Horizon Europe Programme

Standard Application Form (HE EIC PATHFINDER CHALLENGES)

Project proposal – Technical description (Part B)

Version 1.0
04 December 2022

HISTORY OF CHANGES		
Version	**Publication date**	**Changes**
1.0	04.12.2022	▪ Initial version
		▪

Proposal template Part B: technical description

1. Excellence

1.1 Relevance to the Challenge

As the text simplification aims to reduce the linguistic complexity of the text and retain the original meaning. Human being normally simplify by performing several text transformations, such as replacing complex terms by simpler synonyms, reordering words or phrases, removing non-essential information, and splitting long sentences, we want to achieve this kind of simplification in such a way that it is more innovatively, solving social problems or adding benefits such as avoiding the excess of physical documents which normally take up and causes loss of information, takes less time than manual process and due to the partnership with institutions new opportunities will be opened, and with that, some professionals will be employed.

We need to provide best experience of Hungarian text simplification on national and international institutions that deal directly with those with cognitive disabilities such as aphasia, dyslexia, and autism, or with non-native speakers and children with reading difficulties. After completing our project, we want to remove the Hungarian from the list of the 5 most difficult language in the world and surpass 13 million speakers.

1.2 Novelty

GRETEL is progressing beyond the state of the art by the following innovations:

Innovation: Gathering and organizing textual corpora in Finnish and Hungarian languages

Creating textual corpora for certain languages is an essential part of creating complex text simplification systems. Such systems for languages other than English are hard to produce exactly as a result of the need for training and testing data, as datasets for one language cannot be used for another. There is a reasonably large number of datasets for the English language as lots of data can be collected from the internet. such as WikiLarge (Zhang et al., 2017)[1], a dataset consisting of sentences collected from Wikipedia and the simplified Wikipedia.

When creating these corpora numerous aspects must be considered. The sentences and their simplified sentence pairs have to be chosen carefully, to carry as much information as a pair as they possibly can. The different grammatical rules also have to be taken into account, which is why the project requires the participation of institutes of linguistics for both languages. The methods of creating such a corpus for the French language can be found in (Gala et al., LREC 2020)[2].

GRETEL will go beyond the state of the art in creating new large textual corpora for Hungarian and Finnish languages, including sentences selected with high attention to the value they can bring for a text simplification system. Also, the simplified form of these sentences will be included in these new datasets, as these sentences will be the target labels for our data.

Innovation: Improved text simplification systems using transformer networks

Numerous ATS (Automatic Text Simplification)-related research has been published over the past 20 years, as reviewed by Saggion (2017)[3], and Al-Thanyyan and Azmi (2021)[4], just to mention some of them. In short, the field has mostly concentrated on creating techniques for automatically simplifying difficult words (lexical simplification)

[1] ZHANG, Xingxing; LAPATA, Mirella. Sentence simplification with deep reinforcement learning. *arXiv preprint arXiv:1703.10931*, 2017.

[2] Núria Gala, Anaïs Tack, Ludivine Javourey-Drevet, Thomas François, and Johannes C. Ziegler. 2020.

[3] SAGGION, Horacio. Automatic text simplification. Synthesis Lectures on Human Language Technologies, 2017, 10.1: 1-137.

[4] AL-THANYYAN, Suha S.; AZMI, Aqil M. Automated text simplification: a survey. ACM Computing Surveys (CSUR), 2021, 54.2: 1-36.

and/or difficult syntactic structures (syntactic simplification). Tong Wang (2016)[5] described a study of the LSTM based model for text simplification. This study shows several operational rules such as sorting, reversing, replacing sentence pairs, meanwhile in the same year a rule-based text simplification model for the German language was proposed by Julia Suter at the University of Zurich[6]. They use experiments to explain how RNN and LSTM operate. Their model can perform distinct sorting, reversing, and replacement processes. It requires to combine all three procedures to simplify. Later then there a new language representation model was introduced called BERT, which stands for Bidirectional Encoder Representations from Transformers. (Kristina Toutanova, 2019)[7] .

Unlike recent language representation models (Peters et al., 2018a; Radford et al., 2018)[8] [9], BERT is designed to pre-train deep bidirectional representations from unlabelled text by jointly conditioning on both left and right context in all layers.

