Lang Resources & Evaluation
https://doi.org/10.1007/s10579-021-09560-6

ORIGINAL PAPER

Pragmatic evaluations of automated linguistic creativity

Lorenzo Gatti[1] · **Oliviero Stock**[2] · **Carlo Strapparava**[2] · **Gözde Özbal**[2]

Accepted: 2 September 2021

Abstract The optimal innovation hypothesis (OIH) offers a good aesthetic and cognitive reference for the kind of linguistic creativity where minimal variations can have a strong effect on the audience and realize an overall intended goal. The same approach can be the basis for creative systems. The question is how to concretely evaluate not only their quality, but also their pragmatic effect. This paper describes the original evaluations of two systems based on the OIH, one automatically yielding witty headlines for incoming news, the other producing song parodies, varying the song lyrics to evoke a given concept. The goal is to bring attention to the importance of evaluating the pragmatic potential of creative systems, in addition to the quality of their output, and to demonstrate how such evaluations can be done.

Keywords Computational creativity · Natural Language processing applications · Evaluation · Pragmatics

✉ Lorenzo Gatti
l.gatti@utwente.nl

Oliviero Stock
stock@fbk.eu

Carlo Strapparava
strappa@fbk.eu

Gözde Özbal
gozbalde@gmail.com

[1] Human-Media Interaction lab, University of Twente, Enschede, The Netherlands

[2] FBK-irst, Trento, Italy

Published online: 14 September 2021

1 Introduction

Computational creativity is coming of age as the sub-field of AI research concerned with the development of programs that generate creative output and, by extension, that can show intelligent creative behavior (Colton et al., 2009). As for linguistic computational creativity, the literature includes, among the various systems and approaches, tools for storytelling (Pérez & Sharples, 2004), for generating poems (for a recent discussion on the different approaches to poetry generation see Lamb et al., 2017), metaphors (Veale, 2016), riddles and funny acronyms (Ritchie et al., 2006; Stock & Strapparava, 2003), storylines (Veale, 2014) or inspirational sentences (Özbal et al., 2013); some programs are even trying to model creative activities with a clear commercial value, such as creative naming for new products (Özbal & Strapparava, 2013), or the generation of slogans for companies and products (Tomašic et al., 2014). Most systems make use of state-of-the-art computational linguistics techniques, such as sentiment analysis, word embeddings for semantic similarity and language models based on dependency relations. The real issues are the aesthetic quality of the output (Machado & Cardoso, 1998), and—at least for systems concerned with potential applied settings—pragmatics, i.e. the goal the system intends to achieve with the produced linguistic material.

The optimal innovation hypothesis (OIH) is a theory that aims at explaining the pleasurability of stimuli based on variations of familiar material (Giora et al., 2004). Far from proposing a general theory, the OIH is particularly useful as an aesthetic and cognitive reference for the kind of linguistic creativity in which subtle variations can have a strong effect on the audience (as it happens, for example, when the lyrics of a song are repurposed to promote a concept or advertise a product).

Computational systems can also be based on this approach, with the aim of producing pleasurable output with pragmatic effects. By evaluating the aesthetic quality of machine productions, and their pragmatic effects, we can determine if computational systems can indeed have a practical potential. Yet evaluating computational creativity for its effectiveness is not a trivial task and requires original ideas. For creative systems, evaluation is not simply a matter of gold-standards and precision measures: specific experimental designs are needed to measure the output quality and its effects on readers.

The goal of this paper is to draw attention on the evaluation of pragmatic effects of such systems. We will briefly recall two different creative systems we designed, and described previously, that are inspired by the OIH, and demonstrate through their evaluations how computational systems based on this theory can produce creative linguistic output with strong pragmatic potential. The two systems have different pragmatic goals and their evaluations had to be different, but they realize the same underlying methodology. The presented methodology is novel, certainly for the subfield of computational creativity and NLP, in that it aims at measuring the specific pragmatic effects of the systems in an extrinsic way, and at comparing them with human baselines. They speak in terms of cognitive concepts, such as attention, memory, recall, as effects of the application of aesthetic principles. Many creative

and non-creative systems with pragmatic goals, instead, do not focus their evaluation on the effects of the produced output, but just on its intrinsic properties [e.g. measuring BLEU scores (Munigala et al., 2018), or collecting ratings for the system without comparing them to ratings of human producers (Clark et al., 2018)]. With this work we argue that these measures are necessary but not sufficient for a complete evaluation of the systems.

Furthermore, by giving positive evidence of the effectiveness of computational linguistic creativity, this approach can contribute to giving credibility to AI systems in important and rich applied sectors such as adaptive promotion, online journalism and advertising. In fact, systems based on slight variations of familiar expressions are easily adaptable to groups and individuals: you just need to change the set of expressions considered familiar to the target. Personalized, creative and evocative expressions go beyond the possibility of production by professional creative humans.

The paper is organized as follows. In Sect. 2 we introduce the concept of optimal innovation. Section 3 presents two computational systems based on this theory: Heady-Lines, a program for generating witty headlines for the news of the day, and Mockingbird, a system for generating and singing song parodies based on a piece of news. The evaluations of these systems, with their particular focus on pragmatics, are described in Sect. 4, while Sect. 5 compares them with recent works with similar aims, and gives an overview of different proposed methodologies for evaluating creative systems. The conclusions, in Sect. 6, include some considerations about the application of such systems when combined with a model of the "target" of the message.

2 Optimal innovation

During the 2008 Super Bowl, German car maker Audi aired a commercial that spoofed the famous scene from "The Godfather" with the bloody horse head in the bed, the head being replaced by the front of a car leaking oil on the bed sheets. "Indiana Bones and the Temple of Groom" is the name of an existing pet care center, while "Raiders of the Lost Bark" is the title of a surprisingly accurate remake of the Indiana Jones movie, starring a canine archaeologist[1].

These successful creative items have something in common: they are all modifications of something very famous, for instance a movie title or a picture. There is more: they are instances of optimal innovation, i.e., novel creations that variate a known theme or item and remind us of something that we already know. This mechanism is at the basis of mashups, parodies, puns, and much more.

According to the original formulation of Giora et al. (2004), an optimally innovative stimulus involves:

– A novel—less-salient or non-salient—response to a given stimulus.

[1] https://www.youtube.com/watch?v=7ydBgdL5R08.

– At the same time, allows for the automatic recoverability of a salient response related to that stimulus so that both responses make sense.

Salient responses, here, are those coded in the mental lexicon of the perceiver. They are readily accessible thanks to factors such as prototypicality, familiarity, frequency, conventionality, but also "defaultness" (Giora et al., 2017).

The OIH tries to explain why these modifications are so effective and, hence, so common; according to this theory, pleasurability is a function of novelty. In particular:

– Items that are familiar are pleasurable;
– Optimal innovations, however, are much more pleasurable;
– Items that are too novel ("pure innovation") are the least pleasurable.

Optimal innovation thus manages to strike a balance between the clichés of familiar expressions, and the complete novelty we are not prepared for, and for which we might still need to "develop a taste" through repeated exposure.

The theory offers a powerful conceptual tool for computational creativity systems—they mostly aim at producing effective and pleasurable output—even if it does not account for every pleasurable output[2].

In this paper we are focusing on creativity based on variation of known material. Of course this is only one of the aspects of creativity and many theories of human creativity have been proposed, mostly in abstract and general terms. Staying in the language domain, we would like to mention here only one of the most influential: Fauconnier and Turner's conceptual blending (Fauconnier & Turner, 2008), which aims at providing a description of one of the basic mechanisms of the creative process. According to these authors, novel creations can be obtained by merging elements and relations that normally belong to different "mental spaces". Conceptual blending works by projecting structures and qualities of these spaces into a new one, the blended mental space, which "develops structure not provided by the inputs" (Fauconnier & Turner, 1998).

