EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

Brussels, 4 June 2019

COST 048/19

DECISION

Subject: **Memorandum of Understanding for the implementation of the COST Action "Multi3Generation: Multi-task, Multilingual, Multi-modal Language Generation" (Multi3Generation) CA18231**

The COST Member Countries and/or the COST Cooperating State will find attached the Memorandum of Understanding for the COST Action Multi3Generation: Multi-task, Multilingual, Multi-modal Language Generation approved by the Committee of Senior Officials through written procedure on 4 June 2019.

MEMORANDUM OF UNDERSTANDING

For the implementation of a COST Action designated as

COST Action CA18231
MULTI3GENERATION: MULTI-TASK, MULTILINGUAL, MULTI-MODAL LANGUAGE GENERATION (Multi3Generation)

The COST Member Countries and/or the COST Cooperating State, accepting the present Memorandum of Understanding (MoU) wish to undertake joint activities of mutual interest and declare their common intention to participate in the COST Action (the Action), referred to above and described in the Technical Annex of this MoU.

The Action will be carried out in accordance with the set of COST Implementation Rules approved by the Committee of Senior Officials (CSO), or any new document amending or replacing them:

a. "Rules for Participation in and Implementation of COST Activities" (COST 132/14 REV2);
b. "COST Action Proposal Submission, Evaluation, Selection and Approval" (COST 133/14 REV);
c. "COST Action Management, Monitoring and Final Assessment" (COST 134/14 REV2);
d. "COST International Cooperation and Specific Organisations Participation" (COST 135/14 REV).

The main aim and objective of the Action is to foster an interdisciplinary network of research groups working on different aspects of language generation (LG), focussing on 4 themes: grounded multi-modal reasoning and generation; efficient machine learning algorithms, methods, and applications to LG; dialogue, interaction and conversational LG applications; and exploiting large knowledge bases and graphs. This will be achieved through the specific objectives detailed in the Technical Annex.

The economic dimension of the activities carried out under the Action has been estimated, on the basis of information available during the planning of the Action, at EUR 68 million in 2018.

The MoU will enter into force once at least seven (7) COST Member Countries and/or COST Cooperating State have accepted it, and the corresponding Management Committee Members have been appointed, as described in the CSO Decision COST 134/14 REV2.

The COST Action will start from the date of the first Management Committee meeting and shall be implemented for a period of four (4) years, unless an extension is approved by the CSO following the procedure described in the CSO Decision COST 134/14 REV2.

TECHNICAL ANNEX

OVERVIEW

Summary
Language generation (LG) is a crucial technology if machines are to communicate with humans seamlessly using human natural language. A great number of different tasks within Natural Language Processing (NLP) are language generation tasks, and being able to effectively perform these tasks implies (1) that machines are equipped with world knowledge that can require multi-modal processing and reasoning (e.g. textual, visual and auditory inputs, or sensory data streams), and (2) the study of strong, novel Machine Learning (ML) methods (e.g. structured prediction, generative models), since virtually all state-of-the-art NLP models are learned from data. Moreover, human languages can differ wildly in their surface realisation (i.e. scripts) as well as their internal structure (i.e. grammar), which suggests that multilinguality is a central goal if machines are to perform seamless language generation. Language generation technologies would greatly benefit both public and private services offered to EU citizens in a multilingual Europe, and have strong economic and societal impacts.

Areas of Expertise Relevant for the Action	**Keywords**
● Computer and Information Sciences: Machine learning algorithms ● Languages and literature: Linguistics: formal, cognitive, functional and computational linguistics ● Computer and Information Sciences: Artificial intelligence, intelligent systems, multi agent systems	● multi-task ● multilingual ● multi-modal ● natural language processing ● machine learning

Specific Objectives
To achieve the main objective described in this MoU, the following specific objectives shall be accomplished:

Research Coordination
● Foster knowledge exchange by sharing of resources including semantic annotation guidelines, benchmarking corpora, machine learning and alignment tools.
● Create multimodal and multilingual benchmarks for NLG involves experimenting with automatic mapping between existing resources, crawling of Web data, definition of annotation guidelines and launching of crowdsourcing campaigns for bigger datasets (also as games-with-a-purpose).
● Facilitate interactions, collaborations, knowledge building and dissemination between Action participants via online tools, as website, blogs, downloadable publications.
● Promote the generation of novel ideas and introduce the new joint Multi3Generation discipline to other researchers.
● Provide opportunities for joint research projects by Action members on multi-task, multilingual and multi-modal processing during exchange visits of Early Career Investigator, and other activities that encourage young researchers to establish links with industry and more senior academics.
● Disseminate the results of the Action through conferences, scientific and industrial gatherings, which will have substantial impact in the participating countries and beyond.
● Create synergies between participants via joint publications in books, journals and conferences; reports from working group meetings and training materials from training schools.

Capacity Building
● Strengthen the European research on theory, methodology and real-world technology in language generation, particularly in the four Multi3Generation focus research themes.
● Facilitate international collaboration, networking and interdisciplinary community fostering joint activities.
● Drive scientific progress by liaising extensively with industry and end-users, and by increasing joint collaboration and knowledge transfer by the end of the Action.

TECHNICAL ANNEX

1 S&T EXCELLENCE

1.1 SOUNDNESS OF THE CHALLENGE

1.1.1 DESCRIPTION OF THE STATE-OF-THE-ART

Natural Language Generation (NLG) encompasses all Natural Language Processing (NLP) tasks that deal with automatically (or semi-automatically) generating readable text. Texts generated by NLG models usually have a large audience and can be quite application-specific. Those applications can be categorised in two broad areas: *text-to-text* and *data-to-text* generation. The former takes existing texts as input and produces a coherent text as output. Example applications include Machine Translation (MT), Question Answering (QA), text summarisation, and spelling correction. The latter converts non-linguistic data to text, for instance, football commentary, weather and financial report generation, patient information summarisation, and image/video caption generation.

T1 Grounded multi-modal reasoning and generation
Current state-of-the-art multi-modal reasoning and generation systems are generally based on Neural Network (NN) architectures and are commonly trained end-to-end and for a specific task. The simplest neural image captioning model (Vinyals et al., 2015) pretrains image representations with a neural image classification model (Russakovsky et al., 2015) and then initialises a text generator with the image representation. Similar approaches have been also used for multi-modal reasoning and generation for visual question answering (Mostafazadeh et al., 2017) and video description (Donahue et al., 2015). The adoption of attention mechanisms in neural MT (NMT) systems have proved to be greatly useful (Bahdanau et al., 2014; Luong et al., 2015). In an attentive NMT model, the target sentence is generated word by word while at every step the attention mechanism dynamically directs its attention to the relevant parts in the source sentence. More recent developments include multilingual MT where especially low resource languages benefit from the multilingual setting (Firat et al., 2017), and image caption translation, where the decoder can attend both to the source caption as well as the image itself (Huang et al., 2016; Calixto et al., 2017). Similarly, visual question answering systems, where the questions and answers are grounded in an image can benefit from attending to particular parts of the picture depending on the question (Ilievski and Feng, 2017).

Word embeddings pre-trained on large text corpora (Mikolov et al., 2013; Pennington et al., 2014) provide elementary building blocks for various NLP tasks and can be adopted in low-resource settings as pre-training for initialising model parameters. Recently, several approaches for learning contextual word embeddings (Devlin et al., 2018; Peters et al., 2018) and multi-modal embeddings have been proposed, e.g. grounded in images (Kottur et al., 2016) or sounds (Vijayakumar et al., 2017).

