Brussels, 23 June 2017

COST 015/17

DECISION

Subject: **Memorandum of Understanding for the implementation of the COST Action "Distant Reading for European Literary History" (DISTANT-READING) CA16204**

The COST Member Countries and/or the COST Cooperating State will find attached the Memorandum of Understanding for the COST Action Distant Reading for European Literary History approved by the Committee of Senior Officials through written procedure on 23 June 2017.

MEMORANDUM OF UNDERSTANDING

For the implementation of a COST Action designated as

COST Action CA16204
DISTANT READING FOR EUROPEAN LITERARY HISTORY (DISTANT-READING)

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);
b. "COST Action Proposal Submission, Evaluation, Selection and Approval" (COST 133/14);
c. "COST Action Management, Monitoring and Final Assessment" (COST 134/14);
d. "COST International Cooperation and Specific Organisations Participation" (COST 135/14).

The main aim and objective of the Action is to This Action will develop the resources and methods necessary to change the way European literary history is written. Through a shared theoretical and practical framework, it will enable sophisticated computational methods of analysis of large collections of literary texts and foster insight into cross-national, large-scale evolutions across European literary traditions.. 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 44 million in 2016.

The MoU will enter into force once at least five (5) 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.

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.

TECHNICAL ANNEX

OVERVIEW

Summary
This Action's challenge is to create a vibrant and diverse network of researchers jointly developing the resources and methods necessary to change the way European literary history is written. Grounded in the Distant Reading paradigm (i.e. using computational methods of analysis for large collections of literary texts), the Action will create a shared theoretical and practical framework to enable innovative, sophisticated, data-driven, computational methods of literary text analysis across at least 10 European languages. Fostering insight into cross-national, large-scale patterns and evolutions across European literary traditions, the Action will facilitate the creation of a broader, more inclusive and better-grounded account of European literary history and cultural identity. To accomplish this, the Action will:

1. build a multilingual European Literary Text Collection (ELTeC), ultimately containing around 2,500 full-text novels in at least 10 different languages, permitting to test methods and compare results across national traditions;
2. establish and share best practices and develop innovative computational methods of text analysis adapted to Europe's multilingual literary traditions;
3. consider the consequences of such resources and methods for rethinking fundamental concepts in literary theory and history.

The Action will contribute to the development and distribution of methods, competencies, data, best practices, standards and tools relevant to Distant Reading research. This will not only affect the way scholars in the Humanities do research, but also the way institutions like libraries will make their holdings available to researchers in the future. The Action will foster distributed research, the systematic exchange of expertise, and the visibility of all participants, activities and resources.

Areas of Expertise Relevant for the Action	**Keywords**
● Languages and literature: Literary theory and comparative literature, literary styles ● Languages and literature: Linguistics: formal, cognitive, functional and computational linguistics ● Languages and literature: Databases, data mining, data curation, computational modelling	● computational stylistics ● literary history ● computational linguistics ● digital humanities

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

Research Coordination
● To coordinate the creation of a multilingual European Literary Text Collection (ELTeC) in three iterations.
● To use the ELTeC to establish best practices and develop innovative methods of Distant Reading for the multiple European literary traditions.
● To engage in an investigation into the theoretical, methodological and practical consequences of Distant Reading approaches for literary history and literary theory.

Capacity Building
● To foster the acquisition of state-of-the-art methods of Distant Reading, including competencies relating to data curation, standards, best practices and methods of Distant Reading analysis.
● To encourage and support the submission of competitive grant proposals both at the national and European levels.

● To help address the current gender imbalance among practitioners of Distant Reading research.

TECHNICAL ANNEX

1. S&T EXCELLENCE

1.1. Challenge

1.1.1. Description of the Challenge (Main Aim)

The challenge of this Action is to create a vibrant and diverse network of researchers jointly developing the resources and methods necessary to change the way European Literary History is written. Grounded in the Distant Reading paradigm (i.e. using computational methods of analysis for large collections of literary texts), the Action will create a shared theoretical and practical framework to enable innovative, sophisticated, data-driven, computational methods of literary text analysis on a large scale and across at least 10 European languages. Fostering insight into cross-national, large-scale patterns and evolutions across European literature, the Action will facilitate the creation of a broader, more inclusive and better-grounded account of European literary history and cultural identity. To accomplish this, the Action will:

1. build a multilingual European Literary Text Collection (ELTeC), ultimately containing around 2,500 full-text novels in at least 10 different languages, permitting to test methods and compare results across national traditions;
2. establish and share best practices and develop innovative computational methods of text analysis adapted to Europe's multilingual literary traditions;
3. consider the consequences of such resources and methods for rethinking fundamental concepts in literary theory and history.

The Action will contribute to the development and distribution of methods, competencies, data, best practices, standards and tools relevant to Distant Reading research. This will not only affect scholars in the Humanities and the way they do research, but also the way institutions like libraries will make their holdings available to researchers. The Action will foster distributed research, the systematic exchange of expertise, and the visibility of all participants, activities and resources.

1.1.2. Relevance and timeliness

The considerable relevance of the Action results from the fact that **European literary history** is an essential aspect of the history of European cultural production and cultural heritage. The literary culture of modern Europe is marked by the coexistence of literary traditions in many different languages that strongly influenced each other. A firm understanding of the highly complex internal relations and evolutions across centuries and languages is crucial to insight into the foundations, formation, and differentiation of a pluralistic European identity. Creating innovative, new ways of assessing, analysing and comparing this rich cultural production manifested in tens of thousands of literary works in many languages, increasingly available in digital form, is paramount if Europe is to understand its cultural coherence, variety, and differentiations.

In this venture, all **European languages and literatures, including the less-researched and less-resourced ones**, have an essential role to play. One of the main goals of the Action is to detect common trends or patterns of influence among many European literary traditions. This can only be achieved by adapting computational techniques to the wide range of European languages. By building a closely-knit and inclusive network of researchers creating resources, methods and tools for cross-linguistic access to the European digital literary heritage, the Action ultimately aims to facilitate the creation of an account of European literary history that is broader, more inclusive and better-grounded than what traditional research methods can achieve.

The Action is particularly timely because for the first time, **digitization efforts, led in particular by national libraries**, have made available increasingly large parts of the literary production. Examples include the Bibliothèque nationale de France's "Gallica" digital library (with over 4 million documents, 265,000 of which are full-text documents) or Europeana (currently referencing over 1 million full text

documents). Even though digital facsimiles or imperfect full-text are only a starting point for Distant Reading research, the existence of such large parts of the literary production across Europe in digital form is the *conditio sine qua non* for effectively enabling cross-European, comprehensive, computational approaches to the history of European literature for the first time. Taking advantage of the new possibilities of large amounts of digital literary texts is a game-changer for Distant Reading and the beginning of a new research paradigm for literary history.