Blending is a multi-step process, which involves, among other steps, (i) selection of (two or more) input spaces. We can take, for example, the mental space of "car" and the mental space of "yacht". The former includes roads, drivers, petrol, ...; the latter will include the water, sails, luxury, sailors, ...; (ii) selection of which characteristics of the input space that will be projected in the blended space (e.g., the road and the water); and (iii) the creation of a generic space where these characteristics are projected. A mapping between the elements of the two spaces is established, and the final blended space is thus obtained (continuing with the example, the outcome would be a "land yacht" metaphor, for describing a very

[2] Other theories also try to explain why this kind of modifications works. For example, a competing approach is that of Hanks' Theory of Norms and Exploitations (Hanks, 2013). According to Hanks, puns and creative expressions can be seen as a deliberate exploitation of a norm, that the reader can often easily notice. The OIH is more compelling: in the first place it involves familiarity, a broader concept and can work at the level of adaptation to the personal experience; furthermore, it is not limited to the linguistic domain.

expensive car; the yacht corresponds to the luxury car, the road for the car corresponds to the course for the boat, and the driver corresponds to the skipper).

Despite some issues that make it challenging to implement systems based on this theory without introducing additional constraints (Veale, 2019; Li et al., 2012), conceptual blending has been very influential in computational creativity research. Divago (Martins et al., 2019), for example, is a system for automatically blending two different domains among those that are present in its knowledge base. As an example, working on the concepts of "horse" and "bird", depending on which element of the two domains are mapped together, it can generate new animals such as a winged horse, a horse that flies by using its ears as wings, or a "transporter bird" that looks like a bird but can carry humans and be used for cargo and traction. Eppe et al. (2018) present another system based on conceptual blending in which answer set programming is used to find commonalities between the different concepts to merge; they show examples from multiple domains, generating novel metaphors, music chord progressions and even mathematical lemmas.

Part of the popularity of conceptual blending is certainly due to the promise of explaining the creative process. The OIH (Giora et al., 2004), instead, does not describe the process and steps that are necessary to come up with a creative product. The focus, in this case, is the explanation of *why* some creative items are so successful, and in particular the high-level characteristics they need to have to be effective.

3 Automating optimal innovation

As it is the case with many theories of creativity, the OIH is not "algorithmic" enough to allow for a straightforward translation into a computer program. While deciding what is "familiar" might be easy, especially if precise information about the user is obtainable (e.g., from the social media profile), producing a meaningful minimal modification of a linguistic expression that is perceived as novel, while keeping it reminiscent of the original text, is no small feat for a computer. In this section we briefly recall two systems that can produce optimally innovative output. The first one, HEADY-LINES, modifies famous expressions to repurpose them as headlines. In MOCKINGBIRD, song parodies are generated (and then sung by a synthesizer) by replacing words in the lyrics.

These systems share the initial part of the NLP pipeline, as shown in Fig. 1. They both start from the news of the day, downloading them from the RSS feed of the BBC and from the New York Times API. Each downloaded item consists of a headline and a news summary, a snippet of text that concisely describes the event the article is referring to. The headline is discarded[3], while the summary is lemmatized and part-of-speech tagged using Stanford CoreNLP (Manning et al., 2014). Then, the key concepts of the news are identified by considering the term

[3] This is done on the one hand because HEADY-LINES has to generate headlines, and it seemed unrealistic to start from a real headline, and on the other hand because they can sometimes have unusual grammatical structures that can be hard to parse with NLP tools.

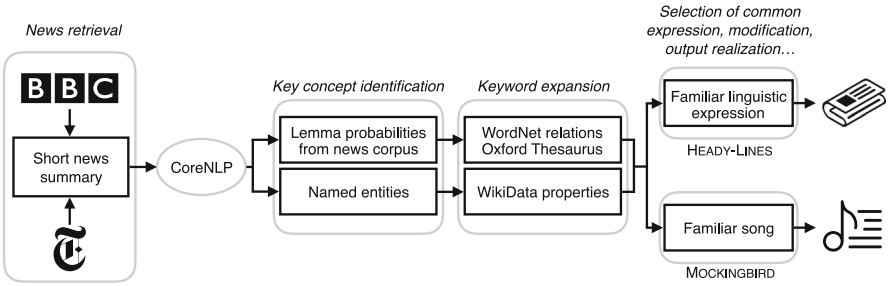

Fig. 1 Overview of the modification process of HEADY-LINES and MOCKINGBIRD

frequency, i.e. the probability of each word to appear in the LDC GigaWord corpus (Parker et al., 2011). Every lemma in the daily news that is above an empirically-determined threshold is discarded. The rationale is that, the more a term appears in the news, the least specific to a news article it is. This process for identifying the key concepts, albeit naïve, has two advantages. For one, probabilities were derived from a large news corpus, thus they are domain-specific. Secondarily, it gives us a "sensitivity" parameter (the probability threshold), which we can easily tune to include or remove concepts not specific enough for the single news event (e.g. "war"). Named entities (also detected with CoreNLP) are considered important and thus never discarded, as they normally indicate either the place where the news event took place, or the participants of the event. Since both Heady-Lines and Mockingbird will use these keywords to modify other expressions, there is a benefit from having more terms to describe a news event, so as to have more possibilities for modification. Synonyms, near-synonyms and derivationally related forms of these previous lemmas are obtained from the WordNet electronic dictionary (Fellbaum, 1998) and from the Oxford Thesaurus (Urdang, 1993), while WikiData is used to obtain properties for the named entities. This way we can incorporate world knowledge and information that is not directly stated in the summary, for example the fact that Trump is both a *tycoon* and *President of the United States*. In particular, we derive capital cities from the name of countries, countries from the name of cities or regions, and demonyms for all the geographical locations, while for people we extract names, surnames, occupations and fields of work.

Due to the different constraints given by the two applications and tasks, the modification step is specific for each system. For example, in HEADY-LINES we allow only one word to be replaced or inserted, as the risk of making the familiar expression unrecognizable increases with every word we change. In MOCKINGBIRD, the music provides such a strong connection with the original song that potentially every word could be changed, but we have additional constraints imposed by the musical properties. In the rest of this section we will describe the background and application scenarios of both systems, and the differences between the two.

3.1 Heady-Lines

Newspapers are often using catchy headlines based on familiar expressions, such as "The dark side of the sun" (for an article about the dangers of tanning booths), "This little LED of mine" (for an article describing LED light bulbs) or "Sick transit's glorious Monday" (referring to the day a plan to fix the ailing NYC transit system was announced), to attract readers. These headlines are clearly examples of optimal innovation, and are only effective for those readers that are familiar with the expression being modified (in the examples above: the album "the dark side of the moon", the gospel "this little light of mine", and the Latin phrase "sic transit gloria mundi"), as the clarity about the content is often traded for the captivating effect. Most of the headlines are either based on simple word substitution (e.g. the first and the second examples) or on puns (such as the last example).

Heady-Lines (Gatti et al., 2015) is a system for the automatic generation of creative headlines following optimal innovation principles. In particular, it combines a well-known expression with a concept coming from the news. The system is composed of five modules that deal with (i) retrieving the news of the day using the BBC and New York Times APIs, (ii) extracting keywords from the news, (iii) expanding them with relevant related concepts, (iv) pairing the news with well-known expressions using word2vec similarity, (v) generating a new headline by merging the well-known expression with a keyword coming from the news, satisfying the lexical and morpho-syntactic constraints enforced by the expression.

The first three modules are shared with Mockingbird, but after having obtained a list of extended concepts Heady-Lines averages their word2vec vectors (Mikolov et al., 2013) to obtain a semantic representation of the news element. This allows us to rank the well-known expressions in our database[4] from the most to the least related to the news. The expressions that do not reach a certain similarity threshold are discarded. This ensures at least a minimum degree of relatedness between the news and the well-known expression. For example, for the news "By any measure, it has been a year from hell for the European Union. And if Britons vote to leave the bloc, next year could be worse.", the most similar well-known expression is the royal anthem "God save the queen", followed by the song "Son of a preacher man". The similarity of the sentences is due to *queen* being related with *Britons*, while *God* and *preacher* are related to *hell*.

Well-known expressions are then modified by taking into account the lexical and syntactic constraints imposed by the original expression. This is accomplished by using a database that stores, for each relation in the dependency treebank of LDC GigaWord corpus, its occurrences with specific "governors" (heads) and "dependents" (modifiers), similarly to the approach of (Özbal et al., 2013). For each lemma in a well-known expression, we determine all the words that are connected to it by a dependency relation. Then, we calculate how likely each keyword coming from the news article can replace a lemma of the same part-of-speech. We can then

[4] Currently the database consists of manually-chosen titles of top-selling books, movies or songs, divided by decade. It also contains proverbs and idiomatic expressions, selected from online lists of "common English expressions".