It has begun to be questioned how historically simplifications are assessed using automated criteria like BLEU[10] or the Flesch-Kincaid Reading Grade Level[11] (Sulem et al., 2018[12]; Tanprasert and Kauchak, 2021[13]; Alva-Manchego et al., 2021[14]).

The goals of ATS research have also been questioned. Stajner (2021)[15] highlights how prospective target groups characteristics have not been considered and urges the creation of more modular ATS systems that can be tailored for certain populations. Some of the most recent approaches, like Maddela et al. (2021)[16] or Sheang and Saggion (2021)[17] for English.

Some of the above-mentioned publications have been used as resources to help us better grasp text-simplification techniques. GRETEL will go beyond the state of the art in developing new criteria using the expert knowledge from linguistic centres and putting high effort into the research of transformer networks for text simplification

Innovation: Accurate text simplification systems for Finnish and Hungarian languages

While some work has also been developed for languages like Spanish, Portuguese, Basque, French or even Japanese, Hungarian has been hardly researched. One of these few Hungarian-related research were Dávid Márk Nemeskey's, published in 2020[18]. It contains huBERT: a variation of BERT models especially for the Hungarian language, using the WebCorpus 2.0, which is worth to mention, because this is the largest Hungarian NLP dataset with over 9 billion words in it.

With the expert knowledge from linguistic centres, we will develop text simplification systems which use language specific rules for simplifying textual corpora. This research can also help in identifying similarities between Hungarian and Finnish languages.

[5] WANG, Tong, et al. An experimental study of LSTM encoder-decoder model for text simplification. arXiv preprint arXiv:1609.03663, 2016.

[6] SUTER, Julia; EBLING, Sarah; VOLK, Martin. Rule-based automatic text simplification for German. 2016.

[7] DEVLIN, Jacob, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[8] PETERS, Matthew E., et al. Deep contextualized word representations. CoRR abs/1802.05365 (2018). arXiv preprint arXiv:1802.05365, 1802.

[9] RADFORD, Alec, et al. Improving language understanding by generative pre-training. 2018.

[10] PAPINENI, Kishore, et al. Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002. p. 311-318.

[11] KINCAID, J. Peter, et al. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Naval Technical Training Command Millington TN Research Branch, 1975.

[12] SULEM, Elior; ABEND, Omri; RAPPOPORT, Ari. BLEU is not suitable for the evaluation of text simplification. arXiv preprint arXiv:1810.05995, 2018.

[13] TANPRASERT, Teerapaun; KAUCHAK, David. Flesch-Kincaid is Not a Text Simplification Evaluation Metric. In: Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021). 2021. p. 1-14.

[14] ALVA-MANCHEGO, Fernando; SCARTON, Carolina; SPECIA, Lucia. The (un) suitability of automatic evaluation metrics for text simplification. Computational Linguistics, 2021, 47.4: 861-889.

[15] STAJNER, Sanja, et al. Lexical simplification benchmarks for English, Portuguese, and Spanish. arXiv preprint arXiv:2209.05301, 2022.

[16] GEHRMANN, Sebastian, et al. The gem benchmark: Natural language generation, its evaluation and metrics. arXiv preprint arXiv:2102.01672, 2021.

[17] SHEANG, Kim Cheng; SAGGION, Horacio. Controllable Sentence Simplification with a Unified Text-to-Text Transfer Transformer. In: Proceedings of the 14th International Conference on Natural Language Generation. 2021. p. 341-352.

[18] NEMESKEY, Dávid Márk. Natural language processing methods for language modeling. 2020.

1.3 Plausibility of methodology

Our methodology consists of 2 main parts. The first part (which includes all the work done in the first 18 months) includes the research done on text simplification and transformer networks, as well as the creation of the datasets needed. The second half of the project will focus on the development of specific simplification systems for the chosen languages.

Our methods of using artificial intelligence will come with high levels of robustness. The text simplification models will be developed to produce simplified text with such accuracy that corresponds and even exceeds the state of the art.

The models will be thoroughly tested to avoid any discrimination towards any groups of people. Such a case when a sentence is simplified in such a manner that is unacceptable should be avoided.

The consortium is assembled in such a way that enables teams to assist each other to achieve the best results they can. The expert knowledge from DeepL and the 3 involved universities combined will enable all teams to develop the best systems we can produce. This will be achieved by sharing any knowledge the other partners may need to progress forward. The creation of corpora will be done in the same manner. The linguistic centers involved will help each other in similar ways

The gender dimension is considered in the project's research in the same way any other discrimination is planned to be avoided. We plan to test our systems to rule out any such discrimination.