The Action is relevant and timely also because we are currently in a **phase of fast-paced methodological development** in many disciplines relevant to Distant Reading, among them Computational Linguistics, Applied Computer Science, and (Digital) Literary Studies. Measures for stylistic similarity of texts are currently being improved, raising the reliability of new authorship attribution studies. Also, more nuanced and intelligent access to text is finally becoming possible. Techniques like Named Entity Recognition or Sentiment Analysis make it possible to move beyond the surface of texts (which to the computer is fundamentally just a sequence of characters) and instead make implicit, latent, and/or semantic structures in texts explicit (such as names of historical or fictional people, or the positive or negative evaluation present in a phrase). This means that the seemingly opposing domains of detailed and interpretive inquiry into literary texts on the one hand, and of computational and quantitative methods on the other, are seeing an unprecedented **opportunity for convergence**.

In addition, the Action is timely because in the last five years or so, a wide range of research activities in Distant Reading has emerged in several European countries and the field currently faces **rapid methodological developments**. Capturing, channelling and coordinating this effervescence of activity seems crucial at this point in time. In this way, bringing together researchers from different European countries and from computational linguistics and digital literary studies, will bring additional momentum to the future development of Distant Reading as an even stronger field, for the emergence of a truly European literary history, and for an innovative, strong and inclusive European Research Area.

1.2. Specific Objectives

1.2.1. Research Coordination Objectives

There are three main research coordination objectives.

Objective 1: to coordinate the **creation of a multilingual European Literary Text Collection (ELTeC)**. The existence of such a collection is an essential condition for the creation of tools and methods of analysis comparable in nature, scope and quality across several European languages. The ELTeC will be built in three iterations:

- 1st iteration: 6 subcollections (100 novels per language) for the period ca. 1850 to 1920, providing a starting point for research.
- 2nd iteration: at least an additional 4 subcollections (100 novels per language) for the same period, completing the “ELTeC core”.
- 3rd iteration: extensions to the “ELTeC core” with at least 6 additional subcollections (a) in additional languages, widening the range of ELTeC, (b) for languages already included, but the earlier period from ca. 1780 to 1850, enabling diachronic views on literary history and (c) with additional, larger but less strictly structured subcollections for languages already included, providing a broader empirical base for specific analyses.

The ELTeC core will contain at least 10 linguistically annotated subcollections of 100 novels comparable in their internal structure in at least 10 different European languages (Dutch, English, French, German, Modern Greek, Italian, Polish, Portuguese, Russian and Spanish), totalling at least 1,000 full-text novels. The extended ELTeC will take the total number of full-text novels to at least 2,500. (Novels have been chosen among major literary genres for availability and size. Chronological limits are due to constraints related to copyright and availability of quality full texts.)

Objective 2: to use the ELTeC to **establish best practices and develop innovative methods of Distant Reading** for the multiple European literary traditions. Distant Reading methods cover a wide range of computational methods for literary text analysis, such as authorship attribution, topic modelling, character network analysis, or stylistic analysis. The Action will foster the adaptation of existing methods to multiple languages and to earlier historical periods and develop strategies to obtain comparable results across languages despite the specifics of each language. Based on new or improved tools and methods tested with the use of ELTeC, the Action hereby enables innovative, data-driven research into the rich European literary traditions.

Objective 3: to engage in a much-needed **investigation into the theoretical, methodological and practical consequences of Distant Reading** approaches for literary history and literary theory. Distant Reading research works in a comparative and multilingual perspective rather than on only one language, bases its research not on a small number of representative and/or outstanding texts but on

a wide spectrum of the literary production, and does so with computational and quantitative methods of text analysis. This has far-reaching consequences for how concepts like authorship, style or genre at the heart of literary theory and phenomena like canonization, intertextuality or periodization central to literary history are understood.
With these objectives, the Action aims to strike a balance between resource creation, methodological innovation and theoretical considerations and to engage with several distinct groups of stakeholders (detailed below, section 2.1.1).

1.2.2. Capacity-building Objectives

In terms of capacity building, the Action pursues several related objectives.
Objective 1: to foster the **acquisition, especially by Early Career Investigators (ECIs), of state-of-the-art methods of Distant Reading**, including competencies relating to data curation, standards, best practices and methods of Distant Reading analysis such as authorship attribution, computational stylistics, network analysis or topic modelling. The main motors of this will be **Training Schools (TS) and Short-Term Scientific Missions (STSM)**, through which researchers at participating institutions will expand their methodological toolbox and boost the range and sophistication of their research, with lasting effects on research quality and career options.
Objective 2: to encourage and support the submission of competitive grant proposals both at the national and European levels, e.g. the Polish National Science Centre grants, the Franco-German DFG-ANR (Deutsche Forschungsgemeinschaft / Agence Nationale de la Recherche) program, or Horizon 2020 calls, specifically Marie Skłodowska-Curie Actions or ERC (European Research Council) Starter or Consolidator Grants. The Action will be able to do so in two ways:

- by encouraging the creation of focused consortia from among the members of the Action's network, joining forces in preparing grant proposals on topics relevant to the Action;
- by supporting participants' relevant funding proposals through their association with the existing, larger network structure of the Action.

This should happen both at the levels of the individual Working Groups (WGs), where more focused grant proposals can be developed, and at the cross-WG level, where more complex and thematically and/or methodologically overarching proposals can be developed.
Objective 3: to help address the current gender imbalance. Most practitioners of Distant Reading are male and many (but not all) have close links with computer science or computational linguistics. Stereotypes surrounding computer scientists, research has shown, may add to the barrier to inclusion for women in related fields. The Action members have seen, however, that the scarce female colleagues in the field do attract more female than male students. The Action will make use of this and adopt a mission statement to significantly enhance participation of women in the Action. This will include concrete measures making it easier for (female and male) ECIs to combine TSs and STSMs with family life (e.g. organizing local child care). This is to the benefit of the field that requires a broad base of practitioners and thrives on diversity and new ideas.
As described above (section 1.2.1), the **creation of the ELTeC** and of the standards and best practices developed for its creation is in itself capacity building for innovation in literary research, because these resources will enable the researchers in the network and beyond:

- to use the new (digital) resources and innovative tools for their research,
- to expand the text collection by building compatible collections around ELTeC, and
- to expand on the tools and methods by building on those developed in this initiative.

In this way, they will take the Action's activities beyond the scope and lifetime of the Action itself.

1.3. Progress beyond the state-of-the-art and Innovation Potential

1.3.1. Description of the state-of-the-art

The computational study of literary texts came to produce its first notable successes in the late 1960s with the availability of digital texts and computers. More recent developments have fostered **the rise of Distant Reading as a specific, growing and successful field** at the intersection of (Digital) Literary Studies, Computational Linguistics, and Applied Computer Science:

- the availability of growing amounts of digitally available literary texts (especially in English);
- the development of ever more sophisticated algorithms and methods as well as software supporting them (in Computational Linguistics and in Text Mining/Machine Learning);
- a developing consciousness that the conceptual basis of literary history and literary theory will need to be adapted to these new developments.