Table 1 HEADY-LINES sample outputs

Summary	[...] Donald Trump's popularity with White America doesn't guarantee him the White House.
Expression	The unexpected virtue of ignorance
Headline	The unexpected popularity of ignorance
Summary	WTO finally reached deals [in a conference where] countries had been split over the path of trade reforms.
Expression	Bridge over troubled water
Headline	Bridge over troubled division
Summary	[Recently appointed] Australia captain Steve Smith [managed to successfully lead his team]
Expression	The empire strikes back
Headline	The captain strikes back
Summary	Shkreli resigns [after his arrest on] fraud charges
Expression	Crime and punishment
Headline	Fraud and punishment

select the slot with the lemma to be replaced, and the best keyword for each news article, by simply maximizing this dependency likelihood. Finally, the morphology of the replaced word is applied to the keyword using MorphoPro (Pianta et al., 2008) and the modified sentence is generated. In addition to these replacements, HEADY-LINES can also insert words, if they form a compound noun with a noun already present in the sentence. For example, for the the well-known expression "the empire strikes back", the system can generate "the British Empire strikes back", as "British Empire" is an existing compound noun in the dependency database. In case of both replacement and insertion, a threshold is enforced, so that sentences that do not reach a satisfactory level of grammaticality are discarded. To rank the final output, the system sorts each modified sentence according to its mean rank with respect to similarity and dependency scores, thus balancing the scores of grammaticality and relatedness to the news. The lower the mean, the better the system considers the headline.

The mechanism allows for the adoption of specific rhetorical strategies, which would intervene in the choice of candidate words. For example, higher precedence could be given to replacement words with a higher similarity level, or to antonyms, or to words that allude to certain features of the concept (e.g. privilege words related to sexual attitude when Berlusconi is the topic) to create explicit contrasts or connections so to produce an ironic, sarcastic or humorous headline. So far we did not want to introduce additional complexity, and in the following we only consider the basic system as described here.

On top of this algorithm, a web-interface (Gatti et al., 2016) has been developed that hides the technical details from the users (ideally "copy writers" , i.e. the creative professionals that, among other tasks, write the headlines for articles to

publish), and collaborates with them in the creative task of generating a good headline and allows them to focus on a subset of expressions that target the readers' specific age and interests. Some examples of the system output can be seen in Table 1.

3.2 Mockingbird

While Heady-Lines deals only with textual information, optimal innovation can also be found in multimodal scenarios. In these cases, multiple modalities can interact in different ways. Take for example Alanis Morissette' cover of "My Humps" (https://youtu.be/wTVbtwe5-NA), originally by The Black Eyed Peas. The lyrics are unchanged, the video is similar, however—due to the change in musical style—the song is perceived as a commentary on the ridiculousness of the original lyrics (SPIN staff, 2007).

Song parodies, i.e. the alteration of a familiar song so as to insert wording with a given communicative intention and character (e.g., evocative, ironical or derogatory), are also a powerful example of optimal innovation. They have been used in advertisement since the '20s and '30s, but they are common in many other forms of creative entertainment as well. An example is the commercial in which the words of Frank Sinatra's *My Way* are changed to *eBay* (https://goo.gl/V8J7ou).

Mockingbird (Gatti et al., 2017a) is a system that aims at automating the creative process underlying these song parodies. The system starts from a news event and tries to get the attention of a target public by singing, using the Vocaloid "song synthesizer" (Kenmochi & Ohshita, 2007), an appropriate well-known song with an alteration in the lyrics, so as to evoke the initial input. Also in this system, the reference to the OIH is clear.

The system uses the corpus developed by Strapparava and Mihalcea (Mihalcea & Strapparava, 2012), consisting of 100 popular songs, such as *Let It Be* by the Beatles, *Dancing Queen* by ABBA and *Alejandro* by Lady Gaga, where notes of the melody are strictly aligned with the corresponding syllables in the lyrics. Every song is annotated with its key (e.g., G major, C minor) and, for each note, its duration, the corresponding syllable in the lyrics, the time code with respect to the beginning of the song, the pitch and the distance of the note from the song key. This annotation was enriched with new tags, indicating the various parts of a song (e.g., chorus and verse), and an attribute that signals the "memorable" part of a song (i.e., the part that most people are supposed to quickly recognize).

As said, the goal of the system is to evoke the content of a particular news. The process for creating the modified song is divided into five main steps: (i) retrieving the daily news; (ii) identifying the most characterizing words of each news piece; (iii) finding new concepts and words evoking the initial text; (iv) altering the original lyrics by replacing words inside the chorus of a song with these concepts, according to musical and linguistic constraints; (v) producing a final output file, aligned with the background music. The files produced by the system are then played with a singing synthesizer, where a virtual voice will actually sing the new lyrics. As mentioned in Sect. 3, the pipeline for the first three steps is shared with Heady-Lines, while the rest of the modules are adapted to the lyrics domain.

In particular, the substitution step is implemented as a set of constraints that decides which word should be replaced in a song. In particular, we compare each content word in the chorus lyrics with each keyword and replace it if the following conditions apply:

– They have the same part of speech,
– They have the same number of syllables,
– The keyword rhymes with it, for song words at the end of song lines.

As with HEADY-LINES, the system applies the morphology of each replaced words to the keywords. Moreover, when multiple substitutions are possible, the Google Web n-grams (Brants & Franz, 2006) are used as a language model to decide which new word fits best with the context. For each word in a song, if the word is at the end of a song line, it will replace it with a related concept only if the concept (i) rhymes with the word; (ii) they both have the same number of syllables. If the word is in any other position, the rhyme constraint is not enforced. The rhyming information[5] is extracted from the CMU pronunciation dictionary (Rudnicky, 2014).

The rationale behind the part of speech, morphology and n-grams rules is to ensure grammaticality of the new song and maximize coherence, while the other constraints are enforced to avoid breaking the rhythmic properties of the lyrics. By keeping the count of syllables constant, we make sure that the synthesizer will sing the word at the same pace of the original. Rhymes at the end of song lines are maintained both to maximize the similarity with the original wording, and to avoid disrupting the rhyming with other line endings. N-grams are chosen instead of the dependency relations we use for HEADY-LINES due to the difficulty of reliably obtaining the dependency parse of song lyrics.

The song in Fig. 2 gives an idea of how the system works: MOCKINGBIRD swapped *day* with *ear*[6], since they have the same part of speech and the same number of syllables. The word *night* at the end of the first song line was replaced by *bite*, since in this position the rhyming constraint must also apply. In case multiple words satisfy the constraints, a language model is used to decide which word to use.

In the end the system produces a Vocaloid file, where each token is aligned with the musical features extracted from the corpus (e.g. pitch and duration). A MIDI track is also added to provide the background instruments. Once this file is opened in Vocaloid, the song created by the system can be sung directly or exported to a file. For a full description of the system, the reader can refer to (Gatti et al., 2017a), while some examples of the parodies produced by MOCKINGBIRD can be listened to at https://youtu.be/AS4w-ovhVJY.

[5] Stress is currently not taken into account, to avoid restricting too much the search space.

[6] The original lyrics are "*It's been a hard day's night, and I've been working like a dog/it's been a hard day's night, I should be sleeping like a log/but when I get home to you/I find the things that you do/will make me feel alright*"

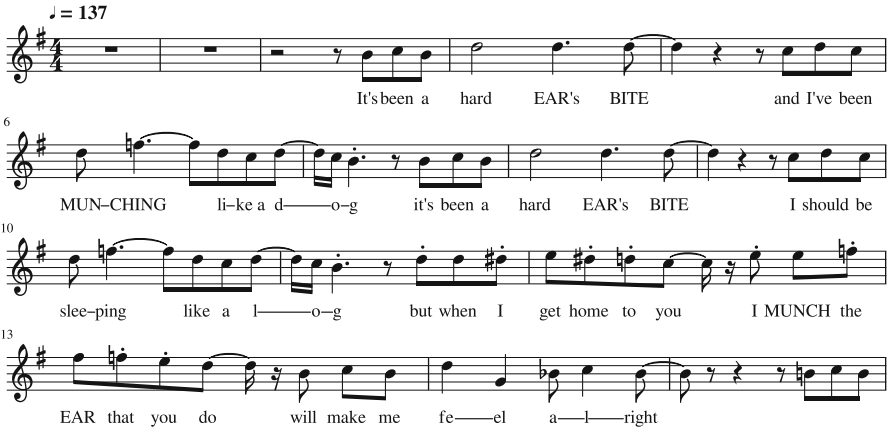

Fig. 2 The song "It's been a hard day's night" by the Beatles, after being altered by the system