T2 Efficient Machine Learning algorithms, methods, and applications to language generation
Neural Networks (NNs) are designed to learn representations at increasing levels of abstraction, and such representations are dense, low-dimensional, and distributed, making them well-suited to capturing grammatical and semantic generalisations (Mikolov et al., 2013; Pennington et al., 2014). They have achieved success in sequential modelling using feedforward networks (Bengio et al., 2003) and Recurrent Neural Networks (RNNs, Mikolov et al., 2010), as well as the more recent Transformer models with positional embeddings (Vaswani et al., 2017). The main advantage of using NNs over feature-based Language Models is that they model a possibly infinite history, while avoiding both data sparseness and an explosion in the number of parameters through the projection of histories into a low-

dimensional space, so that similar histories share representations. An influential architecture is the encoder-decoder framework (Sutskever et al., 2014), where an RNN is used to encode the input into a vector representation, which serves as the auxiliary input to a decoder RNN. This decoupling of encoder and decoder makes it possible in principle to share the encoding vector across multiple NLP tasks in a multi-task learning setting (Dong et al., 2015; Luong et al., 2015).

In multi-modal data-to-text generation it is common to use computer vision methods and Convolutional Neural Networks (CNN) for detecting and labelling objects, attributes, spatial relations, and possibly also action and pose information (Yatskar et al., 2014; Kuznetsova et al., 2014). This is usually followed by mapping outputs to linguistic structures. Holistic scene analysis methods either use a unimodal space to compare a query image to training images before caption retrieval (e.g. Ordonez et al., 2011; Gupta et al., 2012), or exploit a multi-modal space representing proximity between images and captions (e.g. Hodosh et al., 2013; Socher et al., 2014). Moreover, interesting work on vision-language integration is being carried out with deep learning models, e.g. Kiros et al. (2014) predict the next word in a sequence based on both the linguistic context and CNN image features. Finally, recent advances in ML algorithms and techniques include neural models that induce sparsity to better model linguistic bias (Martins and Astudillo, 2016; Niculae et al., 2018), using graph encoding models to incorporate knowledge bases into NLP (Schlichtkrull et al., 2017), and generative models of language to better capture linguistic priors and variation in translation data (Schulz et al., 2018).

T3 Dialogue, interaction and conversational language generation applications
The global success stories, such as *IBM Watson (Ferruci et al. 2010), Apple Siri, Microsoft Cortana, Amazon Alexa* and *IPsoft Amelia*, have resulted in a new wave of research and development in the field of Human-Computer Interaction (HCI). Virtual assistants helping with simple tasks is a reality for mobile users, and are becoming more and more widespread in business applications. Even though there are virtual agent applications available in the market, these technologies are still being actively researched and are considered to be innovative near future technologies. According to Gartner (Gartner, 2014; Gartner 2016) this smart and intelligent machine era will blossom, and it will be “the most disruptive in the history of IT”.

The dominant technique in recently created virtual assistants is the application of deep learning models that learns from texts, examples and other relevant data (e.g. Vinyls & Le, 2015; Serban et al., 2015, Li et al., 2016). However, most current solutions are offered in English, since data in other languages for training conversational agents is sparse. Moreover, models struggle with long-range dependencies, complicated dialogue structures, and only little research exists on combining deep learning techniques with external background knowledge.

T4 Exploiting large knowledge bases and graphs
Common-sense and world knowledge from knowledge bases (KBs) and language resources has always been a key ingredient in NLP and plays a central role in NLG, supporting ML techniques that require expansion, filtering, disambiguation or user adaptation of the generated content. One key aspect of research concerns how to integrate existing language resources and knowledge bases into NN models for NLG - e.g. how to process the relevant content, taking into account different data formats. The textual component of NLG systems is improved by the use of resources that are based on a network of semantically structured knowledge, e.g. WordNet, BabelNet, and ImageNet. Dealing with multi-modal and multilingual NLG models requires the integration of additional layers of information. If a resource is multi-modal, its contents can be integrated at training time. Yagcioglu et al. (2015) propose an average query expansion approach for image captioning based on compositional distributed semantics: the original query is expanded as the average of the distributed representations of retrieved descriptions, weighted by their similarity to the input image. Socher et al. (2014) train a NN on ImageNet for building sentence and image representations mapped into a common embedding space using compositional sentence vector representations: image and word representations are learned in their single modalities, followed by mapping them into a common space.

Even if a KB is not multi-modal it can still provide common sense knowledge for NLG. Researchers have recently attempted to utilise KBs to improve NLP. Examples include Wu et al. (2016), who exploit DBpedia in Visual Question Answering (VQA) to better answer questions that require world knowledge, and Chen et al. (2018), who integrate concepts from WordNet to improve a Natural Language Inference model.

1.1.2 DESCRIPTION OF THE CHALLENGE (MAIN AIM)

Multi3Generation aims to foster an interdisciplinary network of research groups working on different aspects of language generation (LG). We frame LG broadly as the set of tasks where the ultimate goal involves generating language. In contrast to the more classical definition of NLG, this also includes tasks not concerned with LG in an immediate sense, but that could obviously inform or improve LG models. The Action will focus on the four core challenges: (1) Data and information representations: in modern applications, inputs can be different sources such as images, videos, KBs and graphs. (2) Machine Learning (ML): modern ML approaches face additional challenges when applied to LG; inputs should be mapped to different correct outputs, i.e. challenges involve e.g. structured prediction and representation learning. (3) Interaction: Applications of LG, e.g. Dialogue Systems, Conversational Search Interfaces and Human-Robot Interaction pose additional challenges to LG due to uncertainty derived from the changing environment and the non-deterministic fashion of interaction. (4) KB exploitation: structured knowledge is key to many NLP tasks, including NLG, supporting ML methods that require expansion, filtering, disambiguation or user adaptation of generated content. This Action aims to address these challenges by answering the following questions:

1. How can we efficiently exploit common-sense, world knowledge and multi-modal information from various inputs such as knowledge bases, images and videos to address LG tasks such as multi-modal machine translation (MT), video description and summarisation?
2. How can modern ML methods such as multi-task learning (MTL), representation learning and structured prediction be leveraged for LG?
3. How can the models from (1) and (2) be exploited to develop dialogue-based, conversational Human-Computer and Human-Robot interaction methods?

The EU has recently published its Digital Single Market priority *Language Technologies and Big Data,* which aims to promote the intelligent use and management of data sources in Europe to facilitate the provision of innovative, commercial and social solutions. Our proposed action will look into how **different data sources** can be aggregated and exploited and how novel ML methods can **assist in automating** the management of data sources as well as the extraction of important information. In addition to this, Multi3Generation is related to the *Automated Translation* policy which aims to provide innovative solutions for cross-lingual access to digital services by addressing the translation workload. Multi3Generation also speaks to the concerns raised in the recent European Parliament resolution on *Language Equality in the Digital Age (2018/2028(INI)), which is premised on the necessity of a multilingual orientation in the development of language technologies.* NLP, NLG and HCI are also included as part of key strategies for R&D plans in other countries outside the EU. In this manner, *The National Artificial Intelligence (AI) Research and Development Strategic Plan*, defined in 2016, sees AI as an urgent priority, establishing as its second main strategy the development of effective human-AI collaboration methods, among which the need for NLP systems capable of engaging in real-time dialogue with humans are emphasised. These types of systems may make the interactions between humans and AI systems more natural and intuitive. In China or the United Arab Emirates, the AI Research and Development Strategic Plans, both launched in 2017, provide goals for setting up the next generation key AI systems, including NLP and autonomous technology to apply to a wide range of sectors (education, health, transport, etc.). The Digital Transformation Monitor report about USA-China-EU plans for AI concludes that 70% of the global economic impact of AI will be concentrated in North America and China. Europe cannot lag behind, and **Multi3Generation** will set up a network of experienced researchers from different European academic institutions, research centres, and companies, who will share their knowledge and foster cutting-edge research and development in the field of AI, able to compete with the US or China. Crucially, this Action incorporate members from leading research-heavy multinational companies in Asia, which can help bridge the knowledge gap between Europe and the rest of the world and also between academia and industry. Several European industries and markets will benefit from the outcomes of this action, including e-commerce (ability to translate multi-modal content and, as a result, reach wider audiences), smart vehicles (through vision and language interfaces), customer service (through automated assistants), media and broadcasting. Finally, the robotics industry will benefit from this action through the development of innovative Human-Robot Interactions. In fact, the *International Federation of Robotics* reports that the sales for personal/domestic robots assistants increased by 31% in 2017 (~6.1 million units/year) and this number could reach almost 39.5 million units within the period 2019-2021. The trend for this market is positive, is expected to increase substantially in the future, creating new opportunities and challenges for Human-Robot communication, which will be addressed by this Action.

