Building the Legal Knowledge Graph for Smart Compliance Services in Multilingual Europe

D3.2 Intermediate translation services

PROJECT ACRONYM	Lynx
PROJECT TITLE	Building the Legal Knowledge Graph for Smart Compliance Services in Multilingual Europe
GRANT AGREEMENT	H2020-780602
FUNDING SCHEME	ICT-14-2017 - Innovation Action (IA)
STARTING DATE (DURATION)	01/12/2017 (36 months)
PROJECT WEBSITE	http://lynx-project.eu
COORDINATOR	Elena Montiel-Ponsoda (UPM)
RESPONSIBLE AUTHORS	Andis Lagzdiņš (TILDE)
CONTRIBUTORS	Artūrs Vasiļevskis (TILDE), Ēriks Ajausks (TILDE)
REVIEWERS	Julian Moreno-Schneider, Stefanie Hegele (DFKI), Ilan Kernerman (KD)
VERSION | STATUS	V0.4 |Draft
NATURE	Report
DISSEMINATION LEVEL	Public
DOCUMENT DOI	https://zenodo.org/communities/lynx/10.5281/zenodo.2580104
DATE	25/02/2019

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 780602

VERSION	MODIFICATION(S)	DATE	AUTHOR(S)
0.1	Draft of TOC	23/01/2019	Andis Lagzdiņš, Ēriks Ajausks
0.2	Draft of document body	06/02/2019	Andis Lagzdiņš, Ēriks Ajausks, Artūrs Vasiļevskis
0.3	Draft of document body	21/02/2019	Andis Lagzdiņš, Ēriks Ajausks, Artūrs Vasiļevskis
0.4	Draft of document body	25/02/2019	Andis Lagzdiņš, Ēriks Ajausks, Artūrs Vasiļevskis

ACRONYMS LIST

MT:	Machine Translation
NMT:	Neural Machine Translation
SMT:	Statistical Machine Translation
RDF:	Resource Description Framework
NIF:	Natural Language Interchange Format

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

EXECUTIVE SUMMARY

This deliverable summarizes the work done on the intermediate translation services (as part of Work Package 3) within the context of the Lynx project. The aim of the task is to provide a description of the services. There are two types of services available – terminology and translation. Terminology covers a cloud based terminology service for terminology management and terminology annotation. Translation service covers a cloud based machine translation service for customized machine translation systems.

This document serves as reference material about the corresponding technologies and their integration within the Lynx framework. The translation service provides all the facilities for customizing neural machine translation (NMT) engines for specific languages and domains, and includes sophisticated linguistic components. The terminology service provides user-friendly, collaborative, and multilingual terminology services to a broad spectrum of users for terminology work in practical application scenarios. The terminology service simplifies manual management of multilingual terminology including its processing, storage, sharing, and re-use via the rich terminology service functionalities of the Tilde Terminology cloud platform.

1 INTRODUCTION

1.1 PURPOSE OF THIS DOCUMENT

This document provides a description of the provided translation services whose objective is the processing and analyzing of the data collected in WP2 for training Machine Translation (MT) systems, thereby contributing to the creation of the Legal/Lynx Knowledge Graph (LKG). The separation of common services in a separate work package allows the Lynx partners to focus on the performance and robustness of individual generic services without tailoring them to any specific use case.

In this document, custom MT systems will be described with regard to their integration into the Lynx platform. This will be done via the Tilde MT API and used afterwards within the Lynx platform digital services for the Lynx use cases. The project will focus on European languages. The following languages will initially be involved in the pilots: English, Spanish, German, and Dutch. The data processing technologies at Tilde’s disposal are used on top of the data and documents acquired in WP2 to guarantee a reasonable quality of translation of legal texts, regulations, norms, and standards.

1.2 STRUCTURE OF THIS DOCUMENT

The structure of this document outlines intermediate translation services divided into two components – translation service and terminology service. In section 2.1. “Translation service”, we analyze the translation service and focus on describing the cloud platform and MT system training and evaluation. The MT system training and evaluation embeds information about the data used for the process of system training and shows training examples from training exercises. We continue with the translation service analysis by focusing on the general service and a detailed level of analysis of input and output requirements. The final subsection in section 2.1. “Translation service” lays out the description of the API.

In section 2.2. “Terminology extraction service”, we describe the content in a structure similar to the first part of this document. We begin with a description of the cloud platform of terminology services and analyze the main functionalities of the terminology service. In subsection 2.2.2., we give an outline of the extraction process that consists of several independent software modules. Furthermore, in subsection 2.2.3, the API description is summarized. In section 2.3., which is related to further work, we provide insights on further developments and point out certain action points to be considered for improving the services. Finally, the conclusion part sums up the main findings that stem from the whole document.

2 INTERMEDIATE TRANSLATION SERVICES

For the Lynx platform, there are two types of natural language related cloud services available – terminology and translation. The Intermediate translation service consists of two groups of services – translation and terminology extraction. The translation service contains, in its turn, two machine translation (MT) endpoints: one for NIF 2.11 based annotations, the format agreed for the interchange of annotations in Lynx, and one for client documents.

For the terminology service, there is a need for both monolingual data and bilingual (parallel a.k.a. translated) data extraction. The intermediate translation service includes only monolingual terminology extraction, but further work includes the creation of a bilingual terminology extraction workflow, using Tilde's competence and tools available for processing natural language.