Together, these developments have already fostered an increasingly broader scope, increased usefulness and wider uptake of Distant Reading methods. (All references to scholarly work in this section are listed separately in the reference section.)

The above-mentioned digitization efforts, particularly by national libraries across Europe, are making **more and more literary texts available for research**. These initiatives provide the basis for the creation of well-structured, quality text collections with rich, reliable metadata and additional annotations. Distant Reading methods applied across language boundaries create new research opportunities but require such data. While such text collections of multiple European languages exist for non-literary texts (for instance built from European Community materials, such as the *EuroParl* corpus or the *RC-Acquis* multilingual parallel corpus), similar multilingual collections of literary texts, like the one this Action seeks to build, do not exist.

As for the **basic linguistic annotation of such text collections**, although tools for basic levels of linguistic annotation exist for a wide range of languages, these tools are in many cases language-dependent and sometimes do not adhere to shared standards that would make their results comparable across languages. Again, the availability of shared and comparable annotations is a prerequisite for Distant Reading methods applied to a multilingual tradition. In respect both to the availability of corpora and of linguistic annotations, the situation for Distant Reading in Europe is challenging. Concentrating on English texts, researchers in the United States and elsewhere have been able to use well-known authorship attribution cases or shared tasks as benchmark cases, with communities forming around them. Achieving a similar goal is more demanding given the multilingual and multicultural environment in Europe, but extremely rewarding in the long run.

Advances in Computational Linguistics mean that currently, a **much wider range of stylistic, semantic and structural features** than before can be made explicit in texts and used for analysis: today, analysis can go beyond simple surface features (such as characters or words) and use part-of-speech tags, Named Entities, WordNet semantic annotations, temporal expressions, and syntactic structures for analysis. This provides Distant Reading research with layers of information that allow a much more nuanced and targeted access to literary texts, helps to get a better understanding of specific literary phenomena, and enhances the reliability of stylistic observations. This now also includes earlier historical states of (some) languages, opening up a diachronic perspective, but with remaining challenges connected e.g. to spelling normalization. Also, there is little agreement for shared annotation schemas across languages and many languages remain under-resourced with respect to annotation tools.

At the same time, methods newly developed in **Text Mining**, such as Topic Modeling, Sentiment Analysis or Network Analysis, and more generally new machine learning methods, are currently in the process of being adapted for Distant Reading research, in the broader context of the Digital Humanities. This means that the derivation of higher-order features and their use for further analytic questions (for example, the relation between topic models and genre, or network structure and periodization) has also become much more sophisticated. Again, however, the full potential of methods which are language-dependent (such as Sentiment Analysis) or which need to be adapted specifically to literary texts (such as Network Analysis), has not been fully realized yet.

In the area of Distant Reading itself, there are two current trends: The first are **more thorough methodological investigations** in well-established areas. Authorship Attribution is a point in case, where text similarity measures specifically designed for questions of authorship in literary texts have been designed and their reliability continues to be improved. The cross-linguistic evaluation of authorship attribution techniques is an emergent trend (with scholars working with Polish, German and French in addition to English) but would need to be adopted more widely. The second trend is a **widening scope of methods and research questions**. Distant Reading research deals not just with Authorship Attribution any more, but addresses many other fundamental issues in literary theory and history, among them questions of author gender, literary period, literary quality, national traditions, and translator's style. Willard McCarty has designated these and similar efforts as “pioneering work” and calls them “the great exception to the stalemate” (he perceives in some other areas of Digital Humanities).

Distant Reading research is one of the most active areas in Digital Humanities. For this reason, the established field of Literary Studies is increasingly becoming aware of the potential for methodological innovation which Distant Reading carries and of the **opportunity for and usefulness of a convergence** between (interpretive) literary history and (quantitative) Distant Reading. However, much more systematic work needs to be done in order to take stock of the new perspectives opened by Distant Reading for literary theory and history.

A final aspect of the state-of-the-art concerns the **structure of current research practice in Distant Reading research**, which is a rapidly growing, successful research field, that is also, however, highly heterogeneous and fragmented. Co-operations are sporadic and based on direct personal contact between individuals, rather than open, systematic and sustained. Research is often characterised by redundant efforts to establish requirements, a lack of a common theoretical ground and the parallel development of solutions for related problems. This is true on the levels of standards for data, tools for analysis, and theoretical investigations.

1.3.2. Progress beyond the state-of-the-art

The Action will build on the state-of-the-art and help mitigate several of the challenges or current needs for Distant Reading research to thrive.

First, the Action will coordinate the building of a multilingual **European Literary Text Collection (ELTeC)**. Such a collection of freely distributable, public-domain literary texts from several national European literatures will greatly facilitate the establishment of best practices and comparative studies for Distant Reading methods comparable in nature, scope and quality across at least 10 European languages. It will serve as a benchmark collection for common tasks in Distant Reading and create incentives for the development of methods, tools and algorithms for specific issues.

Second, the Action will address the prevailing **need for the adaptation of tools and methods**, especially from better-resourced languages to lesser-resourced ones. The aim is to make a shared set of annotations, language-specific linguistic resources, and tools for analysis available to researchers working with texts from the European literary traditions. A shared, comparable set of annotations will help make analyses based on ELTeC comparable across languages. Analysis tools will allow conducting investigations into several national literary traditions with comparable results for each literary tradition. Adapting existing tools to new domains or languages helps share and spread existing expertise and best practices and rapidly improve the cross-European availability of tools.

The Action will encourage and practice **Distant Reading also in a diachronic perspective**. The ELTeC, in its third iteration, will cover the extended time period of 1780-1920 and will make research into questions of chronology possible. The Action will connect the challenge of change over time (lexical and syntactic change, shifts in the system of literary subgenres, historical development of narrative devices) to the challenge of Distant Reading, thereby opening the field to issues which are currently not widely addressed.

Finally, the Action will foster and coordinate research into the consequences of the use of Distant Reading methods for matters of literary theory and history. Established and fundamental concepts in literary theory (such as style, theme or genre) and literary history (such as periodization, canonization or literary complexity and quality), which are being challenged by Distant Reading methods, will need to be re-conceptualized in the light of these developments.

Generally speaking, the Action will **help connect the existing actors in the field and significantly broaden the field's base** by disseminating methods, competencies, data, best practices, standards and tools to a wide range of researchers. The training and dissemination activities of the Action target specifically ECIs, with special consideration given to female ECIs, in order to foster the emergence of a new generation of scholars in the humanities who tackle old and new research questions with the help of the latest tools and methods for Distant Reading.