1.2 PROGRESS BEYOND THE STATE-OF-THE-ART

1.2.1 APPROACH TO THE CHALLENGE AND PROGRESS BEYOND THE STATE-OF-THE-ART

T1 Grounded multi-modal reasoning and generation
Recent developments in Computer Vision (CV) and NLP have led to a surge of new research problems which move beyond conventional text understanding or visual recognition tasks and which lie at the intersection of these two fields, most notably grounded multi-modal reasoning and generation, which has been applied to visual question answering, multi-modal MT and multi-modal machine comprehension. Despite this, much research on how to integrate the distinct but related modalities is still needed.

Current state-of-the-art methods using deep learning rely on large amounts of annotated training data. The same requirement holds for multi-modal grounding data. One important challenge this Action will address deals with weakening the reliance on large supervised training sets. Topics to address include methods for linking entities of different modalities in an unsupervised way (Yeh et al., 2018), multilingual MT systems without much or any parallel data (Lample et al., 2017), and grounding words using images to train models without parallel data (Nakayama & Nishida 2017).

Although end-to-end neural networks have proven to be very efficient for various multi-modal learning tasks, model interpretability is still a problem. Exploiting attention mechanisms to ground model decisions in various modalities might help to develop more interpretable models (Park et al., 2018) as well as visualisation of the learned representations. Another major challenge is grounding in video, which has only been explored partly (e.g. Haonan et al., 2015) and proves to be very challenging, as e.g. shown by a project on understanding stories in movies through QA (Tapaswi et al., 2016).

Finally, the conditions under which grounding can help generation and vice versa still have to be explored in more depth, as well as the utility of naturally occuring datasets. Existing datasets are mostly created artificially through crowdsourcing, e.g. Multi30K (Elliot et al., 2016), whereas leveraging naturally occurring data, such as done in MovieQA by using movie subtitles and directors' comments, has proven less feasible so far. Promising candidate sources would be those for which videos or images and actions are aligned more clearly, e.g. in teaching videos such as cooking videos. A tangible contribution of this Action network will be the release of two data sets for multi-modal LG: (i) sentences aligned in multiple languages with associated images, used to train multilingual multi-modal machine translation models, and (ii) images with descriptions in multiple languages with additional metadata (e.g. geolocation, emotion tags), which can be used to train richer image description models.

The novel ways to integrate multi-modal data for reasoning and LG to be analysed within this Action will lead to more robust and accurate models. Moreover, knowledge and skills developed by Action participants can drive CV, NLP and related fields. This might, in turn, spawn entirely new innovative applications, such as building autonomous translation systems for e-commerce sites or intelligent agents with the ability for understanding multi-modal data.

T2 Efficient Machine Learning algorithms, methods, and applications to language generation
NLG systems vary in whether they require aligned data or not. As deep learning approaches become better understood, they are likely to feature more heavily in a broader range of NLG tasks. Could NLG be about to witness a renewed emphasis on multi-levelled approaches, with deep architectures whose components learn optimal representations for different sub-tasks? And to what extent would such representations be reusable and knowledge transferable across domains, such as from newswire and social media data? The prospect of learning domain-invariant linguistic representations that facilitate MTL and transfer learning in NLP remains somewhat elusive, despite certain notable successes (Collobert & Weston, 2008; Collobert et al. 2011), not least those scored in the development of distributed word representations. This could well be the next frontier in research on statistical NLG.

Image-to-text generation is one area of NLG with a clear dominance of deep learning methods. This Action will advance the state-of-the-art in this field, by facing a number of current challenges, including generalisation beyond training data (Devlin et al. 2015a); the localisation in images to be able to associate linguistic expressions with parts of images; to produce explanatory descriptions (Hendricks et al., 2016a); moving from static inputs to sequential ones, especially videos (e.g. Venugopalan et al., 2015b, 2015a); or devising generative models of language (Schulz et al., 2018) that are multi-modal.

Within NLP, a recent development is the explosion of interest in social media, including blogs, micro-blogs such as Twitter, and social platforms such as Facebook. These sources together with the extended models proposed in Action will open new lines of application-based innovation scenarios.

Interest in social media could be seen as a natural extension of long-standing topics in NLP, including the desire to deal with language 'in the wild'. However, social media prominently features non-canonical language (Eisenstein, 2013; Baldwin et al., 2013), which most NLP tools struggle with (Plank, 2016), and for which fewer benchmark datasets exist. Through innovations in MTL and transfer learning, the Action will provide more stable end-user ML tools applicable across domains.

There has been a significant shift towards data-driven NN techniques in the past few years, which overall achieve higher results than previous lexical techniques. However, their decisions are hard to understand for humans. NLG can help with this, as an explanation can be automatically generated using the learned NN feature representation (Hendricks et al., 2016). This has the potential to impact a huge range of applications, such as understanding evaluation and tutoring systems in conversational settings for education purposes, or explaining exact procedures hidden in data in eHealth treatment.

To go beyond the state-of-the-art, this Action will research models that exploit training data more efficiently. NNs typically require large amounts of data to achieve good task performance. Being able to utilise transfer between tasks is a way to benefit from data that already exists. This is already done in many situations, notable cases involve transferring knowledge from CNNs pre-trained to classify images to obtain image features. One more way to leverage training data is by devising MTL models that share representations and can benefit from data external to the task. Lines of research that this Action will foster involves transfer learning and the application of MTL for different LG scenarios.

T3 Dialogue, interaction and conversational language generation applications
HCI through conversation in natural language is currently one of most active research fields in which new commercial applications appear almost every day. The so-called artificial intelligence paradigm (in most cases deep learning) is widely used in dialogue-based problem solving. Although current achievements are promising and many global companies and universities have concentrated their efforts in HCI, the challenges of truly intelligent natural language understanding (NLU) and adequate NLG in terms of adequacy of responses and fluency of the generated outputs are not addressed yet. This Action will go beyond the state-of-the-art by applying LG models to Human Computer Interaction (HCI) tasks in several interesting and challenging real-world use-cases, such as conversational search interfaces; grounded dialogue models; real-time dialogue models; and conversational robots.

Due to the complexity of NLU and NLG, much current research focuses on development of end-to-end systems. However, those depend heavily on availability of adequate training data. Lack of necessary datasets (especially speech/multi-modal) limits research in deep learning for HCI. This Action will address this issue by construction of multilingual multi-modal datasets to benchmark text, speech and multi-modal conversational interfaces. As a concrete example, an industry partner will release a dataset with places-of-interest around the globe. Current methods have not solved the problem of having a truly natural conversation. One main issue here is the difficulty of evaluating NLG systems since intrinsic evaluation metrics (e.g. BLEU, ROUGE) are often a poor proxy for model performance (Belz, 2009). This Action will address this issue first by stressing the need for extrinsic evaluation. Concretely, dialogue agents will be tested by industry members in controlled environments.