2.1 TRANSLATION SERVICE

2.1.1 Description of Cloud platform

The Tilde MT platform provides all the facilities for customizing neural machine translation (NMT) engines for specific languages and domains. The platform also includes a suite of sophisticated linguistic components that provide linguistic knowledge for MT systems – tokenizers, sentence breakers, morphological analyzers and synthesizers, lemmatizers, part-of-speech and morpho-syntactic taggers, syntactic/dependency parsers, and named entity recognizers – as well as sophisticated tools for the correct processing of tags and placeholders, including HTML code.

The Tilde MT platform provides a full-service environment for private and publicly available systems that are hosted on the cloud and can be integrated into any platform or application. Alongside public MT systems, the user can convert its raw data into a fully customized MT system for specific language pairs, domains, and use cases.

The MT platform provides a full cycle of MT needs:

1. Data processing
2. Corpora collection
3. MT training and tuning
4. Terminology management
5. Quality estimation
6. Linguistic tool development
7. API integration
8. e-Services
9. Customer support

The platform consists of various components:

1. Administrative web interface - for training and managing translation systems
2. Public web interface - for translating with public MT systems
3. API for accessing translation services using a machine interface (provides communication using the SOAP standard or RESTful approach)
4. MT integration into other platforms:

[1] https://nif.readthedocs.io/en/latest/

   a. HTML/JavaScript based widget
   b. CAT tools
5. ITS 2.0 data categories support
6. Additional output segmentation information (phrases and their translations)

2.1.2 Service Description

The translation service supports the Lynx platform and Legal Knowledge Graph with automated machine translation. The translation service currently provides support for a runtime scenario as well as an endpoint for the Lynx platform asynchronous process in the background. Translation e-service runtime methods ensure translation of texts on-the-fly, while large documents and corpora of documents can be translated also asynchronously – by uploading documents and later downloading a translated document.

There are different translation needs in the Lynx platform. While in some cases there is a need for translation of a whole document, in other cases there could be a need for translation of annotations. The Lynx platform requires that translation is expressed as annotation of documents, ensuring the possibility to load that annotation in the Legal Knowledge Graph.

The translation service provides two separate endpoints that are dependent parts of the Tilde MT platform – the document translation endpoint and RDF/NIF translation endpoint. Both service endpoints use the same centralized MT cloud platform and exploit shared MT systems that are created for the Lynx platform and are in running state.

As input, the RDF/NIF translation endpoint receives plaintext or document annotations expressed by the NIF ontology. The translation service processes the NIF document, extracts out translatable fragments from the NIF document, and passes them to the common MT translation core service in a synchronous way. After all textual strings are translated, the translation service adds translation annotations to the original NIF document while preserving all previous information.

The document translation endpoint (also called – file translation feature) allows Lynx users to upload documents in various formats for translation with the specified MT system. Document translation can be a long running process, so it is scheduled to run offline. The Lynx platform can track status changes, and the translated file can later be downloaded from the MT cloud platform when the translation is completed. Supported file formats are DOC, DOCX, XLSX, PPTX, ODT, ODP, ODS, HTML, HTM, XHTML, XHT, TXT, TMX, XLF, XLIF, XLIFF, SDLXLIFF, TTX, RTF, PAGES. All documents must be in UTF-8 or UTF-16 encoding.

The scenario for document translation is the following (Figure 1):

1. The Lynx platform uploads a file and chooses the running MT system (NMT in the case of Lynx) for document translation.
2. The Tilde translation service forwards the document and the MT system ID to the Application Logic Layer where:
   a. A new repository object is created to store the uploaded file. The MT system ID is stored in the metadata of the repository object.
   b. A file translation task is submitted to the HPC Cluster.
3. The following steps are taken for the file translation task:
   a. Download a file from the repository.
   b. Convert the file to XLIFF using Okapi or another framework.

c. Iterate XLIFF nodes by translating text with inline tags. A text is translated with the translator service of the Logic Layer that handles sentence braking and translator load balancing.
d. Return the translated XLIFF back to the original document format.
e. Store the translated document to the resource repository.