All our results and achievements (new algorithms, scientific methods and datasets) will be shared publicly on Github and the website dedicated to GRETEL.

During the lifetime of the project, we plan to use version control and project management software to improve communication and sharing of data between members. ELTE will be responsible for secure data management and for providing a database where all data can be stored and searched easily.

2. Impact

GRETEL will produce impacts in the field of natural language processing sector as well as on artificial intelligence in general. Research centres, Universities and Linguistic Institutes are a good representation of the potential beneficiaries of the results generated by the project: new technology will be developed and demonstrated, existing technology will be applied in innovative ways, new corpus will be created for further research.

2.1. Potential impact:

Ensure equality of children with reading difficulties.

Equality is one of the basic concepts of the European Union. As far as equality considered amongst children every one of them must have the possibility to read and enjoy anything they want. To support this ideology text simplification methods, help children who have ADHD, learning disabilities, reading difficulties. Reading disorders occur when a person has trouble reading words or understanding what they read. Text simplification helps in a way that it makes complex sentences in books and exercises much more understandable for disabled children by simplifying their structures and replacing rare and difficult words with common synonyms. Using text simplification, we can destroy these borders and they can feel equal moreover they can forget that they have some disorder.

Offer an easily usable tool for language learners.

The potential boost of text simplification amongst language learners is quite prominent. Language learners often encounter the difficulty to find easily understandable content in target language to deepen their knowledge. Since European Union consist of 27 different countries there are several languages spoken in it. European citizens have the possibility to easily move to another country

which in most cases leads to learning a new language besides English. Unfortunately, you cannot always get by using only English especially when you move to another country for a living. Text simplification makes it easier to integrate into a new culture by making language learning easier. In most cases language learners cannot understand different language phrases when they move to a new country. Replacing rare words and reducing sentence complexities improve the communication between language learners and native speakers. In our case we will make Finnish and Hungarian language learning easier and offer a tool to improve one's language skills.

Better understanding of law text and medical papers

For this kind of an impact, we can also mention equality. People who are not really educated have difficulties understanding complex and long laws and medical papers which can contain Latin words. By using text simplification, we can highlight the main meaning of a law as well as replace Latin words with their synonyms. If laws could be transformed into everyday language complexity people would much more likely read them and they would be more aware of the consequences of committing crimes. As far as medical paper simplification is concerned people would feel safer and closer to doctors since Latin words can alienate doctors and confuse patients. This would improve and their mental health and encourage them to regularly visit hospitals for medical screening.

Downside of text simplification usage

Thanks to text simplification methods complex texts can be transformed into simpler forms. This can encourage people read more and be familiar with more stories, news, information. But this benefaction does not come free if text simplification is overused there is a serious risk that people will be less intelligent since they do not have to think about and understand complex texts. People would be lazier and would have a demolished mindset where everything is simple, and the humanity would slowly be in a regression. To evade this dystopic scenario, we must ensure that text simplification usage is restricted and make sure to use it only for people in need of it decrease the inequity between people. This is also an ethic goal to save general intelligence and preserve excellent quality communications and beautiful complex sentence shaping.

GRETEL's potential impact towards future research

GRETEL has great potential to support future research. The establishment of the well preprepared Hungarian and Finnish corpus is a great advance both in Hungarian and Finnish natural language processing. These can be a well-established benchmark for text simplification for these languages. As far as the developed models are concerned, they can be also a good comparison benchmark for future models. Furthermore, the experiment with machine translation could be a good starting point for further developments.

2.2. Innovation potential:

Text simplification's innovation potential in language translation

Nowadays language translation is rather advanced and reliable for simple text. Text simplification has the innovation potential to boost language translation of complex texts. With text simplification we are decreasing the complexity while preserving the meaning of a text. Language translation can benefit from this since new databases can be generated where the translation dictionary would consist of the complex text, simple text and the translated text based on the simple form. Thus, text simplification improves language translation models since it can be applied before translation in order to translate easier texts. It would improve the readability and the fluency of the translation. In this way complex texts could be translated not only to simpler form but into new language with better understandability more precision about the meaning.

