we created an original and a curated version of the riddles. For the first set, the resulting comparisons were randomly selected. For the other set, the four most significant comparisons were manually selected among seven generated using the described process in order to avoid not valid comparisons due to polysemy or semantic contradictions.

The results of the curated version of the riddles are significantly better than the ones of the random version. So, it is evident that a special selection of comparisons is needed in some cases.

Regarding the number of comparisons needed to guess the correct answer, with just a single comparison, there is almost no chance of guessing the target concept. In most cases, people answer at random because there are lots of concepts that share the presented attribute. When providing two comparisons, users are able to multiply by four the number of correct guesses. When they are provided three comparisons, in the case of randomly chosen comparisons, they reach their maximum rate of success. In the case of manually selected comparisons, they guess 55% of riddles, which is almost the maximum success, because the difference with the last phase, where four comparisons are provided, is almost negligible.

This evaluation is the result of an ongoing masters thesis, expected to be finished by the end of 2016, and is further described in a conference paper:

- Paloma Galvan, Virginia Francisco, Raquel Hervas, and Gonzalo Mendez. Riddle generation using word associations. In *Language Resources and Evaluation Conference (LREC 2016)*, Portoroz, Slovenia, 05/2016 2016. [3]

8 Information Extraction

In order to retrieve knowledge from text, it must first be analysed to find the relevant details, and the nature of the language used can greatly impact the quality of the extracted information. We have compared triplets that represent definitions or properties of concepts obtained from three online collaborative resources (English Wikipedia, Simple English Wikipedia and Simple English Wiktionary) and study the differences in the results when Basic English is used instead of common English.

We have evaluated the quality of the triplets extracted from text by assigning them a value based on how strongly related its property is to the concept and how well it respects the relation. The possible values are: 1, when the triplets correctly represent an IS_A or IS relation in which the property defines or is very strongly related to the concept; 0.5, when the property is a less accurate or informative definition of the concept, or when it represents a feature or quality of the concept; and 0, for triplets with properties which are related to the concept but do not respect the relation or which are unrelated to the concept.

The evaluation has been performed manually by human annotators. The final statistics were obtained by using the average of the score given by all of the annotators, following an inter-annotator agreement using a popular metric, Fleiss Kappa. This allows us to know the degree of agreement between the annotators.

This evaluation is further described as a result of two masters theses and has been published in a conference paper:

- Adrian Rabadan. Generacion de lenguaje natural a partir de grafos semanticos. Master's thesis, Computer Science School, Universidad Complutense de Madrid, september 2016. [19]
- Teresa Rodriguez-Ferreira. Content filtering and enrichment using triplets for text generation. Master's thesis, Computer Science School, Universidad Complutense de Madrid, june 2016. [20]
- Teresa Rodriguez-Ferreira, Adrian Rabadan, Raquel Hervas, and Alberto Diaz. Improving information extraction from wikipedia texts using basic english. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may 2016. European Language Resources Association (ELRA). [21]

9 Poetry Generation

Several aspects of Portuguese song lyrics produced by Tra-la-Lyrics [10] and by a specific instantiation of PoeTryMe [7] were assessed. More precisely, two strategies of Tra-la-Lyrics were

tested – RR that only tries to match the text with the rhythm and have occasional rhymes; and GG where, besides the previous, the text follows a coherent syntactic structure and a seed is used here and there, when possible. PoeTryMe was used with a generate & test strategy that considered not only the presence of rhymes and the length of the lines, but also how the rhythm was matched. Moreover, as in other instantiations, lines were produced by filling blanks from a grammar of lines with semantically-related words.

First, an automatic evaluation was conducted, based on lyrics that were generated for 6 different songs with 2/4, 3/4 and 4/4 metres. This included 30 lyrics generated by RR, 180 by GG and 180 by PoeTryMe. For the latter two, lyrics were generated with 6 different seeds (6×30).

Considering that the strongest beats of the melodies should match there were slightly more stressed syllables matching strong beats in RR and GG ($\approx$80%), this was still the trend with PoeTryMe ($>$70%). Rhymes were more frequent in RR – 60 to 72% of the lines ended in rhyme – but, in general, PoeTryMe had a higher rhyme ratio than GG – 11-38% against 28-52%.