4 Evaluation

When confronted with the challenge of evaluating the two systems, we did not have clear, well established references available. It was not a matter of seeing how the systems fare in relation to a given gold standard. Another theoretical option, comparing the system output with the production of a number of professional copy writers or advertising gurus was also not feasible: simply, they would not be available, at least in sufficient numbers. On the other hand, for the given task and level of the technology (for instance one could have noted that the systems do not rely on any serious reasoning on domain knowledge, just on language-based topic coherence), it was a well defined challenge to do equally well as non professional, but creative humans. In order to do that, we needed to design novel, complex evaluations, which included the selection of the best creative material produced by humans, to compare with. As for the multimodal, song-based setting, creative production is even more rare in nature and complex to require from human subjects. In this case we aimed to compare the system productions with other specific music-based options and see if the effectiveness of the former was better than the others. The practical goals of the two systems are different, one aims to attract the target attention to the contents, the other one to favor recall after time. The two evaluations needed to be specific but the underlying approach and methodology are the same: it is characterized by a comparative assessment of effectiveness, without neglecting relevance for the evoked topic and linguistic correctness. Crowdsourcing was the technical base for the experiments.

4.1 Heady-Lines

We evaluated HEADY-LINES with a series of experiments on the Figure-Eight crowdsourcing platform (formerly CrowdFlower). The goal was to determine whether the system can produce relevant, creative and effective headlines, with

minimal user intervention. To do so, we chose to compare the headlines produced by HEADY-LINES for 10 news articles with the best headlines written by a large number of non-professionals for the same articles. In addition to this, we also compared HEADY-LINES' output with the original headlines—written by a real copy writer—the ones that were originally associated with the 10 news.

4.1.1 Experimental setup

The outline of the evaluation procedure is the following: (a) we started by generating 10 headlines with HEADY-LINES. Since we wanted to compare these headlines with those of human authors, (b) we asked annotators to write creative headlines for the same 10 articles. (c) Then, different annotators rated all these headlines, indicating whether or not the headlines are creative and relevant to the original article. In this way, we could obtain an independent, fair selection of human-produced headlines that would include only those rated having good quality. With the same procedure, we also obtained a direct, explicit rating of the headlines generated by the system, as well as of the original headlines. (d) With these good-quality headlines, we then set up an all-play-all tournament where the best human-created headlines are directly compared to the ones the system created. The goal of this tournament is to determine a ranking among the different headlines, and see if HEADY-LINES can produce results that are comparable with human results. A detailed explanation of the experimental setup follows. The reader can also refer to Fig. 3 for an overview of the evaluation steps.

Why an evaluation of this sort, and not an ecological evaluation, with a comparison between headlines generated by the system and headlines written by a professional copy writer? First of all, there is the difficulty of finding multiple professional copy writers to evaluate the system against, on a number of preassigned news articles[7]. Secondarily, we found it unrealistic to only compare the quality of the system with that of the best human professionals (a bit like expecting state-of-the-art automatic poems to be on par with the works of Shakespeare or Milton). The quality of linguistic creativity systems is still limited by poor word-sense disambiguation and lack of world knowledge, among other things; on the other hand, differently from professional creatives, they can generate a great number of outputs in a short time. If the output quality, although not at the level of human excellence, is at least at the level of good nonprofessional creativity, these systems have the potential to create multiple *effective* personalized productions almost in real time, something that no human could ever achieve.

Let us follow the main steps of the evaluation:

(a) We started by choosing 10 news and had HEADY-LINES generate one headline for each of them. News were taken from multiple sources and different categories (such as sports, politics and weather), and the only human interventions were removing some key concepts incorrectly identified by the system, if any, or—in case some concepts were not automatically detected—

[7] Even an eminent literary critic, prof. Stanley Fish, claims he had difficulty in reaching out to copy writers for asking questions about their work (Fish, 2009)

adding concepts, and deciding which of the headlines proposed by the system is the best one for the given article. The "human curator" was not a trained copy writer, and had to choose between 3 and 7 headlines (this depends on how many modifications are "grammatical" and "related to the news", as determined by HEADY-LINES internal thresholds) and proposed by the system for each article. No manual edit was done to the final headline created by the system.

(b) Then, we showed to each Figure-Eight annotator one of the 10 news summaries, and asked him or her to write a creative headline for it. The task instructions contained examples of what we consider "creative headlines" taken from real tabloids and newspapers. The annotators had a limited time (3 min) to read the short summary and produce a headline. The allotted time was chosen to grant a fair comparison with HEADY-LINES in a real scenario, where a copy writer reads and validates the suggested headline in 1 or 2 min. We asked annotators to produce headlines for a single news article.
This was done for each of the 10 news, and for each news event we collected 10 headlines written by 10 different humans. Every headline proposed by annotators was manually checked, and those that were completely nonsensical (e.g. a headline completely off-topic) were discarded, so that the final dataset consisted of 100 news headlines with different degrees of creativity, but at least all marginally relevant.

(c) In the third step, we asked a different group of annotators to judge the headlines produced in (b). Annotators were presented each of the news summary with one of the human-created headlines; they then had to choose whether the headline was relevant or not, and whether it was creative or not. Each news summary was only presented once to each annotator (i.e., no annotator saw different headlines for the same news), to avoid direct comparisons and instead collect absolute "yes" or "no" ratings. The instructions clearly stated that the two dimensions, creativity and relevance, are potentially independent, and headline examples were there provided for the four cases. Each headline was viewed by 5 different annotators. Using the same process we collected also the relevance and creativity ratings of the original headline (the one written by a professional copy writer) and of the one created with HEADY-LINES. To exclude annotators that provide random answers, control questions (created by taking real news/headlines pairs and manipulating the 2 dimensions) were mixed with the real questions.

(d) Finally, we wanted to see how the automatically generated headlines rank, compared to good user-written ones. To do so, we first removed the user-written headlines that did not receive at least 2 positive (out of 5) votes for creativity and 2 (out of 5) positive votes for relevance in step (c). Then, we set up a "round-robin tournament", where each of the remaining 56 headlines, plus the headlines generated by the system and the original headlines (those written by the NYT and BBC copy writers for the 10 news we selected), are directly compared, one against the other. Actually, annotators were shown the news summary, the pair of "real" headlines and also a control headline not