1.3.3. Innovation in tackling the challenge

The ELTeC will be **published with a permissive Creative Commons open access license** and will be free to download, modify and use in its entirety according to researchers' needs. It is unique in that it will contain, for each language, 100 complete novels, rather than text samples. The ELTeC is special also in that it will offer **collections comparable in structure**, as far as possible, with regard to time period, number of different authors, number of texts by each author, author gender distribution, and perceived literary quality across at least 10 languages. It is not a parallel corpus; each collection will contain a different set of novels in their original language. For each novel, **detailed and identically structured metadata** will be made available, notably with regard to authorship (name, date of birth, gender), date of first publication, perceived literary quality (low-/mid-/high-brow), narrative perspective and novelistic subgenre.

Text curation will happen on GitHub (with distributed access, issue tracking and version control); complete collections will be published on the Action's portal (for best visibility) and archived in a dedicated Zenodo.org community (for long-term availability independent from the Action's lifetime). Experience with similar, national initiatives shows the importance of openness, standards and technical sustainability. As the Action progresses, linguistic annotation will be added to the texts (at least, concerning lemmata, part of speech and named entities). A shared format for the representation of document-level metadata and linguistic annotation across subcollections will be used, based on best practices in the field as recommended e.g. by DARIAH (Digital Research Infrastructure for the Arts and Humanities) and CLARIN (Common Language Resources and Technology Infrastructure). The ELTeC will contain linguistic annotation in a cross-linguistically compatible manner by mapping each language-specific tagset onto a coarse-grained, shared tagset.

In this way, the ELTeC will make it possible, for the first time, to **use and evaluate various Distant Reading methods and algorithms across many different languages**. For example, a typical application in the domain of computational authorship attribution is to assess the stylistic similarity of texts. A wide range of methods for this assessment of stylistic similarity exist, and current research

has shown that different similarity measures perform differently for different languages, but a systematic assessment of the underlying mechanisms for these differences is lacking. Evaluation across languages, on the basis of the ELTeC, is a practical means and a strong incentive for the improvement of many other techniques of linguistic annotation or literary analysis, be it named entity recognition, direct/indirect speech identification, or topic modelling. The availability of the ELTeC will make such an assessment possible for the first time.
To support the annotation of the ELTeC and foster cross-linguistic analyses of the ELTeC materials, **work on the adaptation of tools and methods** is necessary and will concern several areas:

- Adaptation from non-fictional text types (such as newspaper articles) to literary texts. For example, Literary Network Analysis (and Named Entity Recognition, on which it relies) need to take into account the specific ways in which literary texts refer to fictional characters and places, ways that are quite distinct from non-fictional real-world figures or places.
- Adaptation from one language (e.g. a better resourced one) to another. For example, the demarcation of direct and indirect speech in novels follows different conventions in various national literary and editorial traditions, and tools aiming to automatically identify direct speech need to take these differences into account.
- Adaptation to texts from earlier literary periods, with their historical spellings, semantic change and syntactic specificities. A part-of-speech tagger's performance, for example, may drop drastically if it is not specifically adapted to the older type of language.

Finally, with regard to literary theory and history, the Action will foster the re-conceptualization **of fundamental concepts and phenomena**. Distant Reading methods require, for reasons of methodological rigour and statistical significance, that concepts be clearly defined and operationalized in a way to permit formal identification and quantification. Many of the concepts central to literary studies, however, are highly complex and composite and the level of consensus in the field is typically quite low. This Action will strive to show that, far from being incompatible with literary studies, Distant Reading's requirements can help deconstruct and rethink established concepts in literary studies. Conversely, such re-conceptualization will provide a firmer ground to the further development of the research agenda of Distant Reading. This is a highly innovative and controversial issue likely to spawn lively debates between proponents of Distant Reading and mainstream literary scholars.

1.4. Added value of networking

1.4.1. In relation to the Challenge

For the future development of European Distant Reading research as a strong field and for the emergence of a truly European literary history to become reality, the coordination of actors from different European countries with their languages and national literary traditions is essential, because no single discipline and no single linguistic community can cover the entire ground alone. The creation of the ELTeC is **possible only collaboratively and internationally**. For each subcollection, the texts need to be quality-checked, metadata collected from existing research, and language-specific tools for linguistic annotation applied and potentially adapted.
The Action will benefit from the wide range of languages and expertise of the participants who embody a significant proportion of the relevant experts worldwide, large enough already to attract further participants and to make a significant impact on best practices in Distant Reading research.
In relation to the challenge of adaptation (described in section 1.3.3), **synergies and self-reinforcing network effects** can be expected to be particularly strong. Experience with the different types of adaptation and for various types of tools can be shared, so that it becomes easier each time another tool needs to be adapted to a new domain or language.
The partners themselves will be rewarded considerably by new opportunities to regularly share and discuss methods and insights with colleagues; by increasing the efficiency of their research and reducing redundant efforts; by being able to work with the ELTeC as a common reference; by taking advantage of training events for themselves and their teams; by providing opportunities for dialogue between established experts and ECIs; by providing expertise and partners for joint research funding proposals; and by creating the opportunity for comparative work beyond language borders exploring the interrelations of a common European literary history.
Finally, each partner's contribution will gain a **much higher visibility** through the network effect produced by the joint platform for the ELTeC, where resources (tools, methods, corpora, teaching materials) curated or created by the partners will be hosted or presented and widely disseminated.

1.4.2. In relation to existing efforts at European and/or international level

Currently, there is a **lively and diverse landscape of relevant projects and activities** at the national and European levels relevant to Distant Reading research. Among them are infrastructure initiatives

for research in the humanities such as **DARIAH and CLARIN**. This Action will liaise closely with their relevant Working Groups. The Action will collaborate with them especially with regard to common data and metadata standards and best practices as well for ensuring the long-term archiving and accessibility of resources (data and tools) produced by the Action. However, the Action's focus is much more specific, building a closely-knit, highly active network around the ELTeC and the research questions it enables.
In parallel, there is a **large number of more focused and more local initiatives or research centres** specifically active in the area of computational methods and digital resources for Distant Reading research, across Europe, many of which are among the proposers of the present Action. These groups, while doing excellent and recognized work in their domain or country, do not usually have the means for coordinated networking activities across Europe. Also, these centres, departments and groups usually have ECIs and would therefore be precisely the ones that would benefit most from the Action's networking and training activities.