2.3. Communication and Dissemination:

New market in book selling

The stakeholders who have the biggest potential to benefit from this model are language institutes, libraries, book publishers. Concerning monetary benefactions book publishers could be the biggest winners since they could open a new market for language learners and children with reading difficulties or even to new target audience who want to read more but less in a sense (more story, less text) by publishing simplified version of already existing books at a bigger rate. To say the truth this approach of simplifying bestseller books is already known but with automated text simplification the process can be speeded significantly.

Language translation with better fluency

Companies who develop language translation software like DeepL could use this to improve their translator's fluency. First transform original language complex text to simplified form then translate to target language. To improve it more inverse transformation can be learnt too with machine translation by learning translation from simple text to complex text. In this approach not only the meaning but the complexity can be also saved in the target language translation.

3. Quality and efficiency of the implementation

3.1 Consortium

Given the objectives and the path we have set out to achieve them, the topic of this proposal can rightly be described as interdisciplinary. The two main areas concerned are computer science, more precisely NLP and deep learning, and the other one is textual linguistics. These two areas are very different, and their effective and productive joint cultivation requires different approaches, knowledge and expertise from the participants in the programme.

University of Strasbourg

Their previous research is the most relevant to this project, based on the ideas and findings set out in HECTOR. As we plan to create corpus for several languages that can help us to solve the text simplification problem in a given language, it is important to start by comparing the corpus construction methods described in the HECTOR research.

The professional excellence of the colleagues is also praised by the text simplification architecture they have created. This gives convincing results based on their research and can be considered as a good basis for a more general text simplification model, but it will probably need to be modified.

The primary contribution of the people delegated by the university is to support their colleagues in simplifying the creation of corpora, and they also have an important added value in the creation of the deep learning model. They will lead work package 1 and will also have a key role in the implementation of work packages 2, 3, 4 and 5.

Hungarian Research Centre of linguistics

The Hungarian Research Centre of linguistics aims to cover all fields of Hungarian linguistics and to carry out general, experimental and applied linguistic research. Their many years of experience in general linguistics are an important contribution to the overall work package, but the real significance of their membership of the Consortium lies in their potential to contribute to elements specific to the Hungarian language. The Institute of Linguistics and Applied Linguistics has as one of its main tasks the research and development of language technology, the construction and maintenance of language resources and language technology infrastructures, the design and compilation of electronic text databases for corpus linguistics research. By coordinating the latter activities with the objectives of this research, we can very effectively produce corpora of sufficient quality and quantity.

University of Helsinki

As can be seen from the description of the packages, work packages 4 and 5 deal with the construction of corpora of the Finnish language and text simplification. models that make use of them. Finnish has been chosen also because recent research by the Hungarian Research Centre of linguistics suggests a close language family relationship between Finnish and Hungarian. The state-of-the-art NLP model published by the University of Helsinki research team is useful in itself for meeting the commitments in Packages 4 and 5, and also allows the development of new and more effective models by exploiting similarities between languages.

ELTE

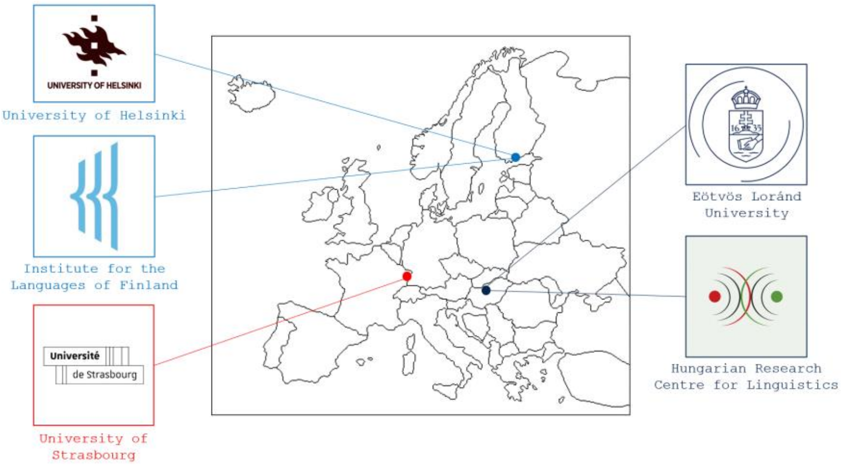
Despite the fact that the colleagues coming from here do not have a long research track record, their commitment to the topic, their youthful drive and their previously published research on Automatic Text Simplification of Hungarian Texts can make them very proactive members of the team. Their primary task will be to create the deep learning model and another important role will be to ensure smooth communication between team members from different fields.