T4 Exploiting large knowledge bases and graphs
Analysing, reviewing and comparing methods about the use of structured KBs into LG models is a prerequisite for improving the output of these models, raising awareness about resources available and possible uses of them. An expected result is to increase the varieties of knowledge resources and language resources used. This is currently a bottleneck in NLG, where much data-driven work tends to be based on a limited number of data sources, such as the WebNLG (Perez-Beltracchini et al., 2016) and E2E (Novikova et al., 2017) datasets. Progress in NLG thus requires exploiting more existing KBs, especially to clarify which KBs contains information useful for grounded multi-modal reasoning. As sources of knowledge, they can be included in ML methodologies that combine different input types. Looking at existing resources with a focus on NLG means also considering how to expand them with multilingual and multi-modal content. An interesting challenge concerns the study of methodologies to make KBs multi-modal through automatic mapping, crowdsourcing or a combination of both.

Large multi-modal KBs include BabelNet (Navigli and Ponzetto, 2012), which includes over 6M concepts spanning 284 languages, 53M images, and additionally already integrates word senses from multilingual WordNets, images from ImageNet, Wikipedia articles, and other resources. Representation learning is an important step, and recent results on learning representations for multi-modal KBs (Pezeshkpour et al., 2018) or using graph NNs to model relational data (Schlichtkrull et al., 2018) are promising. This Action will address the challenge of devising efficient methods to leverage KBs for NLG, which spans

T3 and T4. To process large-scale multi-modal data, challenges around memory efficiency and model parallelism will have to be solved. An important direction will include models to access external KBs and gather common-sense and world knowledge. This would lessen the reliance of LG models on task-driven training data. Among the many open research questions, it involves KB representation learning and inference, entity linking and disambiguation.

1.2.2 OBJECTIVES

1.2.2.1 Research Coordination Objectives

The main objective of Action is the creation of an interdisciplinary European LG research network targeting scientific advances and societal benefits in the focus themes of (T1) Grounded multi-modal reasoning and generation; (T2) Efficient ML algorithms, methods, and applications to LG; (T3) Dialogue, interaction and conversational LG applications; and (T4) Exploiting large KBs and graphs. The following are the research coordination objectives:

1. Foster knowledge exchange by sharing of resources including semantic annotation guidelines, benchmarking corpora (such as the three datasets mentioned in Section 1.2.1, subsections T1 and T3), ML and alignment tools.
2. Create multimodal and multilingual benchmarks for NLG involves experimenting with automatic mapping between existing resources, crawling of Web data, definition of annotation guidelines and launching of crowdsourcing campaigns for bigger datasets (also as games-with-a-purpose).
3. Facilitate interactions, collaborations, knowledge building and dissemination between Action participants via: a website that (i) provides information about members, their activities and contact details, the ongoing research coordinated by the Action, conferences, dissemination activities, shared task challenges and training opportunities; (ii) contains a Multi3Generation blog and forum discussions, and (iii) makes important publications available for download.
4. Promote the generation of novel ideas and on introduce researchers to the new joint Multi3Generation discipline, via the organisation of training schools including summer schools/conferences.
5. Provide opportunities for joint research projects by Action members on multi-task, multilingual and multi-modal processing during exchange visits of ECIs, and other activities that encourage young researchers to establish links with industry and more senior academics.
6. Disseminate the results of the Action through conferences, scientific and industrial gatherings, which will have substantial impact in the participating countries and beyond.
7. Create synergies between participants via joint publications in books, journals and conferences; reports from working group meetings and training materials from training schools.

Joint applications for European and national funding for research projects within the fields covered by the Action, to encourage novel outcomes and establish critical mass, will be made.

1.2.2.2 Capacity-building Objectives

The overall expected impact of Multi3Generation is to bring about a significant change in progress towards effective solutions for computational challenges involving LG with respect to multi-task, multilingual and multi-modal aspects. In particular, Multi3Generation will focus on the integration of these three aspects and how they can benefit LG solutions. Our specific objectives are:

1. Strengthen European research on **theory, methodology and real-world technology in LG**, particularly in the four Multi3Generation focus research themes (T1-T4);
2. Facilitate international collaboration, networking and interdisciplinary community building by yearly **conferences and workshops** and biannual international **training schools**;
3. Drive scientific progress by liaising extensively with **industry and end-users**, and by increasing joint collaboration and knowledge transfer by the end of the action;

To coordinate the development of **benchmark data resources** for tasks relating to the focus themes above and to organise corresponding shared-task competitions

2 NETWORKING EXCELLENCE

2.1 ADDED VALUE OF NETWORKING IN S&T EXCELLENCE

2.1.1 ADDED VALUE IN RELATION TO EXISTING EFFORTS AT EUROPEAN AND/OR INTERNATIONAL LEVEL

Multi3Generation will liaise and interact with European resource building programmes such as META-NET (FP7 249119, 271022, 270893 and 270899) and CLARIN (FP7 212230) for natural language resources. While there are no COST Actions dedicated solely to LG, there are some that address issues tangentially related where interaction will produce synergetic benefits. The recently finished European Network on Integrating Vision and Language (iV&L Net) COST Action IC1307 was dedicated to bridging the gap between research in NLP and CV, and our proposed Action will fully utilise its outputs where appropriate: published datasets and benchmark challenges, research methods, networks of researchers with NLP and CV expertise, etc. A further action related to Multi3Generation is the PARSEME COST Action IC1207, which also focuses on multilingual learning. However, PARSEME is solely focused on parsing and multi-word expressions, and does not consider challenges arising in the multi-modal context, or machine learning solutions centred around multi-task learning. Another such multilingual action is the Distant Reading for European Literary History COST Action CA16204 with a focus on literacy data and the aim of providing such in 10 languages.

Where possible, Multi3Generation will liaise with these other actions by: (i) regularly informing the respective coordinators on the Action's scientific programme and progress; (ii) co-organising, whenever appropriate, events, e.g. workshops at international conferences of the domain; and (iii) promoting the Action's STSMs and Training Schools among ECIs in these projects and actions.

2.2 ADDED VALUE OF NETWORKING IN IMPACT

2.2.1 SECURING THE CRITICAL MASS AND EXPERTISE

Europe has many of the world's leading researchers in diverse fields, such as Linguistics, CV, Speech, ML, MT, Dialogue Systems, Robotics, and HCI. Multi3Generation will tap into this body of expertise to create new strategic partnerships aimed at bringing together complementary strengths from diverse fields to solve different aspects of NLG from different perspectives. A successful network will place in Europe at the forefront of developing NLG solutions with clear commercial potential. A COST Action with its support for coordination of research, networking, exchange of expertise, and forming new research partnerships, is exactly the right framework for the aims and activities outlined in this document, and this particular form of support is not available under other funding schemes.

Every effort has been made to create a wide-reaching network, initially encompassing 19 scientists from 15 countries in Europe who prepared the proposal, 2 of which at SMEs. The network of proposers involved a balanced panel of expertise from different fields corresponding to T1-T4 to come together to tackle the challenges of NLG. Out of these initial countries, 9 are from ITCs. Multi3Generation emphasises participation of ITCs in STSMs. Each STSM undertaken by an ITC will entail close collaboration with one of the more experienced partners to support technology transfer. NLG research is addressed through different perspectives by groups at universities, research centres and companies spread over Europe. This Action will enhance scientific synergies between disciplines and integrate so far single isolated research efforts on multi-task, multilingual and multi-modal LG. The different types of knowledge and expertise brought together will cross-fertilise algorithmic thinking, and will bring about novel ideas on multi-task, multilingual and multi-modal challenges.