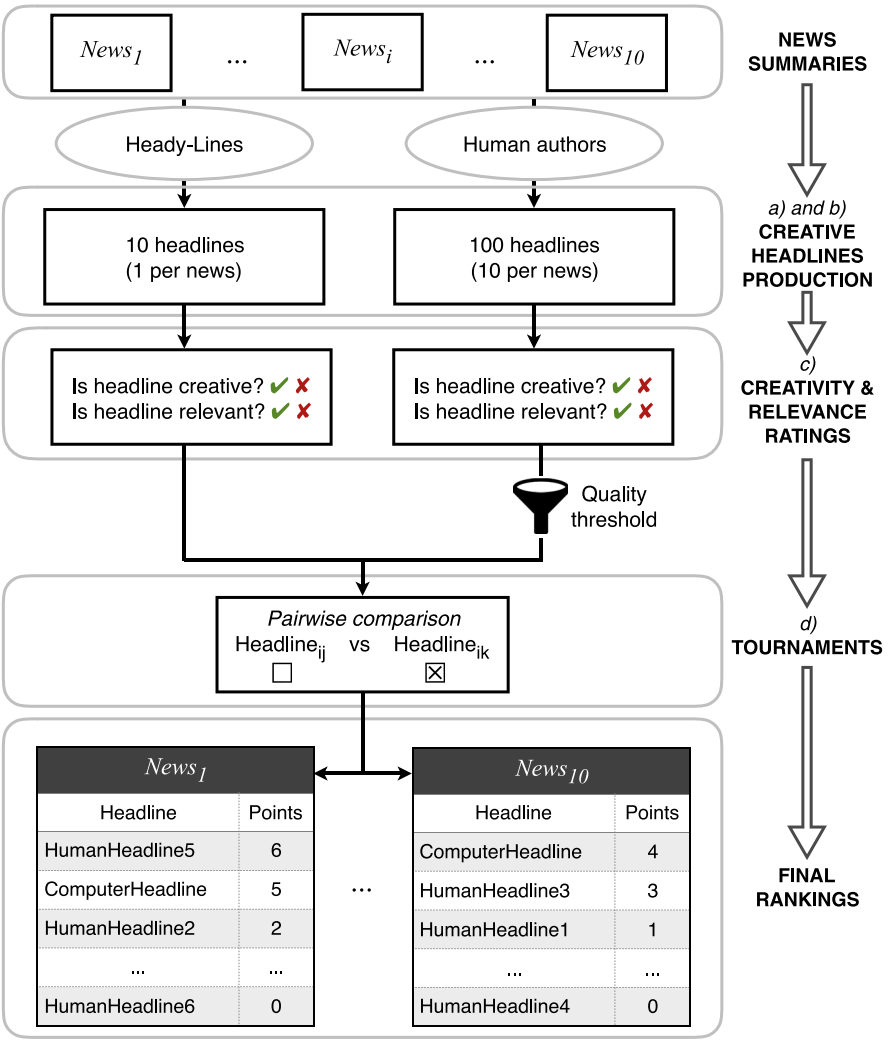

Fig. 3 Overview of the HEADY-LINES evaluation process

related to the article, all presented in a randomized order. Annotators were asked to choose which of the "three headlines of a tabloid" for the given article would attract them to read it. Annotators choosing even just one of the 10 controls were removed all-together from the final dataset. In this last step of the evaluation, the annotator - controls aside - is directly comparing two items, focusing on the effectiveness of the headline. Also in this case we only accepted annotators that did not participate in the previous tasks, and we never showed the same news event to a single annotator twice. For each pairwise comparison we collected 5 judgments, and used majority voting to determine the aggregate score.

4.1.2 Results and discussion

The results of the explicit rating step are shown in Table 2. Out of the 100 headlines written by users that we collected in phase (b), 47% were rated as being creative by at least 3 (out of 5) annotators (Fleiss' kappa = 0.453) in phase (c), while 78% were rated relevant by 3 (out of 5, Fleiss' kappa = 0.316). This gives us a clear indication that producing a creative headline, especially in a short time, is not an easy task for humans. Moreover, the correlation between the two dimensions is modest ($\rho = .22$), confirming that the relevance and creativity are not dependent one on another, but also that annotators understood the distinction between the two concepts. As for the 10 headlines created by HEADY-LINES, 80% were creative and 70% were relevant. Surprisingly, of the original headlines that appeared on the NYT and on the BBC, none was judged creative and only 70% were rated as being relevant to the article summary[8].

We cannot simply compare the creativity ratings of the system (80%) with that of 100 random annotators (47%). Since we did not pick 10 random productions of HEADY-LINES, it is only fair to compare them with the ones that were independently rated creative, as we will do in the next step. Regarding relevance, it is worth noting that both HEADY-LINES and the original headlines received a lower rating (both 70%) than the crowdsourced ones (78%). This might indicate that a clear explanation of the topic of the article is not an absolute requirement for a good headline, just one of the factors taken into account by the BBC and NYT copy writers.

4.1.2.1 Heady-Lines vs. creative humans The round-robin tournaments, one per each news article, were altogether "played" by 66 headlines, 56 written by creative humans (and rated as creative in phase (c) and 10 written by the system. We consider a "victory" each time a headline is preferred to another headline at least 3 times (out of 5 total ratings). Out of 10 tournaments, in 40% of cases HEADY-LINES produced the best headline, while in 60% of the tournaments the HEADY-LINES production is among the top-three headlines. On average, there are 6.6 headlines in each tournament, and the mean position of HEADY-LINES is 2.8; this means that the system is doing better than 58% of the "good" human-written news headlines. The automatic system is not only comparable to, but even substantially better than most of the creative human annotators.

It is worth highlighting that in this experiment we compared the best productions of 56 different human writers with those of a single system. The human-produced creative headlines were written by annotators who only had to focus on a single news element for a short time, hence they were a one-off effort. In fact, it would be

[8] One such headlines is "Gay Marriage Backers Celebrate in Germany: 'We Don't Need to Hide'", and the corresponding news summary is "Marchers filled the streets of Berlin days after a marriage equality law was passed. Supporters say the next step is to change the Constitution" (NYT, 22nd of July 2017). While we cannot know why 3 out of 5 annotators decided to rate the headline as non-relevant, we guess this is due to the fact that it does not mention the newly-approved law and the battle for changing the Constitution.

Table 2 Ratings collected in phase *b*)

Headlines creator	Creative headlines (%)	Relevant headlines (%)
HEADY-LINES	80	70
Professional copy writer	0	70
Human annotator	47	78

unlikely that a single human could produce in a short time 10 creative headlines, one for each of the 10 different articles, maintaining a consistent performance. HEADY-LINES, obviously, can do it. In the applied scenario, the human role would be limited to "choosing", less demanding than "inventing".

4.1.2.2 Heady-Lines vs. professional copy writers We also considered the results obtained when including the original headlines in the tournaments. In this case, the competing headlines are 76, and the mean number of headlines per tournament is 7.6. The average position of HEADY-LINES is 3.7, while the original headlines are, on average, in position 2.7; hence, they belong to the top-35%. The headlines written by professional copy writers are thus more broadly accepted; even if they are not creative (as seen by the ratings of phase (c)), they can be perceived as "the safer choice". It should also be noted that the original headlines were not written by a single person, as they come from different news sources, and we do not know how long their creation process took.

4.1.2.3 Familarity and optimal innovation A point worth considering is that one of the requirements for optimal innovation to happen is to modify something familiar to the reader. We, however, could not check whether the starting expressions were familiar to our judging annotators, meaning it is possible that some of the headlines were not "optimally innovative" for the readers. This would of course skew both the creativity rating of step (c) and the effectiveness of the headline during the tournaments of step (d). A post-hoc analysis with 4 English native speakers, two from the U.K. and two from the U.S., gave interesting insights. We gave instructions explaining the gist of the OIH, and provided examples of creative headlines based on it. We then showed them the 66 creative headlines that participated in the tournaments, and asked them to select the ones based on optimal innovation. For those, they had to write the original familiar expression they recognized. Of the 10 headlines produced by HEADY-LINES, 7 were recognized by all annotators. Each of the 3 remaining ones were still recognized by 3 annotators. This implicitly confirms the quality of the list of well-known expression; however, it also means that some headlines created by the system were not optimally innovative for some Figure-Eight annotators, and the results could be improved, for example by targeting the audience more accurately. One American annotator failed to recognize the British royal anthem "God save the Queen" behind "God save the bloc". While in this example it is hard to attribute the misjudgment to cultural differences across

the two sides of the Atlantic, this is probably the case for the crowdsourced headline "Let's make a WTO deal". Both our American annotators saw this as an optimal innovation on "Let's make a deal", a popular American TV game show, while none of the British annotators recognized anything familiar in it. Similarly, personal education and interests can influence the results: the only annotator without a computer science background did not recognize "Google Earth" in the human-written headline "Google Un-earthed"; for her, this was not a case of optimal innovation. We can also assume one of our annotators was not a fan of Bob Dylan, as he did not find anything familiar in "Blowing in the solar wind".

Out of 66 creative headlines (i.e., those produced by the system and the crowdsourced ones), 7 headlines (all created by HEADY-LINES) were unanimously declared examples of optimal innovation, 5 were selected by exactly 3 annotators, 7 were chosen by 2, and 11 were considered optimally innovative by exactly 1 annotator. Focusing on the 56 user-written headlines, 9 (16%) were selected at least 2 times, while 20 (30%) at least once, meaning that optimal innovation is indeed a strategy often used for producing creative expressions.

4.1.2.4 Optimally innovative headlines Let us now consider the performance of optimally innovative headlines (either created by a human or by the computer) against the other creative headlines. This will tell us whether optimal innovation is a good strategy for producing effective headlines. Considering the first group of tournaments we discussed in this section (66 creative headlines), the "optimal innovation team" consists of 19 headlines (29%, 9 human-written and 10 computer-written headlines). Despite being less than one half the size of the other team, optimally innovative headlines win 5 out of 10 tournaments, and at least an optimally innovative headline is in top-3 position in 9 out of 10 tournaments, clearly showing the effectiveness of optimal innovation headlines. It is also worth noting that, among this "optimal innovation team", 4 victories (out of 5) and 6 top-3 positions (out of 9) are of HEADY-LINES: the contribution of the computational system is substantial. Thus, not only optimal innovation is effective, but optimally innovative headlines of HEADY-LINES were qualitatively better than the ones produced by human annotators.