The chosen research has potential to bring new benefits both to the scientific community and to a wide section of society. High-precision text simplification makes more complex and complex scientific papers more accessible to a wider audience, allowing for a simpler but loss-free transfer of information between the author and the reader. It can be applied to any written scientific article or research, thus spanning the spectrum of R&D and the humanities.

In addition to simplifying scientific writing, it is an excellent solution for EU nationals living in target countries who, due to a difficulty or disability, would not be able to process/interpret texts of greater complexity.

The listed consortium members have all previously worked on research with a subject matter and infrastructure requirement close to those undertaken in this project. It can therefore be said that all the applicant institutions have the basic infrastructure necessary to carry out the activities of the project. As the interdisciplinary meeting of these research areas is not particularly resource-intensive, the existence of the basic conditions may be sufficient in themselves. However, some non-essentials (e.g. powerful video cards/cloud gpu rental for training, large amounts of storage space for the corpora, etc.) may be necessary to ensure that the work can be carried out smoothly and successfully from a technical point of view as well.

Figure 3.1: Map of Europe with distribution of consortium members



3.2 Work plan and resources

Participants:

1. University of Strasbourg
2. Hungarian Research Centre for Linguistics
3. Eötvös Loránd University
4. Institute for the Languages of Finland
5. University of Helsinki
6. DeepL GmBH

Table 3.2a: List of work packages

Work package No	Work Package Title	Lead Participant No	Lead Participant Short Name	Person-Months	Start Month	End month
1	General text simplification model	1	UoS	42	7	18
2	Hungarian corpora creation	2	HRCL	27	1	18
3	Hungarian text simplification model	3	ELTE	45	19	32
4	Finnish corpora creation	4	ILF	27	1	18
5	Finnish text simplification model	5	UoH	45	19	32
6	Exploration of machine translation approach with transformers	1	UoS	36	1	18
				222		

Figure 3.2: Gantt chart of GRETEL

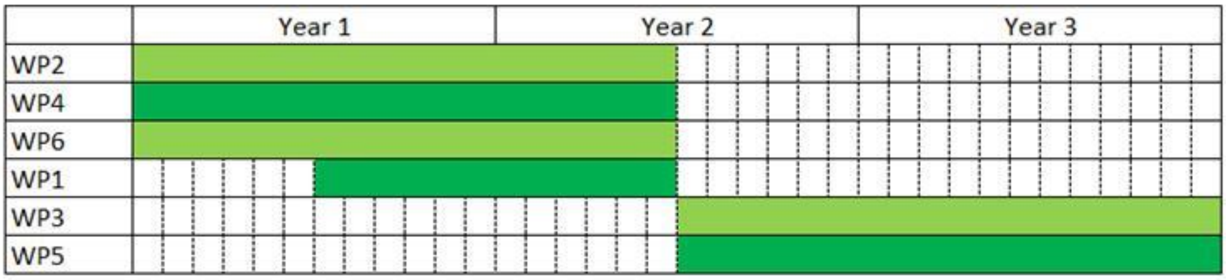


Table 3.2b: Work package description

For each work package:

Work package number	1	**Lead beneficiary**			1		
Work package title	General text simplification model						
Participant number	1	3	5	6.			
Short name of participant	UoS	ELTE	UoH	DeepL			
Person months per participant:	6	24	12	6			
Start month	7			**End month**	18		

Objectives

The package aims to create a general text simplification model to facilitate the creation of specific models later on. This model can be used later for similar research purposes.

Description of work (where appropriate, broken down into tasks), lead partner and role of participants

UoS will come up with the model structure, also syntactic and lexical simplification for English language.
ELTE's goal here is to develop the model using Hybrid approach(syntactic, lexical) for English.
UoH'll test the model exhaustively with BLEU, FKGL in English. They should highlight the errors.
DeepL's desired goal is to optimize their translator. They'll do experiments to see what kind of simplified text are better to optimize the language detection and translation with.

Deliverables (brief description and month of delivery)

The main deliverable for this WP is a general model that can be easily built upon when creating a specific text simplification model later. For ease of use, it is important to produce a well-documented work.