4. The Lynx platform tracks the status changes and downloads the translated file when it is available.

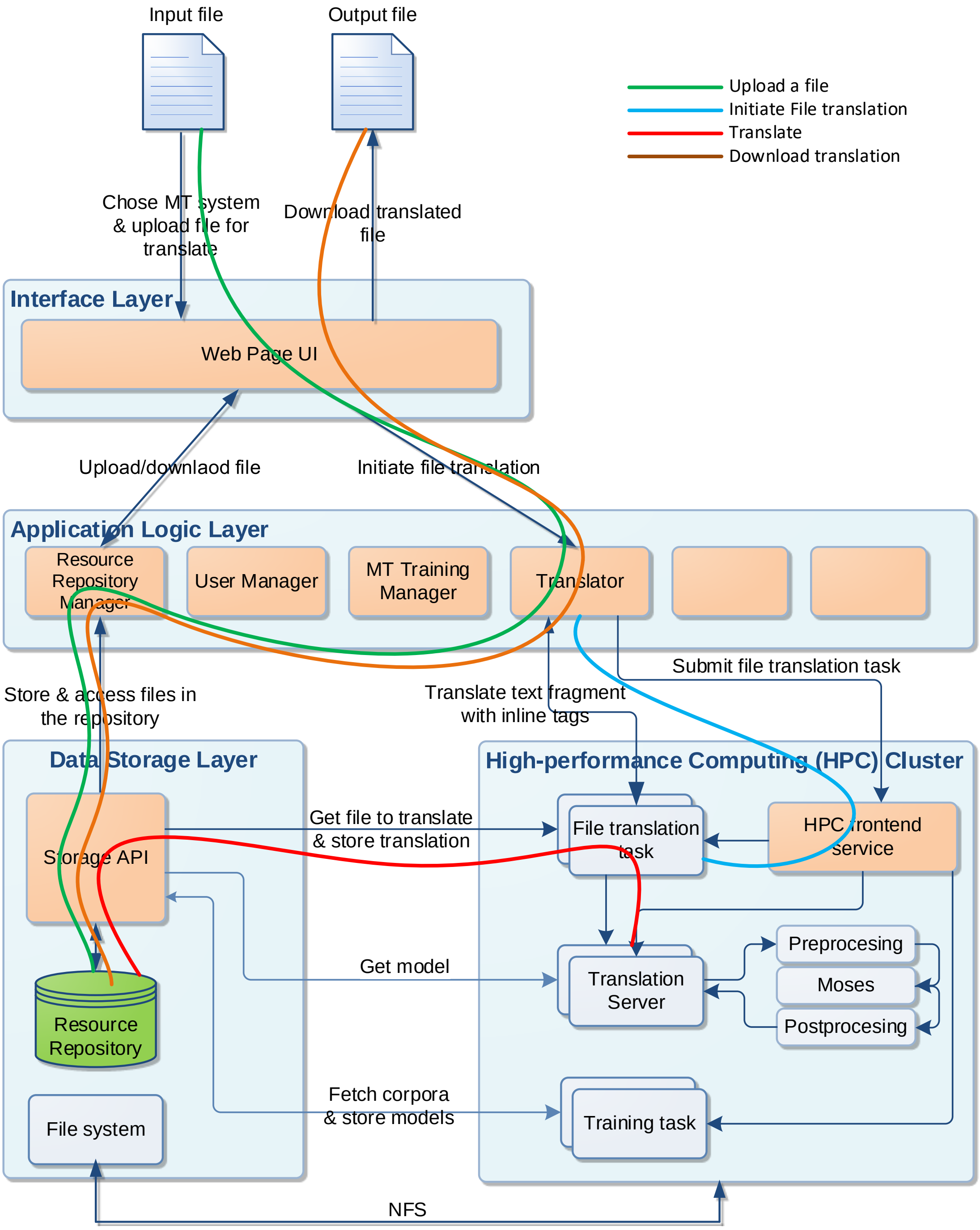


Figure 1 Document translation workflow in the Tilde MT cloud platform

2.1.3 API Description

The Lynx MT systems are mostly in running state, but in some cases they could also be in other states. The following system statuses are available:

- Running - ready for translation
- Waking up - will be ready for translation very soon
- Stand-by - systems in stand-by mode
- Failed - systems with error (this section is normally not visible)

Available systems mean that they are ready for translation, and standby systems mean they need to wake up. Systems will be woken up with the first request and need a few minutes to start up. Some systems will go to sleep if they are inactive for 60 minutes (no translation).

Currently there are six MT systems available for the Lynx project needs. Quality improvements of adjusted MT systems are described in the next chapter. The following MT systems are available (the id of each MT system is in brackets and can be used in the MT endpoint to select the appropriate system for translation):

1. English-German (smt-99b2f71a-1b3b-418e-bd6b-125f61a53feb), adjusted for the Use Case 1 Contracts (see deliverable 4.1)
2. German-English (smt-160de000-f719-4d5b-9daa-34859345e889), adjusted for the Use Case 1 Contracts
3. English-Spanish (smt-7f098605-5838-4f84-b73e-94af698c3e00), adjusted for Use Case 2 Labour Law
4. Spanish-English (*smt-4eafabb9-7cd6-4ae6-9dd6-6b7cc68925bb*), adjusted for Use Case 2 Labour Law
5. Dutch-English (*smt-2eb02c32-1406-45a0-8974-0310becf564b*), adjusted for the needs of Use Case 3 Energy
6. English-Dutch (*smt-8fc59d9e-5566-4e35-af4b-98382578cdf2*), adjusted for the needs of Use Case 3 Energy

RDF NIF translation endpoint

This is a synchronous web service, which provides translation functionality for texts and documents described with the NIF ontology.

There is one REST method for this endpoint, as input and output support plaintext or NIF document in various serializations. The input and output format specifications must be provided in request headers.

The endpoint is protected with Basic Authentication. For the Lynx platform, a user account was created to access the service. Thus, for each request, the Authorization header should be presented with a Basic authentication token.

Request

Action:

POST https://services.tilde.com/lynx/translation

Headers:

- Content-type – provides input format specification for the input
  - “text/plain” – for plaintext documents
  - “application/rdf+xml” – for NIF document in XML serialization

  - "text/turtle" – for NIF document in Turtle serialization
  - "application/n-triples" – for NIF document in Triples serialization
- Accept – provides output format specification for the input
  - "text/plain" – for plaintext documents
  - "application/rdf+xml" – for NIF document in XML serialization
  - "text/turtle" – for NIF document in Turtle serialization
  - "application/n-triples" – for NIF document in Triples serialization
- Authorization – Basic authentication token

Query parameters:

- sourceLang – language of input text
- targetLang – target language

Body:

As body, NIF/RDF document content must be included.