In sum, the evaluation confirmed the validity of the system:

- Headlines produced by HEADY-LINES have the similar relevance rating as headlines written by professional copy writers.
- In a direct comparison between the headlines automatically produced by HEADY-LINES and the best headlines written by humans, we find that the system is still better than the majority of them.

Not only this, but HEADY-LINES can produce these headlines at a steady, consistent rate; something that, we believe, is not achievable by an average human.

4.2 Mockingbird

A first, limited evaluation of the output of Mockingbird was run (Gatti et al., 2017b), to get an indication of its quality and (more implicitly) of the concept extraction and expansion. To do so, we asked 3 Figure-Eight annotators to read 10 news, then compare 10 altered songs with their unmodified version, both sang by Vocaloid. We asked the annotators to decide (i) which version is more grammatical (if any) and why; (ii) whether the modified song is more related to the headline, and why; (iii) whether the new song is fun.

The results were positive: 7/10 modified songs received a good "grammaticality" score, and 9/10 passed the "relatedness" test, also confirmed by a qualitative analysis of the answers. 6 out of 10 parodies were considered fun. This experiment, however, did not account for the main requirement of optimal innovation: familiarity with the song.

We present here a deeper evaluation (previously reported in (Gatti et al., 2017a)), also using a crowdsourced experiment, which addresses this issue, while focusing on the effectiveness of the system in helping recall (i.e., a delayed recall experiment).

4.2.1 Experimental setup

4.2.1.1 Stimulus presentation We first asked 653 subjects to decide which kind of music they know best, from the following categories: '90s rock and modern pop (e.g. Nirvana, Lady Gaga, Train), '60s rock (The Beatles, Bob Dylan), classical music, other ('70s dance, '80s pop, etc.). The task could be completed only by subjects that chose one of the first 2 categories (484 subjects in total), as those are best represented in the song corpus. This preference was saved, and every song during the experiment was then taken from the category of choice[9]. They then saw a distractor task, i.e. we presented a news headline with an altered song and told them that, at the end of the experiment, they would be asked a question about this song. The rest of the task proceeded as follows: participants were shown 5 news[10], one per page, each paired with a condition randomly chosen among:

- No song (i.e. just the headline and description);
- An unmodified unknown song;
- An unmodified known song;
- A randomly modified known song;
- A known song modified by MOCKINGBIRD (i.e. our experimental condition).

This randomization should mitigate the effect of the intrinsic memorability of single news, by spreading it across each condition. To avoid possible confounding effects

[9] This was done to increase the likelihood of songs being familiar to each subject.

[10] Each participant was shown the same 5 news. They were chosen to be at least 2 months old, so that they would not be reinforced outside the experiment, and not particularly memorable, to minimize the likelihood that subjects knew them already.

given by the quality of the songs themselves, each user was assigned one of two different "random" and "known" songs. Thus, our test dataset consisted of: 5 experimental songs, 2 known non-modified songs, 2 randomly modified songs for the "modern" category, other 9 songs for the "'60s rock" category, plus 1 unknown song[11]. All were manually selected to be particularly well-known and in line with the category description. After each song, subjects had to answer a question about its wording, to test whether they really listened to it or not. After 2 wrong answers the subjects were considered as unreliable and removed from the experiment. This was the case for 126 subjects. The news were presented in a randomized order too. After all the five news/conditions were seen by a subject, the second part of the distractor was shown, and we asked the subject to choose which words was in the altered song associated with the distractor news. We then asked the subjects to come back after 6 days for another experiment.

4.2.1.2 Recall test Six days later, we launched the second part of the experiment. We presented a list of 20 headlines: the 5 news presented in the first part, plus 15 distractors manually chosen to be semantically or lexically similar to the experimental news. The subjects were asked to choose from this list the 5 items they had seen in the first experiment. Our hypothesis was that the news which in the first phase were associated with the experimental condition would be remembered more often than the others. To be sure, and validate our "familiarity" assumption, we then presented again the songs (one per page) they heard during the first task, this time in the unmodified version sung by the original artist, and asked if they knew the song before the experiment. Finally, to confirm the reliability of the subjects, we concluded again with a control question where they had to match a headline to its short description, among 5 apparently similar options. In total we collected the answers of 198 subjects, but 33 of these failed the final test and their answers were discarded.

4.2.2 Results and discussion

The results for the 165 valid subjects are presented in Table 3. The "Unknown song" condition had to be excluded from the analysis, since the song was still rated as being familiar by 40% of participants, hence we deemed it not suitable for representing something completely novel. The other songs were familiar to at least 88% of the participants, as expected.

The news that was shown in association with the song altered by the system, i.e. the experimental condition, is the one that was by far remembered the most (64% of the participants). It appears that hearing unrelated songs is even detrimental to recall, since when an unmodified or randomly modified song is presented, the results are lower than when no song is present (47%, 52% and 53% respectively). We used

[11] "All my lovers", by the Australian singer Kylie Minogue, was used. Despite being a hit in 2010, it was deemed less memorable than most other songs in the corpus. Australian workers on Figure-Eight were excluded from the task, to further minimize the likelihood of the "unknown" song being familiar. Unfortunately, given that the song corpus we are using is small and composed of popular songs, it is not easy to choose a real "unknown" song in the format required by Mockingbird.

Cochran's Q test to determine if the differences across conditions are significant, and found that they are ($p < 0.05$). This suggests that a successful automatic modification helps recalling the news, just as similarly found in studies on advertising with songs altered by creative humans Allan (2006).

5 Related work

As previously mentioned, evaluations focused on pragmatics as those presented in Sect. 4 are rare in the field of computational creativity. In other fields, their usage is more common: these task-based evaluations are essential for those researching behavior change support systems (BCSS), but also works with a stronger focus on natural language generation (NLG) have looked at real-world effects of the generated language. For example, the STOP system (Reiter et al., 2003) generated tailored letters for persuading users to quit smoking, and its evaluation determined how many subjects quit smoking (or intended to quit) after 6 month of receiving the letter. More recently, Braun et al. (2015) describe an NLG system that reports on drivers' behavior; in their evaluation, the readers self-report whether the output is encouraging them to change their driving behaviour. While both these examples lie between the boundaries of NLG and behavior change support, SkillSum (Williams & Reiter, 2008) belongs to the former field. SkillSum is a system that produces personalized reports of basic literacy and numeracy skills for low-skilled readers. Despite not directly targeting at influencing behavior, the evaluation of the system also checked whether readers' self-assessment is more accurate after having read the generated report: a pragmatic effect of the feedback.

In this work we argue for evaluating the pragmatic effects of computational creativity systems. It is worth mentioning that a more traditional evaluation, focusing on the other aspects and qualities of these systems, should be considered as well. Jordanous (2019) highlights some of the challenges of these evaluations, by surveying a large number of works describing creative systems. In her work she analyzes some of the evaluation methodologies proposed in the past, and introduces a *Standardised Procedure for Evaluating Creative Systems* (SPECS), a method to specifically evaluate *creativity*. SPECS is a three-step process: developers of creative systems should first choose a definition of creativity that is suitable for the system to be tested; then, the sub-components of this definition should be identified as dimensions to be evaluated; finally, the actual evaluation should be executed on each of these dimensions, using appropriate methods, and the results reported. In addition to determining whether a system is creative or not, it is often useful to be able to compare different versions of the same system, to determine whether the new developments corresponds to actual improvements. To this end, Colton et al. (2014) propose comparing both diagrams and actual output of two versions of a system. The diagrams represent the interplay of programmer and program behaviours, covering both development and execution of the creative system. By comparing the diagrams, it is possible to notice, for example, when new complex behaviors are introduced, or when a task that initially required a substantial human intervention is now entirely done by the system itself. Colton and colleagues also

Table 3 Results of the news recall experiment

Condition	Remembered news	Percentage
Experimental	105	64
Unknown	88	53
No song	87	53
Random	86	52
Known	77	47

advocate for comparing the output to collect *uncreativity* judgements. The rationale behind this uncommon choice is that creativity is often contested, and it is impossible to achieve a consensus on whether a system is creative; *uncreativity*, however, is more easily recognized.