2.2.2 INVOLVEMENT OF STAKEHOLDERS

A central goal of Multi3Generation is to enhance knowledge transfer within academia and between academia and industry. The ultimate beneficiaries of such knowledge transfer are expected to be the end users of systems which, increasingly, rely on the processing of large multi-modal datasets to relay information. With an emphasis on models grounded on multi-modal data, fuelled by state-of-the-art ML machinery, and by definition multilingual, we expect next-generation systems to adapt to the needs of a broader user base, including speakers of low-resource languages. This Action will contribute to this aim by strengthening ties between researchers and industrial partners as they jointly focus on developing such adaptive multi-modal multilingual systems for information dissemination. These developments also have a potential for commercial and economic impact.

At the same time, end-users are in diverse communities, including linguistic communities. Not all languages are equally represented in the digital landscape, as has been repeatedly emphasised in the European context (e.g. through the white papers published by the MetaShare initiative in 2012 (Stelios, 2012) and the recent debate and resolution on language equality in the digital age within the European Parliament). Part of this Action's strategy is to emphasise multilinguality in generation. This includes European languages with small number of speakers (e.g. Welsh, Sami), which are also well-represented in the Consortium. Small under-resourced languages would benefit from advances in multi-task and transfer learning (which raise the possibility of exploiting large-scale resources for neighbouring languages to the benefit of small languages for which training data is harder to come by).

The Consortium's work-plan seeks to address the needs of these three groups of stakeholders -- researchers in academia, industry players and ultimately, end-users -- through a strategy that is both horizontal and vertical. Along the horizontal dimension, our work-plan emphasises synergies among representatives working within the main themes of the Action, enabling transfer and knowledge gap-filling through direct interactions between academics and industrial partners. At the same time, the Action is structured along principal thematic axes that will inculcate a deeper ("vertical") understanding of current trends, requirements and future directions within each working group. Along these two dimensions, the Action will maximise the involvement of stakeholders in the following ways:

- Through regular meetings and Short-Term Scientific Missions, it will enable researchers, especially those in the early stages of their careers, to acquire knowledge and develop skills and networking capabilities by interacting with both academic and industrial partners, thus providing formation opportunities for the next generation in this burgeoning field;
- By developing tools for dissemination and collaboration, through exchanging resources and guidelines, and organising international grand challenges (shared tasks), the Action will enable partners to converge on standards and goals that will facilitate cross-fertilisation;
- By enhancing interaction via conferences and summer schools, as well as encouraging scientific dissemination, it will foster stronger collaborations that will result in solid scientific contributions in the field of multi-task, multi-modal, multilingual generation.
- By involving stakeholders from academia and industry, progress in this field will directly translate into better multi-modal, language-enabled systems, boosting economic impact.

Through its flexible membership structure, the Action will be open to all interested parties. We will actively seek participation of other academic and industrial partners, but also members of NGOs and public bodies for whom there are clear benefits. These include policy-makers and representatives of groups whose needs impact the development of the technologies under consideration, e.g. (a) language councils and other bodies with an interest in multilingual technology; (b) groups representing the needs and interests of individuals with specific needs.

2.2.3 MUTUAL BENEFITS OF THE INVOLVEMENT OF SECONDARY PROPOSERS FROM NEAR NEIGHBOUR OR INTERNATIONAL PARTNER COUNTRIES OR INTERNATIONAL ORGANISATIONS

This Action brings together academic and industrial partners, both European and international, with a strong research mandate and previous record. Among International Organisations (IO) in the industrial sector, this Action includes **Huawei from China** and **Naver from South Korea**. Both companies have a strong presence in Europe: Naver Labs Europe (NLE), a part of Naver Labs, is the biggest industrial research centre in Artificial Intelligence in France and is located in Grenoble, France. Huawei is a large multinational ICT company, and runs 18 research and development (R&D) organisations located in eight European countries (Belgium, Finland, France, Germany, Ireland, Italy, Sweden and the UK), employing 1900 R&D staff only in Europe.

In Huawei Noah's Ark Lab, dialogue modelling is one of the main research areas within NLG, and with many application scenarios, e.g. digital assistants for mobile phones and customer service. Other research topics actively investigated there include interactive story generation from images and videos and question answering, with application scenarios of, for example, improving Huawei's global technical support. Dialogue is directly addressed as one of this Action's four core themes, and Huawei's real-world use cases will greatly benefit not just how to frame the problem but also what are the right ML algorithms and techniques when, for instance, near real-time response time is needed.

Naver's services are an interface between content and people. The ability to present that content succinctly and conveniently to users is a integral part of NLE's research agenda. In particular, NLE is interested in summarising large multi-modal user data, much of which could be redundant or even contradictory. This is an example of an important research challenge where the best solution is still an open question, and NLE plans to partner with universities and public research institutions to tackle this together. This is part of the mandate of NLE, and participating in a large Action network fits it perfectly.

For the participants of this Action in COST member countries, ITC and IPC, it is crucial to have large and research-heavy multinational companies such as Naver and Huawei on board. Both are looking forward to host researchers, as well as sending some of their own for STSMs in European academic institutions. Crucially, this Action network will facilitate the joint collaboration on research project proposals involving Huawei Noah's Ark Lab, NLE, other European industrial partners and academic institutions, and this two-way interaction will strengthen both academic and industry partners.

3 IMPACT

3.1 IMPACT TO SCIENCE, SOCIETY AND COMPETITIVENESS, AND POTENTIAL FOR INNOVATION/BREAK-THROUGHS

3.1.1 SCIENTIFIC, TECHNOLOGICAL, AND/OR SOCIOECONOMIC IMPACTS (INCLUDING POTENTIAL INNOVATIONS AND/OR BREAKTHROUGHS)

For the **scientific community,** the prime benefits of the Action will be (i) advanced methods capable of more accurately processing and generating language, (ii) open access publications of the main scientific findings in world-class conferences and journals, (iii) resources and annotation guidelines relating to such content, and (iv) cross-fertilisation of multi-task, multilingual and multi-modal methods currently in use in the NLP communities focusing on grounded reasoning and generation, effective ML algorithms, human-computer and human-robot generation, and exploiting large knowledge graphs. Moreover, the Action will provide access to other representatives from academia, industry and public institutions, facilitating exchange of knowledge and expertise and fostering new collaborations.

The Action will have substantial impact for **society**. The technologies Multi3Generation will be working towards have the potential to benefit a wide range of different users, especially in a multilingual context such as Europe where official languages, minority languages and not EU languages spoken by immigrants coexist. Improvements in language processing and generation will enhance online search experience in multiple languages, as well as help institutions such as hospitals and police forces to cope with ever growing quantities of images and videos alongside text. The development of ML methodologies applicable to less-resourced languages together with the use of multilingual resources that could act as a bridge between languages can guarantee equal status for every language in the scientific community and in the society. Better understanding of natural language by integrating background common knowledge obtained from the visual medium and knowledge bases has enormous application in language understanding (mapping language to knowledge to be used e.g. in decision making) and access to textual content (e.g., bringing text to life in a virtual world).

The software and services that next-generation NLG facilitate have huge **economic impact**. International and European industry partners in this Action proposal are already invested in research on dialogue modelling, e.g. applied to digital assistants and customer service, and in summarisation of large multi-modal user generated data. These are two concrete examples where this Action will put together academic and industry partners to tackle a problem with a relevant economic impact in these industries' businesses. Chatbots are becoming more and more popular and are being integrated in a

wide variety of domains, e.g. tourism, marketing, or medicine. However, their multi-modality and multiliguality capabilities are still limited. This Action will lead to significant improvements in applications that involve dialogue systems or chatbots, as well as multi-modal LG in general. The team of proposers already includes industry partners, including market leaders; industry and end-user groups will be closely involved in all stages of the Action, including research proposal preparation.