Response

As the response, clients will receive the original NIF document with added translation annotations.

Document translation endpoint

Document translation workflow consists of several steps that are asynchronous – document upload, checking the translation status, and document retrieval.

Document upload

This method allows to upload files and trigger the translation process.

Request

Action:

POST /ws/Service.svc/json/StartDocumentTranslation

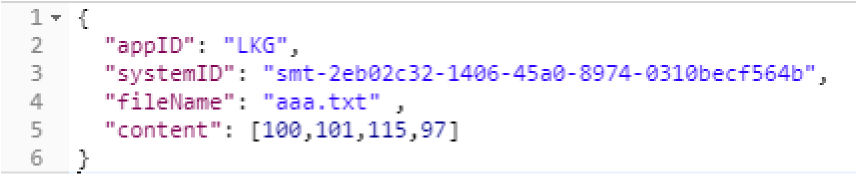
Headers:

- Content-type – "application/json"
- Client-id – secure token issued for Lynx platform (API key)

Body:

As body, a JSON structure must be submitted, where:

- appID – application ID (e.g., "LKG")
- systemID – translation system identification (see the list of systems in the beginning of this chapter)
- filename – filename of the document to be translated
- content – file content with byte array



```
{
  "appID": "LKG",
  "systemID": "smt-2eb02c32-1406-45a0-8974-0310becf564b",
  "fileName": "aaa.txt" ,
  "content": [100,101,115,97]
}
```

Figure 2 Document translation JSON structure

Response

In the response, the method returns the ID of this translation job to track status and retrieve the result later.

Document status

This method allows to retrieve translation status.

Request

Action:

GET https://www.letsmt.eu/ws/Service.svc/json/GetDocumentTranslationState?appID={0}&id={1}

Query parameters:

- appID – application ID (e.g., “LKG”)
- id – job identification for current translation job

Response

In the response, the method returns the JSON structure, where:

- ErrorCode – error code in case of translation error
- ErrorMessage – user friendly message in case of translation error
- FileName – filename of uploaded document
- Id – job ID
- Segments – count of segments in uploaded document
- Size – count of symbols in uploaded document
- Status – current status of the job
- System – translation system’s ID that is requested to translate the document
- TranslatedSegments – count of translated segments within the document

Download translated content

This method allows to retrieve the result – a translated document.

Request

Action:

GET https://www.letsmt.eu/ws/Service.svc/json/DownloadDocumentTranslation?appID={0}&id={1}

Query parameters:

- appID – application ID (e.g., “LKG”)

- id – job identification for current translation job

Response

In the response, the method returns the content of the translated document as a byte array.

2.1.4 MT system training and evaluation

Data

To train the MT systems for Lynx, we initially used all available and eligible data from the Tilde MT platform as general domain data and data from the specific domains (energy, financial, etc.) as in-domain data. Tuning and evaluation data were automatically sampled from the in-domain data. Those initial MT systems were trained using Statistical Machine Translation (SMT) technologies. An overview of the resulting dataset sizes after this early training stage is shown in Table 1.

Table 1 Amount of data used for training MT systems

	Sentence pairs
EN-NL	41,639,299
EN-ES	81,176,632
EN-DE	24,768,821

Baseline SMT Systems

All Spanish ↔ English and Dutch ↔ English SMT systems were trained using the state-of-the-art SMT training technologies that are available in the Tilde MT platform. The systems are phrase-based SMT systems, which feature 7-gram translation models and 3-gram language models. (As the monolingual data is of very large quantity, it was decided to train 3-gram language models so that the systems can be used in practical application scenarios whenever necessary.) All systems were trained using the Moses SMT toolkit (Koehn et al., 2007) on the Tilde MT platform.

Language models were trained with KenLM (Heafield, 2011). The systems were tuned with MERT (Bertoldi et al., 2009) using the tuning sets. The systems were automatically evaluated using a standard metric for automatic evaluation of machine translation - BLEU (Papineni et al., 2002). The automatic evaluation results for all systems are provided in Table 2.

Updated NMT Systems

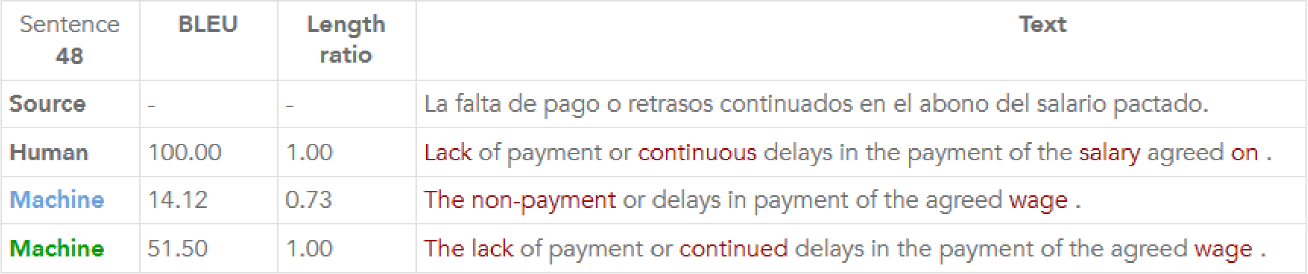
To outperform the baselines, we used all the data from the Lynx SMT training and trained NMT systems with multiplicative long short-term memory (Krause et al., 2017) (MLSTM) and a shared subword unit vocabulary (Sennrich et al., 2016) of 25,000 tokens. The updated systems were trained using the Marian NMT training toolkit (Junczys-Dowmunt et al., 2018). The results in Table 2 show that NMT systems outperform their SMT counterparts in all cases.