Work package number	2	**Lead beneficiary**			2		
Work package title	Hungarian corpora creation						
Participant number	2	3					
Short name of participant	HRCL	ELTE					
Person months per participant:	18	9					
Start month	1			**End month**	18		

Objectives

The main difficulty in creating Hungarian text simplification models is the lack of corpora of sufficient size and quality. Its preparation is critical for the success of subsequent WPs

Description of work (where appropriate, broken down into tasks), lead partner and role of participants

HRCL - Collect complex-simple pairs, volunteer work, native - non-native speakers involved, approve linguistic correctness
ELTE - Make it a database, create sql, csv format, publish it, do some exploration of data (vocabulary from it, correlations, frequency dictionary)

Deliverables (brief description and month of delivery)

A high quality and large database should be created for the Hungarian corpora, which can later be used to create a suitable text simplification model.

Work package number	3	**Lead beneficiary**		3			
Work package title	Hungarian text simplification model						
Participant number	3	1	5	6			
Short name of participant	ELTE	UoS	UoH	DeepL			
Person months per participant:	27	9	9	6			
Start month	19			**End month**	32		

Objectives
Based on the general model created previously, the goal is to create a reliable deep learning-based text simplification model with high accuracy.

Description of work (where appropriate, broken down into tasks), lead partner and role of participants

UoS should the support development process, andmake suggestions about the theory
ELTE will develop the model model and test it
UoH will provide help with the data pipeline
DeepL's desired goal is to optimize their translator. They'll do experiments to see what kind of simplified text are better to optimize the language detection and translation with.

Deliverables (brief description and month of delivery)

The most important deliverable here is the Hungarian language-specific text simplification model, which is able to simplify complex texts with the least loss of information.

Work package number	4		**Lead beneficiary**			4	
Work package title	Finnish corpora creation						
Participant number	4	5					
Short name of participant	ILF	UoH					
Person months per participant:	18	9					
Start month	1			**End month**	18		

Objectives
The main difficulty in creating Finnish text simplification models is the lack of corpora of sufficient size and quality. Its preparation is critical for the success of subsequent WPs

Description of work (where appropriate, broken down into tasks), lead partner and role of participants

The main contribution of ILF will be to collect complex-simple pairs for the dataset. This is the most crucial part of this package. They’ll involve both natve and non native speakers to have a wider range of language skill when determining the suffucient level of simplification. They will watch out to the linguistic correctness.
The technikal part will be done by UoH. Such as making a database out of the data, creating the sql distribution, reordering the data to csv format. Also, the publishing is their task.

Deliverables (brief description and month of delivery)

A high quality and large database should be created for the Finnish corpora, which can later be used to create a suitable text simplification model

Work package number	5	**Lead beneficiary**		5			
Work package title	Finnish text simplification model						
Participant number	5	1	3	6			
Short name of participant	UoH	UoS	ELTE	DeepL			
Person months per participant:	27	9	9	6			
Start month	19			**End month**	32		

Objectives

Finnish text simplification model should be created in this WP

Description of work (where appropriate, broken down into tasks), lead partner and role of participants

UoS will fulfill supporting tasks, and will make suggestions about the theoretical part.
UoH will develop and test the model.
ELTE’s goal will be to build a data pipeline.
DeepL’s desired goal is to optimize their translator. They’ll do experiments to see what kind of simplified text are better to optimize the language detection and translation with.

Deliverables (brief description and month of delivery)

The most important deliverable here is the Finnish language-specific text simplification model, which is able to simplify complex texts with the least loss of information.

Work package number	6	**Lead beneficiary**		1			
Work package title	Exploration of machine translation approach with transformers						
Participant number	1	3	5	6			
Short name of participant	UoS	ELTE	UoH	DeepL			
Person months per participant:	18	9	9	6			
Start month	1			**End month**	18		

Objectives

Doing smaller experiment project with the state-of-the-art machine translation approach in this WP besides the hybrid model that we are using for the main mehtod.

Description of work (where appropriate, broken down into tasks), lead partner and role of participants

Since UoS has a lot of experience in this field, they will come up with the transformer structure and support the development with the theoretical part of it.
ELTE'll try machine translation with hungarian databasees, and they will evaluate the results
UoH'll try machine translation with finnish databasees, and they will evaluate the results
DeepL's desired goal is to optimize their translator. They'll do experiments to see what kind of simplified text are better to optimize the language detection and translation with.