2. IMPACT

2.1. Expected Impact

2.1.1. Short-term and long-term scientific, technological, and/or socioeconomic impacts

Impact in terms of the three main areas of activity:

- The **availability of the ELTeC** will have an immediate impact on Distant Reading research because it will enable, from its first iteration on, innovative cross-linguistic enquiries into the European literary tradition. Many of the dissemination activities (described below) aim at ensuring uptake of the ELTeC in research projects beyond the Action itself. A Training School will be devoted to the question of how to contribute further collections to the ELTeC. In the long term, the ELTeC will influence the way literary full-text collections will be prepared and will demonstrate requirements for digital full-text for research to digital libraries. The Action will rely on Zenodo.org for long-term availability and on DARIAH for sustained impact.
- The establishment of best practices in **Digital Reading methods** will function as an impetus for consolidation and greater visibility of new quantitative methods. Case-studies will be used to showcase the currently possible results and the future potential of Distant Reading methods. Such studies will be published in open access and high-impact journals as well as on the Action's portal. In the long term, this is expected to have systemic impact on how digital and quantitative methods of literary research will be done by showcasing and supporting common European or international standards.
- Work on the **theoretical implication of Distant Reading** research both for established literary theory and history and on the theoretical foundations of Distant Reading itself will have the immediate effect of guiding the research among the Action's participants. In the long term, because it connects computational methods and literary theory, this work will foster acceptance and help mainstream computational methods in literary studies.

Impact in terms of target groups and stakeholders:

- The Action will have immediate impact on **individual researchers and research groups** connected through the planned activities. Initially, 16 research groups from 13 different countries, with around 60 expert researchers and ECIs, will be involved directly in this Action and will benefit through sharing of expertise, tools, methods, and data. They primarily come from the fields of Computational Linguistics and Digital Literary Studies. More Action participants are welcome to join, especially during the first year.
- In terms of immediate impact on **ECIs specifically**, Training Schools will be open to all interested scholars, especially ECIs. A total of 9 events with an estimated 25 participants each, will result in over 200 investigators being trained. They will benefit by building-up specific competencies related to corpus building, Distant Reading methods and literary history and theory. They will learn to train or support other scholars at their institutions, ensuring both sustainability and multiplier effects ("train the trainers"). Through Training Schoolss as well as Short-Term Scientific Missions, they can take advantage of the Europe-wide networking opportunities created by the Action for the development of joint publications or projects. This aspect of human resource development will sustain the Action's impact well beyond its lifetime.
- **National or regional libraries** are another important stakeholder in this Action. The ELTeC will incorporate digital full-text from libraries whenever licenses and full-text quality permit it. More importantly, the Action will showcase what research can be accomplished with such

resources, contributing to the discussion on how text mining methods can enhance libraries in the future. Finally, the Action will demonstrate researchers' data and metadata requirements for materials from libraries and foster much-needed closer collaboration.

- **Another target audience of this Action are scholars in Computer Science**, a field that has a strong focus on developing tools and methods for English only and tends to work with relatively structured data or clearly defined concepts. For this reason, complex, multilingual data requiring a lot of contextual information (such as ELTeCs novels) and fuzzy concepts (as prevalent in literary studies) provide an interesting challenge for Computer Science both in terms of information retrieval and data modelling.

The general scientific impact of this Action will concern increasing innovation, internationalization, integration and inclusiveness of research into the cultural heritage in the European Research Area. Enhanced exchange between national research communities will catalyse methodological and conceptual developments in the field. Fostering the wider acceptance of standardization in terms of best practices and input formats, the Action will facilitate international research collaboration.

Beyond research in the Humanities, the Action will contribute to innovation in statistical text processing and interdisciplinary research, a Horizon2020 goal. Its methodological findings also have relevance for forensic linguistics and language technology. Finally, the Action will demonstrate how digitized cultural heritage helps better understand European culture and demonstrate the usefulness of research into literary texts to a wider public interested in the foundations of European identity.

2.2. Measures to Maximise Impact

2.2.1. Plan for involving the most relevant stakeholders

Many **established researchers** from relevant disciplines are already directly involved in the Action as participants, attracting further colleagues. Expert Meetings will be targeted at them and the more technical publications produced by the Working Groups will also help reach them. Established researchers and their research groups from additional countries will be invited to join the activities and be encouraged to contribute a further language subcollection to ELTeC.

A substantial proportion of **ECIs** will be involved from the beginning through their affiliation with existing participants, who all have teams or relevant projects with ECIs. Further ECIs will be reached through existing channels of communication available to the members of the Action, which cover a wide range of languages, countries, and relevant disciplines. Travel subsidies awarded to ECIs for Training Schools constitute a substantial incentive for ECIs to participate. Participating ECIs will be encouraged to support colleagues at their home institutions by using the openly available teaching materials from the Training Schools.

In order to involve **the library world**, the Action will actively liaise with relevant library initiatives, such as Europeana Research, with the goal of the libraries endorsing certain recommendations for full-text and metadata provision by libraries. Based on the Action's experience, WG 1 and 2 will produce a white paper outlining the requirements concerning the format, granularity and quality of full-text and metadata for research in both computational linguistics and digital literary studies.

Finally, the more **general public interested in European cultural identity** will be reached through several dissemination measures aimed at the larger public, such as interviews, video statements and participation in science fairs (see section 2.2.2).

2.2.2. Dissemination and/or Exploitation Plan

Exploitation plan:

- Based on work on and with the ELTeC, **research publications** will be published in appropriate venues, depending on subject and target audience (see list in WG work plans). Wherever possible, **Open Access** venues of publication will be favoured to increase visibility and uptake independently of individuals' or institutions' journal acquisition budgets. Some relevant journals are *Digital Scholarship in the Humanities*, *Digital Humanities Quarterly, Cultural Analytics, Journal of Quantitative Linguistics, Journal of Machine Learning Research*.
- The ELTeC will be the **foundation for separately-funded follow-up projects** either expanding the text collection or using it for specific research questions, either from Computational Linguistics or from Digital Literary Studies. Dedicated sessions at the Expert Meetings will serve to support Action participants in this.

Dissemination, beyond the publication of the freely available ELTeC (see 1.3.2) and the Training Schools (see 3.1), will be continuous and seek to cover a specific set of intended recipients:

The **Distant Reading community beyond the Action** will be targeted to ensure rapid uptake of the ELTeC in new or ongoing research projects:

- **The Action's portal** will be the entry point for scholars, where news about events and resources, especially the ELTeC, will be available. We will actively pursue a strategy to add information about the Action and the link to the portal to relevant, national and European, initiatives' websites in order to increase visibility. Examples include DARIAH, CLARIN, OpenAIRE and EADH (European Association for Digital Humanities).
- **News, calls and announcements** will be disseminated via a dedicated e-newsletter, on relevant mailing lists such as Corpora-List, Humanist-List, Linguist-List, on social media, (e.g. Twitter) and through the Action participant's professional networks.
- The Action will conduct **two larger joint conferences** in years 2 and 4. The target audience will be a broad range of digital and mainstream researchers from the humanities. The aim will be to present the achievements of the Action and create multiplier effects.
- The Action will produce **Final Action Dissemination materials** summing up the resources, results and achievements of the Action.

Scholars in the Humanities, especially ECIs, will be targeted in order to attract them to the Training Schools and inform them about the ELTeC.