Table 2 Results (BLEU scores – higher is better) from MT system automatic evaluation

	SMT	NMT
EN-NL	30.45	34.12
NL-EN	32.62	43.54
EN-ES	28.26	38.36
ES-EN	25.77	32.52
EN-DE	37.01	38.73
DE-EN	41.54	44.73

Translation Examples

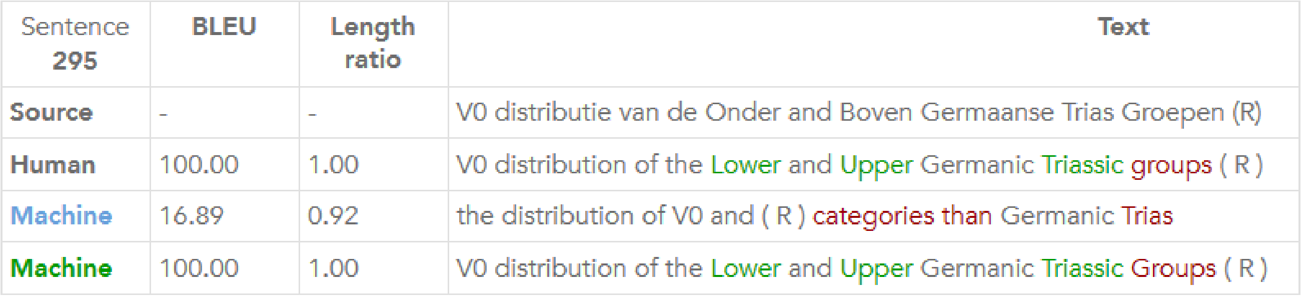
The example sentence shown in Figure 3 shows that the SMT system (blue) has omitted several words from the source sentence and therefore **changed** the whole meaning of the translation. The output from the NMT system (green) is also not a perfect match with the source, but at least it keeps most of the meaning intact.



Sentence 48	BLEU	Length ratio	Text
Source	-	-	La falta de pago o retrasos continuados en el abono del salario pactado.
Human	100.00	1.00	Lack of payment or continuous delays in the payment of the salary agreed on .
Machine	14.12	0.73	The non-payment or delays in payment of the agreed wage .
Machine	51.50	1.00	The lack of payment or continued delays in the payment of the agreed wage .

Figure 3 An example translation from Spanish into English. The upper machine translation is SMT and the lower is NMT

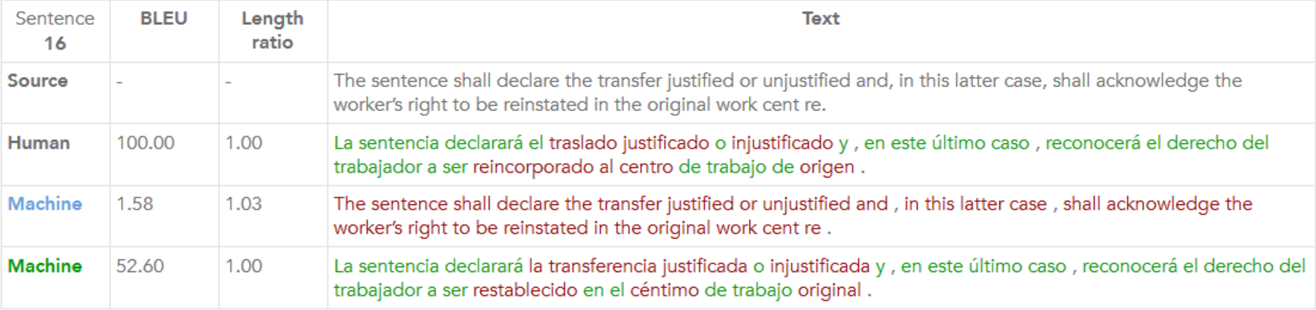
Another example, shown in Figure 4, exhibits the SMT system's difficulty to deal with rare words, while the NMT system handles the whole sentence perfectly.



Sentence 295	BLEU	Length ratio	Text
Source	-	-	V0 distributie van de Onder and Boven Germaanse Trias Groepen (R)
Human	100.00	1.00	V0 distribution of the Lower and Upper Germanic Triassic groups (R)
Machine	16.89	0.92	the distribution of V0 and (R) categories than Germanic Trias
Machine	100.00	1.00	V0 distribution of the Lower and Upper Germanic Triassic Groups (R)

Figure 4 An example translation from Dutch into English. The upper machine translation is SMT and the lower is NMT

In the case of very complex sentences, at times, the SMT system failed to translate anything at all. One such example is given in Figure 5. The updated NMT system can handle such sentences much better.