The semantics of the lyrics by each strategy was validated by measuring the average Pointwise Mutual Information (PMI) [24] between every content word used and, for GG and PoeTryMe, also between every content word and the seed. The PMI for each pair of words was computed based on the co-occurrence of the words in the articles of the Portuguese Wikipedia.

We noticed that the PMI depends highly on the seed used, but, as expected, it is clearly higher for PoeTryMe that for GG, which means that the words used by the former are more semantically similar that those used by the latter.

In addition to the automatic evaluation, the generated song lyrics were uploaded to a crowdsourcing platform, together with the music for which they were generated for, in order to collect human judgements on six aspects – rhythm, rhymes, sound, grammar, meaning, subject – and general quality. Surprisingly, Tra-la-Lyrics got the best scores in all the aspects, though some were really close to the other strategies. More precisely, in a 5-point Likert scale, it got a median score of 3 in all the aspects, while RR and GG also got 3 in all but grammar and meaning, where they got 2. The judges could also select the subject of the lyrics from a list with all the seeds. Only 20% of the judges got the correct subject for PoeTryMe, but this number was lower (14%) for GG.

The previous work is described in the following journal article:

- Hugo Gonçalo Oliveira. Tra-la-lyrics 2.0: Automatic generation of song lyrics on a semantic domain. *Journal of Artificial General Intelligence*, 6(1):87–110, December 2015. Special Issue: Computational Creativity, Concept Invention, and General Intelligence. [6]

Approaches to the automatic validation of computer generated poetry were also explored while using PoeTryMe for producing poetry in three different languages – Portuguese, Spanish and English. Ten seed words were first selected and translated to the three languages, in order to generate nine poems for each combination of language and seed. From those, the average number of syllables per line, rhymes per line, and distinct words were noted down a compared.

Moreover, in order to analyse the structural variation of the lines of poems and also of different poems generated for the same language, ROUGE-based [16] metrics were used to compute the overlapping n-grams, skip-gram, and the longest common subsequence.

Finally, to analyse the topicality of each poem and its association to the original seeds, PMI was again used. This time, we relied on the average PMI between each poem and each seed to check whether we could use it to automatically link the poems with their original seed.

This work is described in the following article, submitted to the Journal of Natural Language Engineering:

- Hugo Gonçalo Oliveira, Raquel Hervás, Alberto Díaz, and Pablo Gervás. Multilingual extension and evaluation of a poetry generator. *Submitted to the Journal of Natural Language Engineering*, 2016. [9]

10 Geometric concept modelling

The concepts produced by a geometric, distributional vector-space model were validated directly in terms of their ability to match lexical concepts in WordNet, showing improved recall over standard lexical embedding models; this has already been reported in Deliverable D3.1 (and since been accepted and published as [17]):

- Stephen McGregor, Kat Agres, Matthew Purver, and Geraint A. Wiggins. From distributional semantics to conceptual spaces: A novel computational method for concept creation. *Journal of Artificial General Intelligence*, 6(1):55–86, December 2015 [17]

Moving beyond this intrinsic evaluation, recent efforts have moved to more extrinsic evaluations, examining how the concepts produced by (variants of) this model fare when used in other tasks. We have examined two uses. First, we used the concept model in poetry generation, to generate candidate vocabulary for a poem, with the results judged by humans on criteria of creativity, quality and meaningfulness; here, the model enabled a comparison of the effect of text descriptions paired with the poems, with some preference shown for descriptions which characterise the generation process in objective terms [18]. Second, we looked at whether this concept modelling approach is suitable for metaphor modelling, by examining the degree to which the conceptual spaces created by the model explain human ratings of metaphoricity of expressions; the model performs significantly better than state-of-the-art distributional lexical models [1]:

- Stephen McGregor, Matthew Purver, and Geraint Wiggins. Process based evaluation of computer generated poetry. In *Proceedings of the INLG 2016 Workshop on Computational Creativity in Natural Language Generation*, pages 51–60, Edinburgh, UK, September 2016. Association for Computational Linguistics [18]
- Kathleen R. Agres, Stephen McGregor, Karolina Rataj, Matthew Purver, and Geraint Wiggins. Modeling metaphor perception with distributional semantics vector space models. In *Proceedings of the ESSLLI Workshop on Computational Creativity, Concept Invention, and General Intelligence (C3GI)*, Bolzano-Bozen, Italy, August 2016 [1]