- A **dedicated flyer** (printed and published online in open access, and updated at least once after year 2) will outline the Action's goals, focusing on the ELTeC and the Training Schools. A policy statement regarding female ECIs will encourage them to join Training Schools.
- The Action will aim to be present at relevant **major conferences** (Dissemination Meetings) such as the annual *Digital Humanities Conference* with presentations about the Action. It will encourage participants to present research they conduct in relation with the Action at national, European and international conferences and distribute the Action's flyer there.

The **wider audience of citizens interested in science and culture** will be targeted to show that the Action's results are interesting and relevant to European cultural history:

- Key publications and results will be presented in **short and lively video statements**, published in a dedicated Youtube Channel, presented on the Action's portal and publicised over the network's dissemination channels.
- Action participants will conduct **interviews with mainstream media** and **present their work at festivals of science** (e.g. 'Athens Science Festival' or 'Festiwal Nauki Krakow'). Media interest in authorship attribution has been significant in the past.

2.3. Potential for Innovation versus Risk Level

2.3.1. Potential for scientific, technological and/or socioeconomic innovation breakthroughs

As outlined above (section 1.1.2), the conditions for scientific breakthroughs in the domain of Distant Reading for European Literary History are extraordinarily great, because we are presently in a phase where several positive developments in relevant areas can be brought to converge. This means that the Action will be able to leverage the ELTeC (the first collection of some 2,500 complete novels in at least 10 languages) and existing Distant Reading methods and tools to support a breakthrough in scope, quality and acceptance for Distant Reading research for European literary history. If this happens, it will not only mean that Distant Reading research in continental Europe catches up with Anglo-American developments, but that it will create a truly multilingual, self-sustained field of enquiry adapted to the cultural and linguistic diversity in Europe. The fundamental risk with this research is mostly related to the fact that it disrupts or challenges many researchers' working habits and thinking styles, potentially leading to less than satisfactory acceptance of this type of research in the literary studies and humanities community. However, the Action members believe this is a risk worth taking at this point and intend to build close links between the Distant Reading and literary history communities. (Operational risks are discussed below, section 3.1.4.)

3. IMPLEMENTATION

3.1. Description of the Work Plan

3.1.1. Description of Working Groups

The Action will be coordinated by a Management Committee (MC). It is the decision-making body and coordinates and manages the Action in line with COST policies. In particular, the MC coordinates the activities, assesses proposals by WGs, plans, manages and allocates budgetary resources, oversees the overall progress of the Action and conducts quality assurance. It develops and oversees the Work and Budget Plans for each of the four Grant Periods and is responsible for submitting the Progress Reports as well as the Final Achievement Report. It will lead the joint organization, with a

local organizer, of the mid-way and final conferences. The MC meets once face-to-face every year and holds at least one additional video conference per year, with a focus once on starting off the new grant period, once on planning the next one. The work plan lists Deliverables (D) and Milestones (M); the eight coordination meetings have been omitted here.

ID	month	description
D-0-2	4	Monitoring Scheme is in place.
M-0-1	18	Progress Report 1.
D-0-3	22	Mid-term revision of Monitoring Scheme.
D-0-4	24	Joint mid-way conference
M-0-2	36	Progress Report 2.
D-0-5	45	Joint final conference.
M-0-3	48	Final Achievement Report.

The Action will further be constituted of 4 Working Groups (WGs) organising and conducting three distinct areas of activity as well as the cross-cutting task of dissemination. The WGs will meet face-to-face once every year and will hold at least two additional video conferences per year.

WG 1, "Scholarly Resources", will coordinate building and publishing the ELTeC. Its tasks will be to advise partners in structuring the subcollections, including linguistic annotation and metadata collection. Members of WG 1 will primarily come from participants active in computational and corpus linguistics. Expert Meetings will mainly concern discussion of the structure, content, data, annotation and metadata standards for the ELTeC.

ID	month	description
D-1-1	4	Expert Meeting (EM) to agree on a basic common framework (data and metadata) for the creation of the ELTeC (in support of D-1-2, Guidelines)
D-1-2	7	Guidelines (version 1) on a common framework (data and metadata formats) for the creation of the first iteration of subcollections of the ELTeC.
D-1-3	9	Training School (TS) on corpus building, including data and metadata standards and requirements for linguistic and literary research.
M-1-1	12	The ELTeC, first iteration: subcollections for 6 languages.
D-1-4	18	EM to agree on common framework for subcollection annotation (in support of D-1-5, Guidelines; also involving WG 2)
D-1-5	18	TS on building corpora generally and contributing to the ELTeC specifically, including using the linguistic annotation framework.
D-1-6	21	Guidelines (version 2) elaborating on the common framework for subcollection creation, including shared framework for linguistic annotation.
M-1-2	24	The ELTeC, 2nd iteration: subcollections in at least 4 additional languages.
M-1-3	32	The ELTeC, 3rd iteration: at least 6 expansion subcollections.
D-1-7	37	EM on data and metadata quality requirements for linguistic & literary research (with WG 2 and library experts; for D-1-9)
D-1-8	37	TS on cross-language linguistic annotation and strategies for compatibility.
D-1-9	42	White paper on full-text data and metadata requirements directed at resource providers such as digital libraries (with WG 2)

WG 2, "Methods and tools", will coordinate activities related to sharing, evaluating and improving methods and tools for Distant Reading research, with a focus (1) on tool and method adaptation and (2) on establishing best practices across Europe. Strengths and weaknesses of existing tools will be established and best practise guidelines for their application will be published. Members of WG 2 will come from participants active in computational linguistics, text mining, computational stylistics, and digital literary studies.

ID	month	description
D-2-1	7	EM to discuss and coordinate adaptation of tools and methods (in preparation of D-2-3).
D-2-2	7	TS on adaptation of tools and methods from one language to another.
D-2-3	11	White paper on best practices and recommendations for adaptation of tools and methods.
D-2-4	18	EM on comparing performance of Distance Reading measures across several languages using ELTeC (work towards M-2-1).
D-2-5	18	TS on using Distant Reading across several language.
M-2-1	22	Journal paper on benchmarking and language-dependent performance evaluation of existing methods using ELTeC.
D-2-6	30	EM to discuss the state-of-the-art and coordinate improvement of tools for multilingual Distant Reading (work towards M-2-2).
D-2-7	30	TS on tools and methods for multilingual Distant Reading.
M-2-2	34	Journal paper summarizing the results of evaluating existing Distant Reading tools and

		strategies for improvement.
M-2-3	46	“Lessons-learned”-report about cross-linguistic use of Distant Reading Methods.

WG 3, “Literary Theory and History”, will coordinate work related to applying Distant Reading methods to research questions from literary history and to clarify the theoretical and methodological bases and implications of Distant Reading and data-driven literary history. Its tasks will be (1) to assess and redefine fundamental concepts for literary history, such as style, genre or authorship in relation to Distant Reading research; (2) to explore the consequences for periodization and canonization in literary history that new results based on Distant Reading have; and (3) to support the theoretical foundations of Distant Reading research itself. Members of WG 3 will come from partners active in digital literary studies and mainstream literary history and theory.