Sentence 16	BLEU	Length ratio	Text
Source	-	-	The sentence shall declare the transfer justified or unjustified and, in this latter case, shall acknowledge the worker's right to be reinstated in the original work cent re.
Human	100.00	1.00	La sentencia declarará el traslado justificado o injustificado y , en este último caso , reconocerá el derecho del trabajador a ser reincorporado al centro de trabajo de origen .
Machine	1.58	1.03	The sentence shall declare the transfer justified or unjustified and , in this latter case , shall acknowledge the worker's right to be reinstated in the original work cent re .
Machine	52.60	1.00	La sentencia declarará la transferencia justificada o injustificada y , en este último caso , reconocerá el derecho del trabajador a ser restablecido en el céntimo de trabajo original .

Figure 5 An example translation from English into Spanish. The upper machine translation is SMT and the lower is NMT

2.2 TERMINOLOGY EXTRACTION SERVICE

2.2.1 Description of the Cloud platform

Tilde Terminology is a cloud platform for terminology work. The platform provides user-friendly, collaborative, and multilingual terminology services to a wide range of users, including human and machine users, to facilitate terminology work in practical application scenarios. Tilde Terminology simplifies manual management of multilingual terminology including its processing, storage, sharing, and re-use via the rich terminology service functionalities of the Tilde Terminology cloud platform. The portal also provides secure access to terminology collections using API methods.

Tilde Terminology also provides powerful functionality for automated term extraction from full-text documents, automated translation look-up (from online term bases), term approval and post-editing, term glossary management, term sharing with other linguists and team members, term re-use, term export as TBX, TSV, CSV, and MT system training.

The main functionalities of the terminology portal include:

1. Search for terms in various sources
2. Identify terms automatically in documents
3. Look up translation candidates for terms
4. Collaborate with colleagues
5. Integrate terms in CAT and MT

2.2.2 Service Description

A new terminology extraction API endpoint has been created; the terminology extraction service extracts term candidates from uploaded corpora. The extraction process consists of several steps: upload of multiple documents, analyze their format, extract plaintext, mark term candidates in plaintext version, extract and group all term candidates, and finally make a single terminology collection.

The extraction process consists of several independent software modules that are organized in a workflow (see Figure 6). The extraction workflow consists of the following steps:

- Plain text extraction
- Monolingual term candidate extraction
- Collection creation
- Result conversion

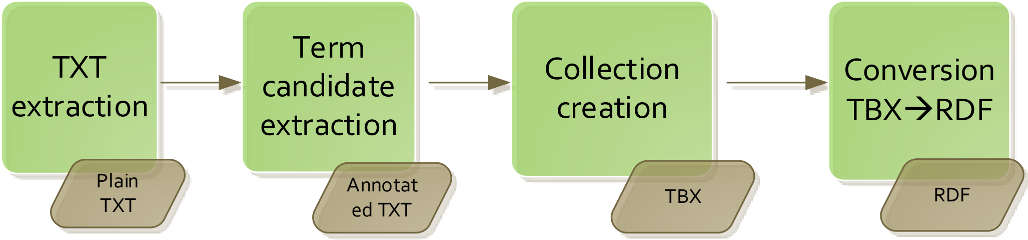


Figure 6 Terminology extraction workflow

Plain text extraction

The first step of the terminology extraction process is the plain text extraction from the documents uploaded to the Lynx platform. As the Lynx platform allows to receive different types of files in different encodings, this module is responsible for the correct interpretation of a file structure, encoding, and content. First, the main module of plain text detects the file format and encoding against the Lynx provided language parameter and automatically decides which tool will be used for the text extraction. The plain text extraction process is executed for each file, and the result is a TXT file that is saved in the Terminology File Store. This text is an intermediate result and is used later as an input for the later software modules in the workflow.

Term candidate extraction

Monolingual term candidate extraction includes a set of software modules that are responsible for the acquisition of monolingual term candidates in the source language selected by the user. As input, these modules receive the plain text from the source documents and the project properties provided by the user. The result of these modules is a monolingual TBX document. Source term candidate extraction modules include (see Figure 7):

- *Term markers in plain text* – annotates text with possible term candidates
- *Term normalizer* – appends canonical forms and morpho-syntactic information about canonical forms to the marked terms of documents
- *Single Collection Creator* – compounds all terms from tagged documents into a single terminology collection. The terms are grouped by canonical forms and morpho-syntactic information, and a single monolingual terminology collection is created including references to the source files and collocations

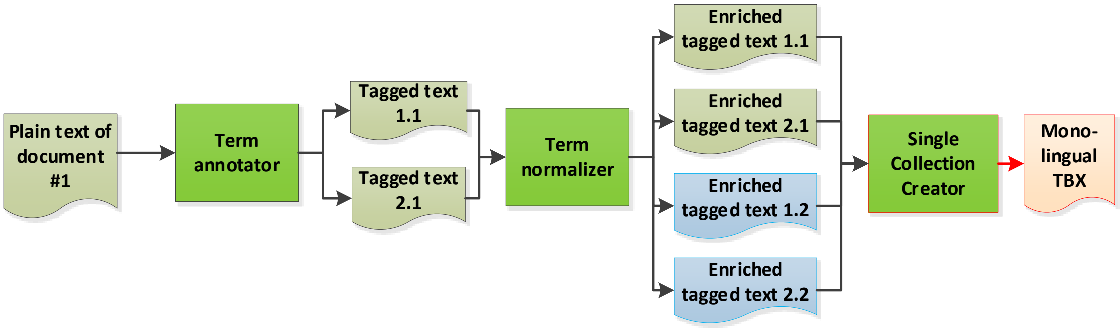


Figure 7 Monolingual term extraction

The terminology extraction service is planned to be used before the Entity extraction service. Entity extraction requires the data to be in RDF SKOS format, so the last step is conversion from TBX to RDF. An example of the terminology extraction result is provided in the Annex of this document.