11 Conclusion

The research carried out over the ConCreTe project has resulted in a significant number of prototypes and applications. Many of these have been described in other deliverables of the project. The present deliverable covers the different approaches to validation of these prototypes that have been employed throughout the project. These include very diverse solutions such as crowsourcing – used for poetry and lexical replacement humour –, automated metrics – used for TraLaLa lyrics and PoeTryMe –, online surveys – used for metaphor interpretation and riddle generation –, comparison with human performance – used for meme generation –, traditional recall measures over standard solutions - used for geometric concept modelling –, and specific evaluation methods based on interaction design – used for co-creativity. This broad range of approaches indicates there are many aspects that require validation in the set of applications that are being considered. This is taken to be an indication that significant research is still needed in the future on the evaluation of applications of this type.

References

[1] Kathleen R. Agres, Stephen McGregor, Karolina Rataj, Matthew Purver, and Geraint Wiggins. Modeling metaphor perception with distributional semantics vector space models. In *Proceedings of the ESSLLI Workshop on Computational Creativity, Concept Invention, and General Intelligence (C3GI)*, Bolzano-Bozen, Italy, August 2016.

[2] Tao Chen and Min-Yen Kan. Creating a live, public short message service corpus: the NUS SMS Corpus. *Language Resources and Evaluation*, 74(2):299–335, 2013.

[3] Paloma Galvan, Virginia Francisco, Raquel Hervas, and Gonzalo Mendez. Riddle generation using word associations. In *Language Resources and Evaluation Conference (LREC 2016)*, Portoroz, Slovenia, 05/2016 2016.

[4] Paloma Galvan, Virginia Francisco, Raquel Hervas, Gonzalo Mendez, and Pablo Gervas. Exploring the role of word associations in the construction of rhetorical figures. In *7th International Conference on Computational Creativity (ICCC 2016)*, Paris, France, 06/2016 2016.

[5] George Giannakopoulos. Multi-document multilingual summarization and evaluation tracks in acl 2013 multiling workshop. In *Proceedings of the MultiLing 2013 Workshop on Multilingual Multi-document Summarization*, pages 20–28, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[6] Hugo Gonçalo Oliveira. Tra-la-lyrics 2.0: Automatic generation of song lyrics on a semantic domain. *Journal of Artificial General Intelligence*, 6(1):87–110, December 2015. Special Issue: Computational Creativity, Concept Invention, and General Intelligence.

[7] Hugo Gonçalo Oliveira and Amílcar Cardoso. Poetry generation with PoeTryMe. In T. R. Besold, M. Schorlemmer, and A. Smaill, editors, *Computational Creativity Research: Towards Creative Machines*, Atlantis Thinking Machines, chapter 12, pages 243–266. Atlantis-Springer, 2015.

[8] Hugo Gonçalo Oliveira, Diogo Costa, and Alexandre Miguel Pinto. One does not simply produce funny memes! – explorations on the automatic generation of internet humor. In *Proceedings of 7th International Conference on Computational Creativity*, ICCC 2016, Paris, France, 2016.

[9] Hugo Gonçalo Oliveira, Raquel Hervás, Alberto Díaz, and Pablo Gervás. Multilingual extension and evaluation of a poetry generator. *Submitted to the Journal of Natural Language Engineering*, 2016.

[10] Hugo R. Gonçalo Oliveira, F. Amílcar Cardoso, and Francisco C. Pereira. Tra-la-Lyrics: an approach to generate text based on rhythm. In *Proceedings of the 4th International Joint Workshop on Computational Creativity*, pages 47–55, London, UK, June 2007. IJWCC 2007.

[11] Oskar Gross, Antoine Doucet, and Hannu Toivonen. Document summarization based on word associations. In *The 37th Annual ACM SIGIR Conference*, pages 1023–1026, Gold Coast, Australia, 2014.

[12] Oskar Gross, Antoine Doucet, and Hannu Toivonen. Language-independent multi-document text summarization with document specific word associations, 2016.