ID	month	description
D-3-1	6	EM on concepts like genre, style and authorship in computational and non-computational literary history (work towards M-3-1)
D-3-2	6	TS on concepts like genre, style and authorship in computational and non-computational literary history
M-3-1	15	Journal paper outlining concepts like genre, style and authorship in computational and non-computational literary history.
D-3-3	21	EM on different approaches to literary periodization and canonization with respect to literary history and Distant Reading (for M-3-2)
D-3-4	21	TS on different approaches to literary periodization and canonization with respect to literary history and Distant Reading.
M-3-2	28	Journal paper outlining different approaches to literary periodization and canonization in relation to literary history and Distant Reading.
D-3-5	35	EM on theoretical assumptions and foundations of Distant Reading research (for M-3-3)
D-3-6	35	TS on theoretical assumptions and foundations of Distant Reading research.
M-3-3	42	Journal paper on theoretical assumptions and foundations of Distant Reading research.

WG 4, “Dissemination”, will (1) provide the Action with the research portal on which the ELTeC and all other resources of the Action will be available; the portal will serve as a virtual meeting point of the Action participants and of the larger community of researchers using Distant Reading methods and (2) coordinate the various dissemination activities of the Action, in close collaboration with the other WGs and the MC. This WG will therefore provide the other WGs with a technical infrastructure and the organisational support that helps them achieve their goals. WG 4 holds Coordination Meetings to plan its work, jointly with MC meetings whenever possible.

ID	month	description
D-4-1	3	Coord. Meeting year 1: portal structure & dissemination strategy (for M-4-1/M-4-2)
D-4-2	4	The project-internal wiki is functional.
M-4-1	4	Guidelines for a dissemination strategy, supporting the other WGs and the MC.
M-4-2	6	The Action’s portal is functional and online, ready to host the ELTeC.
D-4-3	9	The flyer (initial version) is available in print and online.
D-4-4	13	Coordination Meeting year 2: strategy for resource integration into the portal.
D-4-5	17	Portal presents additional resources (TS materials, profiles, etc.)
D-4-6	25	Coord. Meeting year 3: update dissemination strategy, update flyer.
D-4-7	28	The flyer (updated version) is available in print and online.
D-4-8	37	Coord. Meeting, year 4: preparing final impact report for MC; work towards M-4-3.
D-4-9	45	Portal has been converted to a static website documenting the Action.
M-4-3	46	The Final Action Dissemination materials are available.

Among the joint activities are Training Schools (TS), Expert Meetings (EM), and Short Term Scientific Missions (STSM).

The Action will conduct 9 Training Schools (TS), organized by the Working Groups (WGs) and targeting PhD candidates and ECIs from all disciplines involved. TSs serve to share and enhance methodological and practical competencies for using the ELTeC and associated Distant Reading methods. TSs last 3-4 days and, whenever possible, will be co-hosted with EMs to reduce travel, involve expert researchers in teaching and foster cross-generational exchange. Training materials used will be published in the Action’s portal and will be open for reference and re-use by others.

The Action will conduct 9 Expert Meetings (EM), organized by the WGs and targeting experienced ECIs and established researchers. EMs serve to coordinate work on the ELTeC, on evaluation and adaptation of methods and on theory development. EMs also serve to prepare joint journal publications or guidelines and to coordinate joint funding proposals through specific sessions. EMs last 2-3 days and, if not leading to a formal publication, will be documented through white papers.

Short Term Scientific Missions (STSM) will be offered (primarily) to ECIs. The aim of STSMs is to increase ECIs' (especially intra-European) mobility with the aim of helping them acquire new research competencies with regards to data curation, method adaptation and cross-language use, and fostering more intense exchange and collaboration between the participating institutions. The extended STSMs are expected to lead to visible outcomes such as joint (data, tool or journal) publications. Budget permitting, the Action aims to provide **approximately 300 days of STSM per grant period**. These will be divided flexibly between the participants, under the supervision of a STSM coordinator. Each year, 8 STSMs of very short duration (10 to 15 days) will be offered as well as 2 more extended STSMs (around 100 days, aimed at ECIs exclusively). This amounts to a total 1200 days of STSM over the course of the Action.

3.1.2. GANTT Diagram

Year 1 is considered the preparation phase, years 2 and 3 the operational phase and year 4 the consolidation phase.

	year 1 (preparation)												year 2 (operation)												year 3 (operation)												year 4 (consolidation)											
type month	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48
MC / Cross-WG																																															
MC Meetings	Mt								Mt				Mt								Mt				Mt								Mt				Mt								Mt			
Reports																		P																P												P	
Conferences																					C																								C			
Monitoring Scheme									Mi													Mi																											
STSMs																																																
WG 1																																															
Expert Meetings				EM														EM																				EM										
Training Schools									TS									TS																				TS										
ELTeC (3 iterations)												Mi												Mi								Mi																
Publications/reports							P														P																						P					
WG 2																																															
Expert Meetings						EM												EM											EM																		
Training Schools						TS												TS											TS																		
Publications/reports										P											P												P										P				
WG 3																																															
Expert Meetings						EM															EM												EM														
Training Schools						TS															TS												TS														
Publications															P													P														P						
WG 4																																															
Coordination Meetings			Mt										Mt												Mt												Mt										
Portal (3 milestones)						Mi												Mi																											Mi			
Flyer (2 versions)									P																			P																		
Dissem. Scheme (2 milest.)				Mi																																										P		

Grey areas indicate phases of specific preparatory or continuous activities; Black areas indicate months of completion or realization. Abbreviations used: EM=Expert Meeting, TS= Training School, P=Publication/Report, Mt=Meeting, Mi=(other) Milestone, C=Conference.

3.1.3. PERT Chart (optional)

3.1.4. Risk and Contingency Plans

The following issues could hinder the Action from reaching its goals. Frequent virtual meetings of Action Chair and WG leaders will ensure the Action can rapidly detect any issues and react quickly.

- Technical issues of harmonization when bringing together texts from various sources, in multiple languages and with multiple annotations, could pose a risk to the timely implementation of the ELTeC. We are confident that the extensive experience of the partners with building corpora (most of them annotated, some of them multilingual) will help us overcome such difficulties. If problems persist, the Action has fall-back options (core vs. extensions of ELTeC, simpler annotations, less strictly comparable corpora, less detailed metadata). The Action members do not expect any issues arising from copyright law, because all the texts to be included in the ELTeC are in the public domain.
- It may turn out to be difficult to find enough full-text novels for some of the lesser-resourced languages for those subcollections to have a structure and size exactly comparable to the others. However, work in WG 2 can proceed as soon as the 1st iteration of ELTeC is available. If the subcollection for a given language is delayed, of smaller size, or structured differently from the other collections, work in WG 2 would be somewhat limited in scope, but the fundamental feasibility would not be jeopardized.