Language support

The Tilde Terminology functionality for term candidate extraction is a linguistic process that requires different linguistic tools to achieve higher term candidate extraction quality. Therefore, depending on different language specific tools and resources, languages can be divided into four categories – A level, B level, C level, and D level. Lynx focuses on these languages: English, Dutch, Spanish, German, and Italian. The Lynx language support levels are highlighted below.

A level languages have the highest level of support in the Tilde Terminology platform for term tagging and term normalization. *A level languages* are **English** and Latvian. The following linguistic tools are available for *A level languages*:

- Part-of-speech (or morpho-syntactic) taggers trained on (high quality) human annotated training data
- Lemmatizers, which allow performing better statistical analysis for term candidate extraction
- Morphological analyzers and synthesizers, which are required for term normalization
- Rule-based term normalizers, which allow reducing redundancy in the extracted term candidate lists

B level languages have the highest level of support in the Tilde Terminology platform for term tagging; however, they do not have a term normalization tool. *B level languages* are **German, Spanish**, Estonian, French, Hungarian, **Italian**, Lithuanian, and **Dutch**. The following linguistic tools are available for *B level languages*:

- Part-of-speech (or morpho-syntactic) taggers trained on (high quality) human annotated training data
- Lemmatizers, which allow performing better statistical analysis for term candidate extraction and provide basic support for redundancy reduction in the extracted term candidate lists

C level languages have basic support in the Tilde Terminology platform for term tagging, but they do not have a term normalization tool. The linguistic tools are part-of-speech taggers trained on (lower quality) automatically annotated training data. *C level languages* are Bulgarian, Czech, Danish, Greek, Finnish, Croatian, Maltese, Polish, Portuguese, Romanian, Russian, Slovak, Slovenian, and Swedish.

D level languages have no linguistic tool support in the Tilde Terminology platform for term tagging, and they do not have a term normalization tool. Term candidate extraction for these languages is based on language independent methods. *D level languages* are Irish, Turkish, Norwegian (Nynorsk), Catalan, and Galician.

2.2.3 API Description

Monolingual terminology extraction is an asynchronous process, and the workflow is the following:

- Creating a project and retrieving the job id for later reference
- Upload of files while full corpus is uploaded to the terminology share. If there are huge amounts of files, FTP or another solution could be used.
- Starting the extraction process and providing workflow extraction parameters
- Retrieval of extraction status
- Retrieval of extracted collection

The endpoint is protected with Basic Authentication. For the Lynx platform, a user account is created to access the service. Thus, for each request, the Authorization header should be present with a Basic authentication token.

Extraction project creation

This method allows creating a new project and getting its unique reference.

<u>Request</u>

Action:

GET https://term.tilde.com/svc/extraction/project/create

Headers:

- Authorization – Basic authentication token

Response

In the response, the method returns the ID of this just created project.

File upload to the extraction project

This method allows to upload one or more files to the extraction project.

Request

Action:

POST https://term.tilde.com/svc/extraction/files/{projectID}

Headers:

- Authorization – Basic authentication token
- Content-Type – "application/x-www-form-urlencoded"

URL Parameters:

ProjectID – ID of extraction project

Response

In the response, the method returns HTTP status code 200 OK in case of success.

Start term extraction

This method allows starting the terminology extraction workflow.

Request

Action:

POST https://term.tilde.com/svc/extraction/project/{projectID}/start?lang={lang}

Headers:

- Authorization – Basic authentication token

URL Parameters:

- projectID – ID of extraction project
- lang – ISO 2-symbol language code, language of text corpora, the source language

Response

In the response, the method returns HTTP status code 200 OK in case of success.

Retrieve term extraction status

This method allows getting the status of the extraction process.

Request

Action:

POST https://term.tilde.com/svc/extraction/project/{projectID}/status

Headers:

- Authorization – Basic authentication token

URL Parameters:

- projectID – ID of extraction project

Response

In the response, the method returns HTTP status code 200 OK and a status message in the response body. Possible values are the following: NotStarted, Waiting, Executing, Completed, Crashed.

Stop term extraction

In case of incorrect parameters or update of source files, it is possible to stop the ongoing executing process and start again later.

Request

Action:

POST https://term.tilde.com/svc/extraction/project/{projectID}/stop

Headers:

- Authorization – Basic authentication token

URL Parameters:

- projectID – ID of extraction project

Response

In the response, the method returns HTTP status code 200 OK in case of success.

Get term extraction results

This method returns the results of the extraction process, providing a valid answer only if the extraction process status is "completed". It can return data in TBX or RDF format (use Accept header to specify format).

Request

Action:

POST https://term.tilde.com/svc/extraction/project/{projectID}/result

Headers:

- Authorization – Basic authentication token
- Accept – “application/xml” for TBX format, “text/turtle” for SKOS/RDF result

URL Parameters:

- projectID – ID of extraction project

Response

In the response, the method returns a terminology collection in TBX or SKOS format.