[13] C.F. Hempelmann, J.M. Taylor, and V. Raskin. Tightening up joke structure: Not by length alone. In *Proceedings of the 34th Annual Meeting of the Cognitive Science Society 2012 (CogSci 2012)*, Sapporo, Japan, 2012.

[14] Anna Kantosalo, Jukka M Toivanen, and Hannu Toivonen. Interaction evaluation for human-computer co-creativity: A case study. In *Proceedings of the Sixth International Conference on Computational Creativity*, pages 276–283, 2015.

[15] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, pages 25–26, Barcelona, Spain, July 2004.

[16] Chin-Yew Lin and Franz Josef Och. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL 2004. ACL Press, 2004.

[17] Stephen McGregor, Kat Agres, Matthew Purver, and Geraint A. Wiggins. From distributional semantics to conceptual spaces: A novel computational method for concept creation. *Journal of Artificial General Intelligence*, 6(1):55–86, December 2015.

[18] Stephen McGregor, Matthew Purver, and Geraint Wiggins. Process based evaluation of computer generated poetry. In *Proceedings of the INLG 2016 Workshop on Computational Creativity in Natural Language Generation*, pages 51–60, Edinburgh, UK, September 2016. Association for Computational Linguistics.

[19] Adrian Rabadan. Generacion de lenguaje natural a partir de grafos semanticos. Master's thesis, Computer Science School, Universidad Complutense de Madrid, september 2016.

[20] Teresa Rodriguez-Ferreira. Content filtering and enrichment using triplets for text generation. Master's thesis, Computer Science School, Universidad Complutense de Madrid, june 2016.

[21] Teresa Rodriguez-Ferreira, Adrian Rabadan, Raquel Hervas, and Alberto Diaz. Improving information extraction from wikipedia texts using basic english. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may 2016. European Language Resources Association (ELRA).

[22] Y. Rogers, H. Sharp, and J. Preece. *Interaction Design: Beyond Human Computer Interaction.* Wiley, 3rd edition, 2011.

[23] Antonio F. G. Sevilla, Alberto Fernández-Isabel, and Alberto Díaz. Enriched semantic graphs for extractive text summarization. In *Proceedings of Conference of the Spanish Association for Artificial Intelligence (CAEPIA)*, pages 217–226, Salamanca, Spain, September 2016. Springer.

[24] Peter D. Turney. Mining the web for synonyms: PMI–IR versus LSA on TOEFL. In *Proceedings of 12th European Conference on Machine Learning, ECML 2001*, volume 2167 of *LNCS*, pages 491–502, Freiburg, Germany, 2001. Springer.

[25] Alessandro Valitutti, Antoine Doucet, Jukka M. Toivanen, and Hannu Toivonen. Computational generation and dissection of lexical replacement humor. *Natural Language Engineering*, 22(5):727–749, 2016.

Compilation of published papers

The compilation of all the papers published so far as described in this deliverable follows. In some cases, the final versions could not be included in a public deliverable due to copyright restrictions. In those cases, a pre-review version of the corresponding papers has been included. Links to the final version in the publisher's web site have been provided in the corresponding section of the deliverable, though payment may be required to view them. Every effort will be made to purchase open access for these papers using project funds, though this process still has to be put in motion.

Papers are presented in alphabetical order of the surname of the first author.

- Kathleen R. Agres, Stephen McGregor, Karolina Rataj, Matthew Purver, and Geraint Wiggins. Modeling metaphor perception with distributional semantics vector space models. In *Proceedings of the ESSLLI Workshop on Computational Creativity, Concept Invention, and General Intelligence (C3GI)*, Bolzano-Bozen, Italy, August 2016.
  http://www.eecs.qmul.ac.uk/~mpurver/papers/agres-et-al16c3gi.pdf.

- Paloma Galvan, Virginia Francisco, Raquel Hervas, Gonzalo Mendez, and Pablo Gervas. Exploring the role of word associations in the construction of rhetorical figures. In *7th International Conference on Computational Creativity (ICCC 2016)*, Paris, France, 06/2016 2016.
  http://nil.fdi.ucm.es/sites/default/files/GalvanEtAl.pdf.

- Paloma Galvan, Virginia Francisco, Raquel Hervas, and Gonzalo Mendez. Riddle generation using word associations. In *Language Resources and Evaluation Conference (LREC 2016)*, Portoroz, Slovenia, 05/2016 2016.
  http://nil.fdi.ucm.es/sites/default/files/riddlesLREC.pdf.