3.2. Management structures and procedures

The Management Committee (MC) includes representatives of all participating member countries as nominated by the COST National Coordinators. The MC is headed by the **Action Chair and Vice-Chair** (elected at the first MC meeting) and supported by the Scientific Representative of the Grant Holder and by the Working Group (WG) leaders. An STSM Coordinator supporting the Action Chair will be appointed. Also, a Quality Assessment Coordinator will develop qualitative and quantitative criteria consensually with the WGs and support the MC in monitoring the quality and success of the Action's activities and deliverables. An Action-internal wiki will support the coordination of the Action as a shared space for meeting agendas, minutes, contact details, documents, and task management. Meetings will be held together with EMs, whenever possible, to help connect the MC and the WGs. Action Chair, Vice-Chair and WG leads will hold more frequent virtual meetings to ensure optimal communication. The mid-way conference, beyond its dissemination function, will also serve to bring all WGs together.

The **Working Groups** will each be led by a chair experienced in coordinating interdisciplinary, international research groups and co-led by a vice-chair, ideally an ECI for whom this position will be a valuable learning experience. The WGs develop proposals for activities and coordinate research conducted within the framework of the Action. Membership of individual participating scholars is not restricted to a single WG; rather, engagement in multiple WGs is encouraged to foster exchange between WGs. The WGs make proposals for activities and events to the MC, in accordance with the WG tasks and objectives. The WGs report to the MC at the end of each year. Whenever feasible, WG (as well as MC) meetings will be held in conjunction with other events organized by the network to reduce travel, increase cost effectiveness, and foster exchange.

3.3. Network as a whole

This Action brings together a critical mass of researchers in terms of both geographical distribution and spectrum of expertise necessary to achieve the Action's objectives. The **initial network of proposers** includes researchers from 16 institutions in 13 countries. 11 of them are COST Member Countries (Belgium, France, Germany, Greece, Ireland, Israel, Italy, the Netherlands, Norway, Poland and Spain), among them one COST Inclusiveness Target Country.

Also part of the initial network of proposers are three **vital partners from further countries**. These include established scholars from Israel (a COST Cooperating State), the United States and Australia (COST International Partner Countries) who have outstanding expertise in Distant Reading methodology. With issues of multilingualism, multiculturalism and multiple literary traditions of their own, these countries have insights to gain from and experience to contribute to this Action centred around an investigation into Europe's rich multilingual literary traditions.

The Action's network embodies computational linguists and digital literary scholars working on literature in 10 different languages (Dutch, English, French, German, Italian, Modern Greek, Polish, Portuguese, Spanish and Russian, at least 3 of of which are less-resourced languages), as well as computer scientists, most of whom have already worked with several of the relevant languages and with Distant Reading methods. The network incorporates many researchers that have extensive experience with corpus building and/or Distant Reading methodology and have published significant contributions on these issues during the past years.

With a significant number of ECIs as well as a relatively high proportion of female researchers among the initial proposers, the Action is in a good starting position to achieve its goals with respect to inclusiveness and involving higher-than-usual numbers of ECIs and female researchers.

The Action proposers are committed to the Action's objectives and have agreed to pool resources to enable the successful operation of the network, in particular by contributing texts and expertise to the creation of the ELTeC. They are already working together on an ad-hoc basis and wish to broaden, intensify and sustain this collaboration through the COST Action. Many of the Action's proposers have extensive experience in developing successful third-party funding proposals and are in a position to make in-kind co-funding decisions (e.g. staff time). Finally, many partners bring to the network considerable experience gained from participation and/or lead roles in diverse, interdisciplinary and international research consortia, on the level of research projects, networking efforts and infrastructure initiatives. Taken together, all of these factors will ensure making this COST Action a success.

References

Argamon, Shlomo. 2008. "Interpreting Burrows's Delta: Geometric and Probabilistic Foundations." *Literary and Linguistic Computing* 23 (2): 131–47. doi:10.1093/llc/fqn003.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3 (March): 993–1022.

Bouillier, Dominique, and Audrey Lohard. 2012. *Opinion mining et Sentiment analysis*. Marseille: OpenEdition. http://press.openedition.org/198.

Burrows, John. 2002. "'Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship." *Literary and Linguistic Computing* 17 (3): 267–87. doi:10.1093/llc/17.3.267.

Craig, Hugh, and Arthur F. Kinney, eds. 2009. *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge: Cambridge University Press.

Cheryan, Sapna, Victoria C. Plaut, Caitlin Handron, and Lauren Hudson. 2013. „The Stereotypical Computer Scientist: Gendered Media Representations as a Barrier to Inclusion for Women". In: *Sex Roles* 69 (2013): 58–71. doi: 10.1007/s11199-013-0296-x

Stefan Evert, Thomas Proisl, Fotis Jannidis, Steffen Pielström, Christof Schöch and Thorsten Vitt. 2015. „Towards a better understanding of Burrows's Delta in literary authorship attribution", in: *North American Chapter of the Association for Computational Linguistics – Human Language Technologies* (NAACL HLT 2015), Denver, Colorado, USA.

Herrmann, J. Berenike, Christof Schöch, and Karina van Dalen-Oskam. 2015. "Revisiting Style, a Key Concept in Literary Studies." *Journal of Literary Theory* 9 (1): 25–52.

Jockers, Matthew L. 2013. *Macroanalysis: Digital Methods and Literary History*. Topics in the Digital Humanities. University of Illinois Press.

McCarty, Willard. 2014. "Getting There from Here. Remembering the Future of Digital Humanities Roberto Busa Award Lecture 2013." *Literary and Linguistic Computing* 29 (3): 283–306.

Moretti, Franco. 1999. *Atlas of the European Novel, 1800-1900*. Verso.

Moretti, Franco. 2005. *Graphs, Maps, Trees: Abstract Models for a Literary History*. Verso.

Newman, M. E. J. 2010. *Networks: An Introduction*. Oxford; New York: Oxford University Press.

Murphy, Amanda. 2012. 'Corpus Analysis of European Union Documents', in *The Encyclopedia of Applied Linguistics*. London: Blackwell.

Rybicki, Jan, and Maciej Eder. 2011. "Deeper Delta across Genres and Languages: Do We Really Need the Most Frequent Words?" *Literary and Linguistic Computing* 26 (3): 315–21. doi:10.1093/llc/fqr031.

Smith, Peter W. H., and W. Aldridge. 2011. "Improving Authorship Attribution: Optimizing Burrows' Delta Method." *Journal of Quantitative Linguistics* 18 (1): 63–88.

Søgaard, Anders. 2013. *Semi-Supervised Learning and Domain Adaptation in Natural Language Processing*. Morgan & Claypool.

Wilkens, Matthew. 2015. „Digital Humanities and Its Application in the Study of Literature and Culture", in: Comparative Literature, 67/1, 11-20.