3.2 MEASURES TO MAXIMISE IMPACT

3.2.1 KNOWLEDGE CREATION, TRANSFER OF KNOWLEDGE AND CAREER DEVELOPMENT

The main benefit of this Action will be to bring substantial progress in LG with respect to multi-task, multilingual and multi-modal aspects. For the scientific community, cohesion of European research is increased by creating a network of specialists dedicated to these aspects. Existing approaches have not cast a general view on the multi-faceted challenges (Section 1.1 & 1.3), and often resulted in rather fragmented research. Bringing unity into this landscape, by establishing the proposed network, is exactly the right framework to foster collaboration to reach breakthroughs in the NLP and NLG. New interoperable resources, tools, and training schools will serve a large and versatile community.

The initial team of proposers are participating in many national and European projects related to the themes of the Action. Its participants thus have the necessary skills and experience as well as funds and facilities to achieve the objectives listed in Section 1.2. In their institutions, they are currently supervising a large number of graduate students and postdocs. Therefore, active participation in the training schools, workshops, conferences and STSMs, fostering transfer of knowledge, is expected.

Multi3Generation will help shape the next generation of multilingual technology that learns and adapts to its users. As MT plays an important role in multilingual e-commerce, the research outcomes of the Action are expected to attract the attention of many e-commerce companies. In addition, LG for human-robot interaction is expected to attract attention of the robotics industry. The latter is important as new assistive technology has started to be deployed more in the market, from online search, automatic translation to personal home assistants. While gains are high due to the direct application of the technology via industry liaisons, reaching these intended users can be challenging. Although the Action includes relevant industry participants, two dedicated SMEs with direct access to end users, it will issue periodic newsletters, organise events and activities to initiate collaborations with industry and to communicate and engage with European politicians and stakeholders in the European Commission.

3.2.2 PLAN FOR DISSEMINATION AND/OR EXPLOITATION AND DIALOGUE WITH THE GENERAL PUBLIC OR POLICY

Dissemination is a continuous process to be carried out throughout the entire project duration. The project partners will use the established scientific channels of workshop, conference, and journal publications. In addition, software, data sets, and linguistic resources will be made available to the public to encourage uptake of the research issues addressed in this project and further disseminated through dedicated training schools. Some components of this project will also participate at international benchmarking events via shared tasks for measuring the competitiveness of methods developed within the project against the international state-of-the-art. Multi3Generation will pursue a two-way exploitation strategy: *jointly* via the COST members (e.g., joint technical white papers on the state-of-the art, release of novel benchmark datasets) as well as *individually* by each member (via publications). Results will be disseminated and exploited via the following activities and channels:

- The Action website will be set up with clearly visible and regularly updated project information. It will contain a private section to distribute internal information; and a public part to inform external stakeholders of main events and outcomes. Website news categories will include:
    - A database of the involved researchers;
    - All the Action events (meetings, open conferences, training schools, call for participations, STSMs calls, etc.) to increase the participation in the Action;
    - All training / research material produced by the Action (presentations, results of brainstorming sessions and hackathons, lecturers) for educational and instructional purposes, as well as emanating from within the Action consortia membership.
    - An online service for finding the 'right' collaborators for members looking for partners, with special emphasis on the academia-industry partnerships.

- A repository of all the publications and reports produced by the Action, along with a public repository of data resources and software tools.

- Presentation of the project and its results at thematic national and international workshops and conferences. These targeted publications aim to reach possible national and international stakeholders and will be led by the Action board, but all members will be actively involved in it.
- Participation and organisation of meetings and events with attendants from industry and government promoting the reuse of the Action outcomes in professional applications.
- Development of scenarios, use cases, benchmarks, prototypes and libraries of source code, thus providing to the interested people a flavour of the Action work.
- Publications in leading scientific journals (all partners).
- Training activities will especially target ECIs, including Master and PhD students, who will benefit from the educational and networking opportunities offered by the Action and STSMs. The training events organised by the Action will be the most helpful to ECIs for better understanding the research opportunities offered by all disciplines contributing to LG.
- The Action will harness social media to reach a broader European audience and inform citizens of how COST's investment is used to promote science, technology and society. A mailing list will be created where members can post questions and announcements, open to other interested researchers and professionals. A periodic newsletter will be issued with a wide distribution to academic and industry partners as well as to European politicians and stakeholders in the European Commission. Communication and collaboration among Action participants will be enhanced by technical meetings and visits (50% reserved for ECIs).
- Multi3Generation will produce one joint report or 'white paper' on the state of the field and progress achieved in each of its themes (T1-T4) per year. These will be collaboratively authored; the co-authoring process itself will help to cement relationships, increase mutual awareness and create a sense of the collaborative nature of the COST Action. Moreover, the Action will organise a series of training events on ML and LG, as per our timetable in 3.2.1. These are aimed at the intersection between the contributing disciplines, helping researchers to learn more about the other disciplines that contribute to the field as a whole.
- At end of Year 4, the Action will organise a widely advertised final conference for disseminating the main findings, targeting those interested in NLP, ML, CV, information retrieval and HCI, discussing concrete applications and possible roadmaps for these fields.
- The combination of technical meetings, partner visits and exchanges, white papers, annual workshops with shared task competitions, training events for ECIs, final dissemination event and scientific publications represent a broad dissemination portfolio, complementary in nature. Many dissemination activities will continue beyond the initial 4-year lifetime of the Action.

4 IMPLEMENTATION

4.1 COHERENCE AND EFFECTIVENESS OF THE WORK PLAN

4.1.1 DESCRIPTION OF WORKING GROUPS, TASKS AND ACTIVITIES

The Action will be coordinated by the Multi3Generation Management Committee (MC). MC Chair, Vice-Chair and Secretary will be elected at the Action's Kick-off Meeting. Chair and Vice-Chair will preside over the MC and oversee the work of the five WGs, each of which is managed by a WG Coordinator and Deputy Coordinator. The MC will hold two meetings per year, in conjunction with the WGs meetings. There will be three further MC roles, Scientific Coordinator, STSM Coordinator and Dissemination Coordinator, also to be elected at the Kick-off Meeting.

The Scientific Coordinator will periodically report to the MC on the scientific progress of the Action; oversee the implementation of a shared repository of materials/work; liaise with the WG Coordinators about work to be done; provide advice on research topics based on the Memorandum of Understanding; in collaboration with the WG Coordinators, develop scientific/technical programmes for Action Workshops; together with the Dissemination Coordinator, organise the publication of a series of books/proceedings resulting from the Workshops/Conferences; and generally oversee the scientific activities of the Action, identifying any need for improvement/discussion at MC Meetings.

The STSM Coordinator will, together with the WG Coordinators, oversee the STSMs application process, including calls, selection criteria, reviewing, and ensuring required ECI representation.

The Dissemination Coordinator's tasks will include forming a Dissemination Committee; drafting a Dissemination Plan in the first quarter of the Action; overseeing the design and production of dissemination materials; coordinating the promotional activities for Workshops/Conferences to ensure broad participation; liaising with key industrial and academic partners; periodically publishing an Action Newsletter; maintaining a regularly updated list of Action Participants, stakeholders, end-users and other target audiences and keeping them informed about the Action's activities; generally overseeing the dissemination activities and identifying any needs for improvement/discussion at MC Meetings.

Research coordination will focus around the four main research themes and structured around the five dedicated WGs described in Section 3.1.1. Membership of WGs will be open to all participants and not mutually exclusive. Each WG will be coordinated by its Coordinator and Deputy Coordinator, who will be appointed at the Kick-off Meeting. They will (i) organise and chair WG Meetings, prepare meeting agendas and minutes; (ii) coordinate and review scientific/technical work; (iii) ensure continuous/efficient communication within and across WGs; (iv) periodically prepare reports to the MC; and (v) regularly communicate with the Scientific Coordinator and Dissemination Coordinator.