- Hugo Gonçalo Oliveira, Diogo Costa, and Alexandre Miguel Pinto. One does not simply produce funny memes! – explorations on the automatic generation of internet humor. In *Proceedings of 7th International Conference on Computational Creativity*, ICCC 2016, Paris, France, 2016.
  http://www.computationalcreativity.net/iccc2016/wp-content/uploads/2016/01/-One-does-not-simply-produce-funny-memes.pdf

- Hugo Gonçalo Oliveira. Tra-la-lyrics 2.0: Automatic generation of song lyrics on a semantic domain. *Journal of Artificial General Intelligence*, 6(1):87–110, December 2015. Special Issue: Computational Creativity, Concept Invention, and General Intelligence.
  https://www.degruyter.com/downloadpdf/j/jagi.2015.6.issue-1/jagi-2015-0005/-jagi-2015-0005.xml

- Oskar Gross, Antoine Doucet, and Hannu Toivonen. Document summarization based on word associations. In *The 37th Annual ACM SIGIR Conference*, pages 1023–1026, Gold Coast, Australia, 2014.
  https://www.cs.helsinki.fi/u/htoivone/pubs/document-summarization_SIGIR_2014.pdf

- Oskar Gross, Antoine Doucet, and Hannu Toivonen. Language-independent multi-document text summarization with document specific word associations, 2016.
  https://www.cs.helsinki.fi/u/htoivone/pubs/summarization_SAC_2016.pdf

- Anna Kantosalo, Jukka M Toivanen, and Hannu Toivonen. Interaction evaluation for human-computer co-creativity: A case study. In *Proceedings of the Sixth International Conference on Computational Creativity*, pages 276–283, 2015.
  http://computationalcreativity.net/iccc2015/proceedings/13_2Kantosalo.pdf

- Stephen McGregor, Kat Agres, Matthew Purver, and Geraint A. Wiggins. From distributional semantics to conceptual spaces: A novel computational method for concept creation. *Journal of Artificial General Intelligence*, 6(1):55–86, December 2015. http://dx.doi.org/10.1515/jagi-2015-0004.
- Stephen McGregor, Matthew Purver, and Geraint Wiggins. Process based evaluation of computer generated poetry. In *Proceedings of the INLG 2016 Workshop on Computational Creativity in Natural Language Generation*, pages 51–60, Edinburgh, UK, September 2016. Association for Computational Linguistics. http://www.eecs.qmul.ac.uk/~mpurver/papers/mcgregor-et-al16ccnlg.pdf.
- Adrian Rabadan. Generacion de lenguaje natural a partir de grafos semanticos. Master's thesis, Computer Science School, Universidad Complutense de Madrid, september 2016. http://nil.fdi.ucm.es/sites/default/files/20160926-TFMAdrian-Final.pdf.
- Teresa Rodriguez-Ferreira, Adrian Rabadan, Raquel Hervas, and Alberto Diaz. Improving information extraction from wikipedia texts using basic english. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may 2016. European Language Resources Association (ELRA). http://nil.fdi.ucm.es/sites/default/files/LRECPaper-AdrianTeresa.pdf.
- Teresa Rodriguez-Ferreira. Content filtering and enrichment using triplets for text generation. Master's thesis, Computer Science School, Universidad Complutense de Madrid, june 2016. http://nil.fdi.ucm.es/sites/default/files/TFM-TeresaRodriguez.pdf.
- Antonio F. G. Sevilla, Alberto Fernández-Isabel, and Alberto Díaz. Enriched semantic graphs for extractive text summarization. In *Proceedings of Conference of the Spanish Association for Artificial Intelligence (CAEPIA)*, pages 217–226, Salamanca, Spain, September 2016. Springer. http://dx.doi.org/10.1007/978-3-319-44636-3_20.
- Alessandro Valitutti, Antoine Doucet, Jukka M. Toivanen, and Hannu Toivonen. Computational generation and dissection of lexical replacement humor. *Natural Language Engineering*, 22(5):727–749, 2016. https://www.cs.helsinki.fi/u/doucet/papers/JNLE2015.pdf