Deliverables (brief description and month of delivery)

The deliverable is to publish the results of the comparison and merging of the hybrid model we have used so far and machine translation.

Table 3.2c: List of Deliverables

Only include deliverables that you consider essential for effective project monitoring.

Deliverable (number)	Deliverable name	Work package number	Short name of lead participant	Type	Dissemination level	Delivery date (in months)
1	General Model	1	UoS	OTHER	PU	18
2	Hungarian Corpora	2	HRCL	DATA	PU	18
3	Hungarian Model	3	ELTE	OTHER	PU	32
4	Finnish Corpora	4	ILF	DATA	PU	18
5	Finnish Model	5	UoH	OTHER	PU	32
6	Machine translation + hybrid model	6	UoS	R	PU	18

KEY
Deliverable numbers in order of delivery dates. Please use the numbering convention <WP number>.<number of deliverable within that WP>.
For example, deliverable 4.2 would be the second deliverable from work package 4.

Type:
Use one of the following codes:

R: Document, report (excluding the periodic and final reports)
DEM: Demonstrator, pilot, prototype, plan designs
DEC: Websites, patents filing, press & media actions, videos, etc.
DATA: Data sets, microdata, etc.
DMP: Data management plan
ETHICS: Deliverables related to ethics issues.
SECURITY: Deliverables related to security issues
OTHER: Software, technical diagram, algorithms, models, etc.

Dissemination level:
Use one of the following codes:

PU – Public, fully open, e.g. web (Deliverables flagged as public will be automatically published in CORDIS project's page)
SEN – Sensitive, limited under the conditions of the Grant Agreement
Classified R-UE/EU-R – EU RESTRICTED under the Commission Decision No2015/444
Classified C-UE/EU-C – EU CONFIDENTIAL under the Commission Decision No2015/444
Classified S-UE/EU-S – EU SECRET under the Commission Decision No2015/444

Delivery date
Measured in months from the project start date (month 1)

Table 3.2d: List of milestones

Milestone number	Milestone name	Related work package(s)	Due date (in month)	Means of verification
1	A general text simplification model with the required amount of precision is finished	1	18	The model achieves the appropriate quality measures.
2	A sufficient corpus for the Hungarian text simplification model has been collected	2	18	The corpus reaches a level where it can be used for training and evaluating models
3	A simplification model for Hungarian language is created that corresponds to the given quality measurements.	3	32	The model achieves the appropriate quality measures.
4	A sufficient corpus for the Finnish text simplification model has been collected	4	18	The corpus reaches a level where it can be used for teaching and evaluating models
5	A simplification model for Finnish language is created that corresponds to the given quality measurements.	5	32	The model achieves the appropriate quality measures.
6	The machine translation approach with transformers has been investigated in combination with the hybrid model used so far, and substantial conclusions can be drawn for further research	6	18	A conclusion can be drawn from the experiment, which has not been available to the scientific community until now, regarding the relationship between the two approaches

KEY
Due date
Measured in months from the project start date (month 1)

Means of verification
Show how you will confirm that the milestone has been attained. Refer to indicators if appropriate. For example: a laboratory prototype that is ‘up and running’; software released and validated by a user group; field survey complete and data quality validated.

Table 3.2e: Critical risks for implementation

Description of risk (indicate level of (i) likelihood, and (ii) severity: Low/Medium/High)	**Work package(s) involved**	**Proposed risk-mitigation measures**
The creation of the corpora takes too much time, and there won't be enough time for the models to becmone completed (0.2, Low)	2,3,4,5	Strict adherence to the timetable
The general model cannot be developed in time, so the whole project will be delayed (0.1, Low)	2,3,4,5,6	Strict adherence to the timetable

Definition critical risk:
A critical risk is a plausible event or issue that could have a high adverse impact on the ability of the project to achieve its objectives.

Level of likelihood to occur: Low/medium/high
The likelihood is the estimated probability that the risk will materialise even after taking account of the mitigating measures put in place.

Level of severity: Low/medium/high
The relative seriousness of the risk and the significance of its effect.

Table 3.2f: Summary of staff effort

Please indicate the number of person/months over the whole duration of the planned work, for each work package, for each participant. Identify the work-package leader for each WP by showing the relevant person-month figure in bold.