Multi3Generation will set up an Industrial Advisory Board, organise Annual Conferences, Technical Meetings, Industrial Secondments and Partner Visits (50% reserved for ECIs), and a flexible membership structure. WGs 1-5 will produce reports on the state of the field and progress achieved, coordinate data creation and shared-task competitions in the core research themes, and conduct a road-mapping exercise in the interdisciplinary fields brought together to tackle NLG. This will also take into account developments outside Europe, in view of developments discussed in Section 1.1.2.

Coordinated research under Multi3Generation is structured into **5 working groups (WGs)** each with content and tangible outputs. All WGs are supported by the following types of Action activities:

Advancing Science (AS) activities address research planning and consolidating existing knowledge, as well as identifying research gaps. Initially, the diverse body of knowledge underlying the network will be consolidated, to serve as a basis for creating an ambitious and comprehensive research agenda. The key focus is the cross-fertilisation of ideas through the organisation of thematic, consolidation and research roadmap workshops. ***Education & Training (ET)*** activities are intended to address difficulties posed by the multidisciplinary nature of the Action. The goal is to provide effective mechanisms to bridge gaps between sub-disciplines and help researchers come up to speed research areas. It will be targeted at researchers and graduate students. ***Shared Resources for the Community (SR)*** activities focus on providing a Web-based repository of material that will assist the Action's community in research and in education; it will also increase the outside visibility of LG.

WG1 Grounded multi-modal reasoning and generation. WG1 focuses on utilising multi-modal input (e.g. images, speech) to improve LG, since grounding text on other modalities has shown promising results. Important tasks will be to consolidate existing knowledge on multi-modal MT, multilingual video/image description, develop multi-modal comprehension methods, define annotation standards, and to draft a research roadmap. The complexity of the area due to its currently fragmented make-up and its emerging nature calls for an innovative approach tightly focused on scientific progress.

WG2 Efficient Machine Learning algorithms, methods, and applications to language generation. WG2 focuses on the ML machinery behind state-of-the-art LG models, since much of the improvements we observe across different LG tasks nowadays are due to the application of varied deep neural network architectures. This involves multi-task and transfer learning, representation learning, structured prediction and generative models, among others. Moreover, WG2 investigates integration strategies for multi-modal data, which is of critical importance for Multi3Generation.

WG3 Dialogue, interaction and conversational language generation applications. WG3 focuses on applications of real-time and/or interactive LG models, and should strongly align with the interests of the industrial partners taking part in the Action. We anticipate that text-based dialogue models, should be interesting and challenging real-world use-cases. This involves conversational search interfaces, grounded dialogue models, real-time dialogue models and conversational robots.

WG4 Exploiting large knowledge bases and graphs. This WG focuses on the exploitation of diverse structured KBs and language resources for multi-modal LG tasks. The analysis of how to efficiently integrate these (multi-modal) KBs is paired with the study of theoretical models of semantics and semantic processing that can accommodate linguistic and perceptual information.

WG5 Industry and End-User Liaison. This WG develops links with industry and end-users. In addition to industry Action participants, company stakeholders will be invited to join an *Industrial Advisory Board* with the functions of (i) advising and informing the MC's activities, and (ii) fostering collaboration between industrial and non-industrial participants, including placements for ECIs and co-organisation of shared tasks including construction of benchmark datasets. WG activity embraces collaboration between academic and industrial partners, both on academic projects and on real-world product development, and it will stimulate ideas for novel cross-modal applications. Furthermore, WG5 will coordinate user requirement surveys and other methods for obtaining end-user input.

4.1.2 DESCRIPTION OF DELIVERABLES AND TIMEFRAME

Working groups tangible outcomes (deliverables):

Tangible outcomes	WG1	WG2	WG3	WG4	WG5
Advancing Science (AS)					
In-depth analysis and review of methods for incorporating structured knowledge bases into language generation models				X	
Literature review and state-of-the-art on linguistic theories combining linguistic and perceptual information in semantic representations				X	
Algorithms and strategies resulting in a standard methodology for the analysis of text-based dialogue models			X		
Survey and position papers	X	X	X	X	
Joint journal and conference publications	X	X	X	X	X
Cross-topic research roadmap identifying where research effort is most needed and the research challenges that need to be addressed	X	X	X	X	X
Shared Resources for the Community (SR)					
Semantic annotation guidelines and standards for multi-modal data	X				X
Repository of open source software for processing language and visual content, inc. a directory of sources of materials or components		X			
Construction of training and benchmarking datasets for multi-modal MT, multilingual video/image description and dialogue modelling	X		X		X
Reports on requirements surveys					X
Education & Training (ET)					
Curriculum for a graduate course on multi-modal NLP and LG	X				
System application demos accessible via the Action website					X
Other					
Organisation of training and summer schools with courseware, including a short course on Topics on ML for LG for Multi3Generation members, and courses for the wider community (ET, SR)		X		X	
Organisation of an international workshop or shared task (AS, ET)	X				X
Organisation of Industrial Placements (AS, ET)					X
Articles for general readership (ET, SR)					X
Establishment of an Industrial Advisory Board					X
Reports from WG meetings and STSMs	X	X	X	X	X

The Action milestones are the following:

M1 [end of Month 3] The Action is up and running: All planning documents (Dissemination Plan, WG Work Plans, ECI Action Plan, Gender Balance Direct and Indirect Action Plans, etc.) are available, initial WG recruitment drive is complete, Action Website and the social media channels are up and running, first MC Meeting has taken place.

M2 [end of Year 1] MC and WG targets and deliverables for year 1 have been achieved: Reports/publications from the workshops and conferences held; establishment of methodologies for jointly processing vision and text reflected in reports and training courseware; benchmarking data sets, shared-task definitions, and required annotations for one selected application are available.

M3 [end of Year 2] MC and WG targets and deliverables for year 2 have been achieved: Reports/publications from workshops and conferences held; establishment of methodologies for jointly processing vision and text reflected in reports and training courseware; benchmarking datasets, shared-task definitions, and required annotations for two selected applications are available; first set of tools and demonstrators covering selected applications are available.

M4 [end of Year 3] MC and WG targets and deliverables for year 3 have been achieved: Reports/publications from the workshops and conferences held; establishment of methodologies for jointly processing vision and text reflected in reports and training courseware; benchmarking data sets, shared-task definitions, and required annotations for three selected applications are available; expanded set of tools and demonstrators covering selected applications are available.

M5 [end of Year 4] MC and WG targets and deliverables for year 4 have been achieved: Final conference has been held; reports/publications from the workshops/conferences held; establishment of methodologies for jointly processing vision and text reflected in reports and training courseware; benchmarking data sets, shared-task definitions, required annotations for three selected applications are available; final set of tools and demonstrators covering selected applications are available.

4.1.3 RISK ANALYSIS AND CONTINGENCY PLANS

The MC will be responsible for ensuring that the Action runs according to plan, reacting quickly to take immediate contingency measures. Every WG will have a leader and a deputy leader who will assure the timely production of planned outputs and outcomes and facilitate the MC in the coordination tasks by continuously monitoring the WG's performance in terms of involvement of the participants, and the scientific and the industrial activities that are being organised. The initial proposers include researchers and industrial partners with deep expertise in their respective areas and experience in collaborative projects. This will help the network to detect and address the risks. Potential risks are:

Risk	**Contingency measure**
Unlikely	
Low participation in training schools & STSMs	Increase promotion of these activities and stimulate attendance by increasing the financial support.
Shortage of linguistic or evaluation expertise for some languages.	Input from a smaller number of experts is solicited. Liaise with language councils at national level to identify support mechanisms and data sources.
Possible	
Shortage of linguistic or evaluation expertise for some languages.	Input from a smaller number of experts is solicited. Liaise with language councils at national level to identify support mechanisms and data sources.
Delays in deliverables & milestones	Invite people with the necessary expertise and encourage them to collaborate with the network.
Insufficient engagement of industry & end users	Organise additional public and industry-led events to promote their participation. Organise activities within each partner country to promote the Action among local industry stakeholders.
Bottlenecks in data collection & resource building	Investigate automatic data augmentation techniques, including Web-based retrieval, to bootstrap from small datasets. Exploit crowdsourcing platforms for short-term data collection initiatives.
Under-resourced languages	For under-resourced languages, strategies such as active learning,

cannot be adequately treated	transfer learning and fine-tuning of models for better-resourced languages can help overcome the bottleneck.
MTL proves ineffective given the current state of the art	Consider the relative gains to be made by viewing NLG as a collection of sub-tasks, compared to the current tendency to exploit end-to-end models.