Let us now discuss some recent work that deal with issues related to headline generation. Xu et al. (2010) extract keywords from news articles and recombine them into a new headline. They start by downloading Wikipedia pages related to an article, and deriving word features from their inlink, outlink, category and infobox information. These features (plus others derived from the article itself) are then fed into a classifier that decides which of the words are article keywords. The keywords are then recombined using the process described in (Zhou & Hovy, 2003). More recently, given the great contribution of deep learning models to NLP, also headline generation has shifted to neural networks. A survey by Ayana et al. (2017) presents the standard architecture for neural headline generation. This is usually composed by an encoder, which computes a representation of the article as a single vector or as a sequence of vectors, and a decoder which actually generates the headline, emitting one word at a time. While there are many different ways to encode the input data to feed into the encoder (e.g., using standard pre-trained word embeddings, or including information about PoS tags, or TF/IDF statistics), and different encoder (e.g., CNNs or RNNs) and decoder models (from simple neural language models to recurrent models using attention), these are considered encoder/decoder models. In all cases, the difference with Heady-Lines is not simply the architecture, but more importantly the goal and the evaluation. The task for the work reported in (Xu et al., 2010; Zhou & Hovy, 2003; Ayana et al., 2017) is to produce a headline describing the news. In their case, the difference with text summarization is mainly the grammatical structure of headlines. The goal of Heady-Lines is producing a creative headline instead, and for this purpose creativity trumps clarity (i.e., creative headlines are allowed to be more ambiguous or less descriptive of the event). As far as the evaluation is concerned, most works collect grammaticality or readability judgements, and in some cases the similarity with real headlines is considered (e.g. comparing automated scores such as ROUGE or BLEU with a reference corpus of human-written headlines). This is in stark contrast with Heady-Lines, where we investigated whether people think that the computer-generated headlines are fit for being published.

Two recent works by Alnajjar and colleagues follow in the tracks of our Heady-Lines experimentation. In Alnajjar et al. (2019), the authors describe a systems for generating "colorful headlines", starting from non-creative automatically-generated

headlines prepending them with a coherent well-known phrase (a proverb or a movie title), or by adding figurative language. While the former strategy may look reminiscent of optimal innovation, they do not explicitly use it as a framework of reference and indeed the well-known phrase is not modified in any way. The system presented in Alnajjar and Toivonen (2020), instead, generates a slogan using properties of the target concept to remember, and placing them in the "skeleton" (the PoS-tags sequence) of an existing slogan. Also in this case, the goal is not to produce "optimal innovations", hence the "donor" slogan need not be recognizable in the output. In the evaluation of both systems, the authors considered pragmatic aspects by asking users to rate the "catchiness, attractiveness, memorability" (in addition to other dimensions) of the generated headlines (compared to automatically-generated non-creative headlines) and slogans (compared to randomly-selected real slogans). As mentioned, creative NLG systems are rarely evaluated in these terms.

As for MOCKINGBIRD, its task is related to lyrics generation, which in turn shares some connections with poetry generation. In both these tasks the systems need to produce text considering not only grammars and semantics, but also metrical properties and phonetics. Among the works dealing with these tasks, the one of Barbieri et al. (2012) is worth mentioning due to the similarities with MOCKINGBIRD. The goal of their system is to generate lyrics for a song, imitating the style of a particular songwriter. To do so, they use a Constrained Markov Process, with unary constraints on meter (using "rhythmic templates" based on the stresses of words), rhyme (forcing rhyme patterns at the end of verses), syntax (using PoS templates automatically derived from a corpus of lyrics) and semantics (calculating the Wikipedia Link-Based similarity between a chosen "theme word" and the words to be inserted into the templates). MOCKINGBIRD goes through a similar process, although with slightly different constraints (e.g., meter and rhyme constraints do not consider the stress on words), a different output (i.e. lyrics in a phonetic form that can be sung by Vocaloid) and with a set of keywords that comes from the news. As it was the case for most headline generation systems, their evaluation is only concerned with syntactic correctness and semantic relatedness of title and lyrics. More importantly, MOCKINGBIRD does not generate the whole text from scratch, but only replaces content word (and only if the constraints can be satisfied), to keep an even stronger association between the original song and the new lyrics. The delayed-recall test used for the evaluation indicates that this method can lead to strong pragmatics effects, allowing easier retrieval of the topic (the news) presented in the song.

6 Conclusions

The aesthetic dimension is of particular importance when producing creative content (Colton et al., 2012). At the same time, behind many human creative acts there is not only the intent to create something aesthetically pleasing, but also a desired pragmatic effect. Also creative machines should support the achievement of these pragmatic effects, in addition to presenting something aesthetically pleasing.

Evaluating the output of creative systems is often a challenge, however, and attempts to define a general approach (such as (Ritchie, 2007)) have always been complex. In the literature we can find evaluations that mostly focus on two dimensions: on the one hand the linguistic and semantic properties of the output, and on the other hand its creative qualities. The former are the most well-studied, and they can benefit from decades of research in natural language generation. The latter are more challenging and require a careful design, hence the different methodologies mentioned in Sect. 5. In this paper, we presented as a case study two system for applied creativity, based on the OIH and minimal variations of familiar expressions. We emphasized that their evaluations add a third dimension, which is based on the effects of the productions on the reader. From the evaluations, we can draw some general conclusions for other computational creativity systems, especially those that aim at producing output with pragmatic potential.

The first one is that, when comparing the quality of automatically generated output and human output, it is worth considering what classifies as reasonable human output. In the case of HEADY-LINES, we compared the system headlines with those written by professionals, and with creative headlines produced by "the average human". Not only is the comparison more realistic given the current capabilities of computers, but we also believe that, in many real-world scenarios where creative systems could have an impact, there is no need for "outstanding" creativity: the sheer rate of creative productions paired to an above-average quality level would already make creative computers powerful tools to assist human writers.

Furthermore, we want to stress the importance of designing evaluations that try to measure whether the pragmatic effect of the output is achieved. These evaluations are currently extremely rare in computational creativity research; in most cases, only the intrinsic properties of the output are studied. We think, however, that computational linguistic creativity has a strong potential for many applied settings. In our case study we have focused on the news, but advertising and education easily come to mind as well. To ensure that autonomous and semi-autonomous creative systems can be useful in these fields, we need to be able to creatively measure their effectiveness in terms of pragmatics.

References

Allan, D. (2006). Effects of popular music in advertising on attention and memory. *Journal of Advertising Research, 46*(4), 434–444.

Alnajjar, K., & Toivonen, H. (2020). Computational generation of slogans. *Natural Language Engineering*. https://doi.org/10.1017/S1351324920000236

Alnajjar, K., Leppänen, L. & Toivonen, H. (2019). No time like the present: Methods for generating colourful and factual multilingual news headlines. In *Proceedings of the 10th International Conference on Computational Creativity (ICCC 2019)*.

Ayana, Shen, S. Q., Lin, Y., Tu, C. C., Zhao, Y., Liu, Z., & Sun, M. (2017). Recent advances on neural headline generation. *Journal of Computer Science and Technology, 32,* 768–784.

Barbieri, G. , Pachet, F. , Roy, P. & Degli Esposti, M. (2012). Markov constraints for generating lyrics with style. In *Proceedings of the 20th European Conference on Artificial Intelligence (ECAI 2012)* (Vol. 242, pp. 115–120).

Brants, T., & Franz, A. (2006). *Web 1T 5-gram version 1 LDC2006T13. DVD*. Linguistic Data Consortium.

Braun, D., Reiter, E. & Siddharthan, A. (2015). Creating textual driver feedback from telemetric data. In *Proceedings of the 15th European workshop on Natural Language Generation (ENLG 2015)* (pp. 156–165).

Clark, E., Ross, A. S., Tan, C., Ji, Y. & Smith, N. A. (2018). Creative writing with a machine in the loop: Case studies on slogans and stories. In *Proceedings of the 23rd international conference on Intelligent User Interfaces (IUI '18)* (pp. 329–340).

Colton, S., Charnley, J. W. & Pease, A.(2012). Computational creativity theory: The FACE and IDEA descriptive models. In *Proceedings of the 2nd International Conference on Computational Creativity (ICCC 2011)* (pp. 90–95).

Colton, S., de Mántaras, R. L., & Stock, O. (2009). Computational creativity: Coming of age. *AI Magazine, 30*(3), 11–14.