	WP1	WP2	WP3	WP4	WP5	WP6	Total Person-Months per Participant
UoS	**6**	0	9	0	9	**18**	**42**
HRCL	0	**18**	0	0	0	0	**18**
ELTE	24	9	**27**	0	9	9	**78**
ILF	0	0	0	**18**	0	0	**18**
UoH	12	0	9	9	**27**	9	**66**
Total Person Months	**42**	**27**	**45**	**27**	**45**	**36**	***222***

Table 3.2g: 'Subcontracting costs' items

For each participant describe and justify the tasks to be subcontracted (please note that core tasks of the project should not be sub-contracted).

University of Strassbourg		
	Cost (€)	**Description of tasks and justification**
Subcontracting	25.000	Contributing to the development of general and language-specific deep learning models

Hungarian Research Centre for Linguistics		
	Cost (€)	**Description of tasks and justification**
Subcontracting	15.000	Contribute to the collection and creation of language-specific corpora and help with linguistic questions that arise along the way.

Eötvös Loránd University		
	Cost (€)	**Description of tasks and justification**
Subcontracting	25.000	Contributing to the development of general and language-specific deep learning models

Institute for the Languages of Finland		
	Cost (€)	**Description of tasks and justification**
Subcontracting	15.000	Contribute to the collection and creation of language-specific corpora and help with linguistic questions that arise along the way.

University of Helsinki		
	Cost (€)	**Description of tasks and justification**
Subcontracting	20.000	Contributing to the development of general and language-specific deep learning models

Table 3.2h: ‘Purchase costs’ items (travel and subsistence, equipment and other goods, works and services)

Please complete the table below for each participant if the purchase costs (i.e. the sum of the costs for ’travel and subsistence’, ‘equipment’, and ‘other goods, works and services’) exceeds 15% of the personnel costs for that participant (according to the budget table in proposal part A). The record must list cost items in order of costs and starting with the largest cost item, up to the level that the remaining, costs are below 15% of personnel costs.

University of Strassbourg		
	Cost (€)	**Justification**
Travel and subsistence	4.000	Travel to distant countries requires the amount to be
Equipment	5.000	The reason why such a large amount was spent on the purchase of blue equipment is that the price of computing power has risen recently
Other goods, works and services	13.001	Salaries to support researchers and contributors
Remaining purchase costs (<15% of pers. Costs)	2.999	
Total	25.000	

Hungarian Research Centre for Linguistics		
	Cost (€)	**Justification**
Travel and subsistence	3.000	Travel to distant countries requires the amount to be
Equipment	1.000	The cost required here is for the operation of online systems
Other goods, works and services	11.000	Salaries to support researchers and contributors
Remaining purchase costs (<15% of pers. Costs)	0	
Total	15.000	

Eötvös Loránd University		
	Cost (€)	**Justification**
Travel and subsistence	4.000	Travel to distant countries requires the amount to be
Equipment	5.000	The reason why such a large amount was spent on the purchase of blue equipment is that the price of computing power has risen recently
Other goods, works and services	13.001	Salaries to support researchers and contributors
Remaining purchase costs (<15% of pers. Costs)	2999	
Total	25.000	

Institute for the Languages of Finland		
	Cost (€)	**Justification**
Travel and subsistence	3.000	Travel to distant countries requires the amount to be
Equipment	1.000	The cost required here is for the operation of online systems
Other goods, works and services	11.000	Salaries to support researchers and contributors
Remaining purchase costs (<15% of pers. Costs)	0	
Total	15.000	

University of Helsinki		
	Cost (€)	**Justification**
Travel and subsistence	4.000	Travel to distant countries requires the amount to be
Equipment	5.000	The reason why such a large amount was spent on the purchase of blue equipment is that the price of computing power has risen recently
Other goods, works and services	11.000	Salaries to support researchers and contributors
Remaining purchase costs (<15% of pers. Costs)	0	
Total	20.000	

STANDARD MODULAR EXTENSION OF PROPOSAL TEMPLATE:

1. **CLINICAL TRIALS**
    - **PART A: Additional question**
    - **PART B: Add an additional annex with information on clinical trials**

2. **CALLS FLAGGED AS SECURITY SENSITIVE**
    - **PART A: No additions**
    - **Part B: Add an additional annex with information on security**