4.1.4 GANTT DIAGRAM

The Action duration is 4 years, and the table below provides an overview of activities tentatively planned.

MC meetings: The initial MC meeting is the kick-off meeting and will officially start the Action. 2 MC meetings will be organised annually, at least 1 of which will be co-located with other activities. **WG meetings:** Each WG will organise 8 WG meetings. WG meetings at the end of Years 2 and 4 will be organised jointly. Year 1 and 3 meetings will be co-located with an international conference. **Training schools**: A short training session for Action participants will be organised jointly with the kick-off. Two subsequent training schools will be organised in Year 2 and Year 4, on ML and LG. **Workshops, shared tasks and final conference**: Three open access workshops will be organised in conjunction with international conferences in Years 1-3. During the final open access conference at the end of the Action, international experts whose research is relevant to the Action will be invited. The Action will organise one shared task yearly following the working groups themes. **Other dissemination activities:** Dissemination via the public website and social media will start immediately and communication/project management tools will be installed on the internal website. **STSMs**: Throughout the Action, STSMs will be set up within and between WGs. At least four STSMs per year will take place to stimulate collaboration between members of different WGs (WG1 to WG5).

Activity	**Year 1**				**Year 2**				**Year 3**				**Year 4**			
Management Committee meetings	x		x		x		x		x		x		x			x
Working Groups meetings	x		x		x		x		x		x		x			x
Training schools							x						x			
Workshops and Conferences			x				x				x					x
Other disseminations	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Short-Term Scientific Missions	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x

References

Adolfo, B., Lao, J., Rivera, J. P., Talens, J., Ong, E., (2017), Generating Children's Stories from Character and Event Models. International Workshop on Multi-disciplinary Trends in Artificial Intelligence, pp. 266-280.

Barsalou L.W. 2008 Grounded cognition. Annu. Rev. Psychol. 59, 617–645.

Basile, V. (2014), WordNet as an Ontology for Generation. WebNLG 2015 1st International Workshop on Natural Language Generation from the Semantic Web, Jun 2015, Nancy, France.

Bernardi, R. and Cakici, R. and Elliott, D. and Erdem, A. and Erdem, E. and Ikizler-Cinbis, N. and Keller, F. and Muscat, A. and Plank, B. (2016). Automatic Description Generation from Images: A Survey of Models, Datasets, and Evaluation Measures. JAIR.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. Journal of Machine Learning Research, 12(Aug), 2493-2537.

Collobert, Ronan, and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. Proc. of ICML, 2008.

Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In Proc. of CVPR, pp. 2625-2634.

Ferrucci, D. Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A. Kalyanpur, Adam Lally, J. William Murdock, Eric Nyberg, John Prager, Nico Schlaefer, and Chris Welty. Building Watson: An Overview of the DeepQA Project. Published in AI Magazine Fall, 2010.

Firat, O., Cho, K., Sankaran, B., Vural, F. T. Y., & Bengio, Y. (2017). Multi-way, multilingual neural machine translation. Computer Speech & Language, 45, 236-252.

Gerani, S., Y. Mehdad, G. Carenini, R. T. Ng, and B. Nejat. 2014. Abstractive summarization of product reviews using discourse structure. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).

Huang, P. Y., Liu, F., Shiang, S. R., Oh, J., & Dyer, C. (2016). Attention-based multimodal neural machine translation. In Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers (Vol. 2, pp. 639-645).

Ilievski, I., & Feng, J. (2017). Multimodal Learning and Reasoning for Visual Question Answering. In NIPS, pp. 551-562.

Jing, H. (1998), Usage of wordnet in natural language generation. In Proc. of COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems.

Kottur, S., Vedantam, R., Moura, J. M., & Parikh, D. (2016, June). Visual word2vec (vis-w2v): Learning visually grounded word embeddings using abstract scenes. In Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on(pp. 4985-4994). IEEE.

Lample, G., Denoyer, L., & Ranzato, M. A. (2017). Unsupervised Machine Translation Using Monolingual Corpora Only. arXiv preprint arXiv:1711.00043.

Luong, T., Pham, H., & Manning, C. D. (2015). Effective Approaches to Attention-based Neural Machine Translation. In Proc. of EMNLP.

Martins, André F. T. and Astudillo, Ramon (2016). From Softmax to Sparsemax: A Sparse Model of Attention and Multi-Label Classification. In International Conference on Machine Learning (ICML).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in Neural Information Processing Systems (NIPS), pp. 3111-3119.

Mostafazadeh, N., Brockett, C., Dolan, B., Galley, M., Gao, J., Spithourakis, G., & Vanderwende, L. (2017). Image-Grounded Conversations: Multimodal Context for Natural Question and Response Generation. In Proc. of IJCNLP.

Nakayama, H., & Nishida, N. (2017). Zero-resource machine translation by multimodal encoder–decoder network with multimedia pivot. Machine Translation, 31(1-2), 49-64.

Navigli, R. and S. Ponzetto (2012), BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. Artificial Intelligence, 193, Elsevier, pp. 217-250.

Park, D. H., Hendricks, L. A., Akata, Z., Rohrbach, A., Schiele, B., Darrell, T., & Rohrbach, M. (2018). Multimodal Explanations: Justifying Decisions and Pointing to the Evidence. In Proc. of CVPR.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In Proc. of EMNLP, pp. 1532-1543.

Piperidis, Stelios. The META-SHARE Language Resources Sharing Infrastructure: Principles, Challenges, Solutions. LREC. 2012.

Schlichtkrull, M. and Kipf, T. N. and Bloem, P. and Berg, R. van den and Titov, I. and Welling, M. (2017). Modeling Relational Data with Graph Convolutional Networks. arXiv preprint arXiv:1703.06103, 2017.

Schulz, P. and Aziz, W. and Cohn, T. (2018). A Stochastic Decoder for Neural Machine Translation. Proc. of ACL, 2018.

Socher, R., Karpathy, A., Le, Q. V., Manning, C. D., & Ng, A. (2014). Grounded Compositional Semantics for Finding and Describing Images with Sentences. Transactions of the Association for Computational Linguistics, 2, 207–218.

Tapaswi, Makarand, et al. MovieQA: Understanding stories in movies through question-answering. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

Vijayakumar, A., Vedantam, R., & Parikh, D. (2017). Sound-Word2Vec: Learning Word Representations Grounded in Sounds. In Proc. of EMNLP, pp. 920-925.

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015, June). Show and tell: A neural image caption generator. In CVPR, pp. 3156-3164.

Yagcioglu, S., Erdem, E., Erdem, A., & Cakici, R. (2015). A Distributed Representation Based Query Expansion Approach for Image Captioning. In Proc. of ACL.

Yeh, R. A., Do, M. N., & Schwing, A. G. (2018). Unsupervised Textual Grounding: Linking Words to Image Concepts. arXiv preprint arXiv:1803.11185.

Yu, Haonan, et al. A Compositional Framework for Grounding Language Inference, Generation, and Acquisition in Video. J. Artif. Intell. Res.(JAIR) 52 (2015): 601-713.