Colton, S., Pease, A., Corneli, J., Cook, M. & Llano, T. (2014). Assessing progress in building autonomously creative systems. In *Proceedings of the 5th International Conference on Computational Creativity (ICCC 2014)* (pp. 137–145).

Eppe, M., Maclean, E., Confalonieri, R., Kutz, O., Schorlemmer, M., Plaza, E., & Kühnberger, K. U. (2018). A computational framework for conceptual blending. *Artificial Intelligence, 256*, 105–129. https://doi.org/10.1016/j.artint.2017.11.005

Fauconnier, G., & Turner, M. (1998). Conceptual integration networks. *Cognitive Science, 22*(2), 133–187.

Fauconnier, G., & Turner, M. (2008). *The way we think: Conceptual blending and the mind's hidden complexities*. Basic Books.

Fellbaum, C. (1998). *WordNet*. Wiley.

Fish, S. (2009). Headline art. *The New York Times*. Retrieved March, 10, 2021, from http://web.archive.org/web/20210310114310/

Gatti, L., Özbal, G., Guerini, M., Stock, O. & Strapparava, C. (2015). Slogans are not forever: Adapting linguistic expressions to the news. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI 2015)* (pp. 2452–2458).

Gatti, L., Özbal, G., Guerini, M., Stock, O. & Strapparava, C. (2016). Heady-Lines: A creative generator of newspaper headlines. In *Companion publication of the 2016 International Conference on Intelligent User Interfaces (IUI '16)* (pp. 79–83).

Gatti, L., Özbal, G., Stock, O. & Strapparava, C. (2017a). Automatic generation of lyrics parodies. In *Proceedings of the 25th ACM Multimedia conference (ACMMM-2017)* (pp. 485–491).

Gatti, L., Özbal, G., Stock, O. & Strapparava, C. (2017b). To sing like a mockingbird. In *Proceedings of the 16th conference of the European chapter of the Association for Computational Linguistics (EACL-2017)*.

Giora, R., Fein, O., Kronrod, A., Elnatan, I., Shuval, N., & Zur, A. (2004). Weapons of mass distraction: Optimal innovation and pleasure ratings. *Metaphor and Symbol, 19*(2), 115–141.

Giora, R., Givoni, S., Heruti, V., & Fein, O. (2017). The role of defaultness in affecting pleasure: The optimal innovation hypothesis revisited. *Metaphor and Symbol, 32*(1), 1–18.

Hanks, P. (2013). *Lexical analysis: Norms and exploitations*. Mit Press.

Jordanous, A. (2019). Evaluating evaluation: Assessing progress and practices in computational creativity research. In T. Veale & F. A. Cardoso (Eds.), *Computational creativity: The philosophy and engineering of autonomously creative systems* (pp. 211–236). Springer. https://doi.org/10.1007/978-3-319-43610-4_10

Kenmochi, H. Ohshita, H. (2007). VOCALOID—Commercial singing synthesizer based on sample concatenation. In *Proceedings of the 8th annual conference of the International Speech Communication Association (INTERSPEECH'07)* (pp. 4009–4010).

Lamb, C., Brown, D. G., & Clarke, C. L. (2017). A taxonomy of generative poetry techniques. *Journal of Mathematics and the Arts, 11*(3), 159–179.

Li, B., Zook, A., Davis, N. & Riedl, M. O. (2012). Goal-driven conceptual blending: A computational approach for creativity. In *Proceedings of the 2012 International Conference on Computational Creativity (ICCC 2012)* (pp. 3–16).

Machado, P. Cardoso, A. (1998). Computing aesthetics. In *Proceedings of the 12th Brazilian Symposium on Artificial Intelligence (SBIA '98)* (pp. 219–228).

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S. & McClosky, D. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of the demo track of the 52nd annual meeting of the Association for Computational Linguistics (ACL 2014)* (pp. 55–60).

Martins, P., Pereira, F. C., & Cardoso, A. F. (2019). The nuts and bolts of conceptual blending: Multidomain concept creation with Divago. In T. Veale & F. A. Cardoso (Eds.), *Computational creativity: The philosophy and engineering of autonomously creative systems* (pp. 91–119). Cham: Springer. https://doi.org/10.1007/978-3-319-43610-4_5

Mihalcea, R. Strapparava, C. (2012). Lyrics, music, and emotions. In *Proceedings of the 2012 joint conference on Empirical Methods on Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL '12)* (pp. 590–599).

Mikolov, T , Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th conference on advances in Neural Information Processing Systems (NIPS 2013)* (pp. 3111–3119).

Munigala, V., Mishra, A., Tamilselvam, S. G., Khare, S., Dasgupta, R. & Sankaran, A. (2018). PersuAIDE an adaptive persuasive text generation system for fashion domain. In *Proceedings of ACM the Web conference (WWW '18)* (pp. 335–342).

Özbal, G. & Strapparava, C. (2013). Namelette: A tasteful supporter for creative naming. In *Companion publication of the 2013 international conference on Intelligent User Interfaces (IUI '13)* (pp. 55–56).

Özbal, G., Pighin, D. & Strapparava, C. (2013). BrainSup: Brainstorming support for creative sentence generation. In *Proceedings of the 51st annual meeting of the Association for Computational Linguistics (ACL 2013)* (pp. 1446–1455).

Parker, R., Graff, D., Kong, J., Chen, K., & Maeda, K. (2011). *English GigaWord fifth edition, LDC2011T07 DVD*. Linguistic Data Consortium.

Pérez y Pérez, R. & Sharples, M. (2004). Three computer-based models of storytelling: BRUTUS, MINSTREL and MEXICA. *Knowledge-based systems, 17*(1), 15–29.

Pianta, E., Girardi, C. & Zanoli, R. (2008). The TextPro tool suite. In *Proceedings of the 6th international conference Language Resources and Evaluation– (LREC 08)* (pp. 2603–2607).

Reiter, E., Robertson, R., & Osman, L. M. (2003). Lessons from a failure: Generating tailored smoking cessation letters. *Artificial Intelligence, 144*(1), 41–58.

Ritchie, G. D. (2007). Some empirical criteria for attributing creativity to a computer program. *Minds and Machines, 17*(1), 67–99.

Ritchie, G. D., Manurung, R., Pain, H., Waller, A., & O'Mara, D. (2006). The STANDUP interactive riddle builder. *IEEE Intelligent Systems, 21*(2), 67–69.

Rudnicky, A. (2014). *The CMU pronouncing dictionary release 07b*. Carnegie Mellon University.

SPIN staff. (2007). Alanis Morissette 'My Humps' Video. SPIN. Retrieved March, 10, 2021, from https://www.spin.com/2007/04/alanis-morissette-my-humps-video/

Stock, O. Strapparava, C. (2003). Getting serious about the development of computational humor. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI-03)* (Vol. 3, pp. 59–64).

Tomašic, P., Znidaršic, M. & Papa, G. (2014). Implementation of a slogan generator. In *Proceedings of 5th International Conference on Computational Creativity (ICCC 2014)* (pp. 340–343).

Urdang, L. (1993). *The Oxford thesaurus: An AZ dictionary of synonyms*. Clarendon Press.

Veale, T. (2014). Coming good and breaking bad: Generating transformative character arcs for use in compelling stories. In *Proceedings of the 5th International Conference on Computational Creativity (ICCC 2014)*.

Veale, T. (2016) Round up the usual suspects: Knowledge-based metaphor generation. In *Proceedings of the 4th workshop on Metaphor in NLP (Metaphor 2016)* (pp. 34–41).

Veale, T. (2019). From conceptual mash-ups to badass blends: A robust computational model of conceptual blending. In T. Veale & F. A. Cardoso (Eds.), Computational creativity: The philosophy

and engineering of autonomously creative systems (pp. 71–89). Springer. https://doi.org/10.1007/978-3-319-43610-4_4

Williams, S. & Reiter, E. (2008). Generating basic skills reports for low-skilled readers. *Natural Language Engineering, 14*(4), 495–525. https://doi.org/10.1017/S1351324908004725

Xu, S., Yang, S. & Lau, F. C. M. (2010). Keyword extraction and headline generation using novel word features. In *Proceedings of the 24th conference on Artificial Intelligence (AAAI-10)* (pp. 1461–1466).

Zhou, L. & Hovy, E. (2003). Headline summarization at ISI. In *Proceedings of the Document Understanding Conference (DUC-2003).*