2.3 FURTHER WORK

Regarding further work – both translation services and all their endpoints will be analysed to see how they are used in the Lynx platform. Any other use case requirements that might appear in relation to the provided translation services will also be analysed. These analyses will lead to improve the scalability of the system, quality, and integrity with the Lynx platform.

All services will be adjusted to fit best with the technologies in the Lynx framework – authentication, authorization, API semantics, and data formats extension.

In order to make further development, the following action points will be developed for Lynx services.

Translation services (until M24):

- Retrain MT systems. System improvements will be based on WP2 data and feedback provided by partners on MT system quality performance.
- Train new MT system language pairs English-Italian-English. Other language combinations and/or domains will be trained on partner’s request.

Terminology extraction service (until M24):

- Provide multilingual support, extract term candidates from comparable and/or parallel data
- Tune and extend data returned in SKOS format

The final version of translation services will be described in Deliverable D3.6 “Translation services” (implementation and report) in M27.

CONCLUSIONS

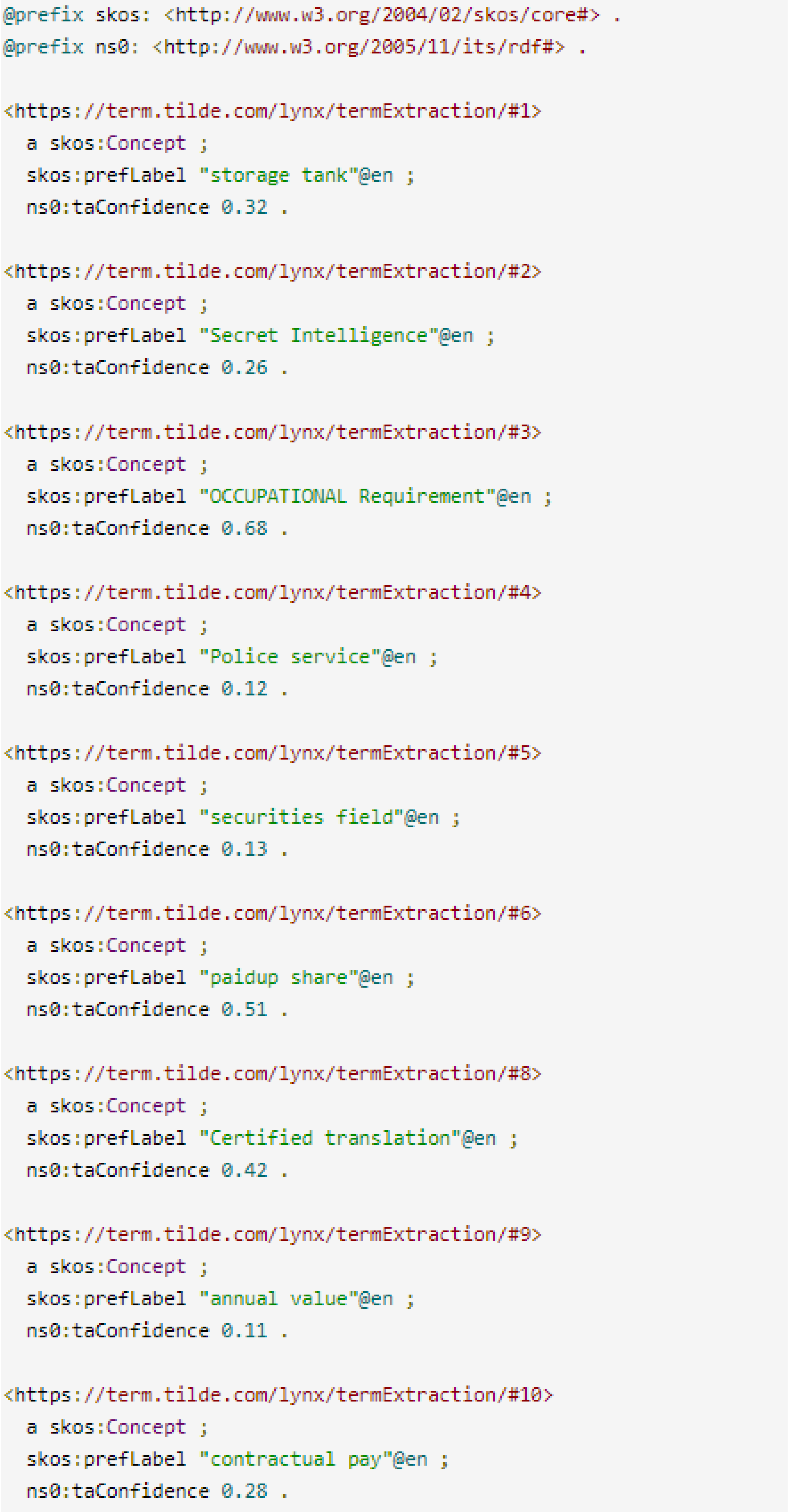
This report provides a description of the intermediate translation services (as defined in D3.2). Two services are described: translation service and terminology service. Both services include the following parts:

- Description of the Cloud platform in general
- Lynx Service Description
- Lynx APIs Description

Future work for both services includes analysis from the perspective of usability in the Lynx platform. Usability based on other use case requirements might be applied, if relevant. The future analysis will be based on principles of scalability, quality, and integrity with the Lynx platform. The final version of both services will be described in the D3.6 report “Translation services”.

ANNEX

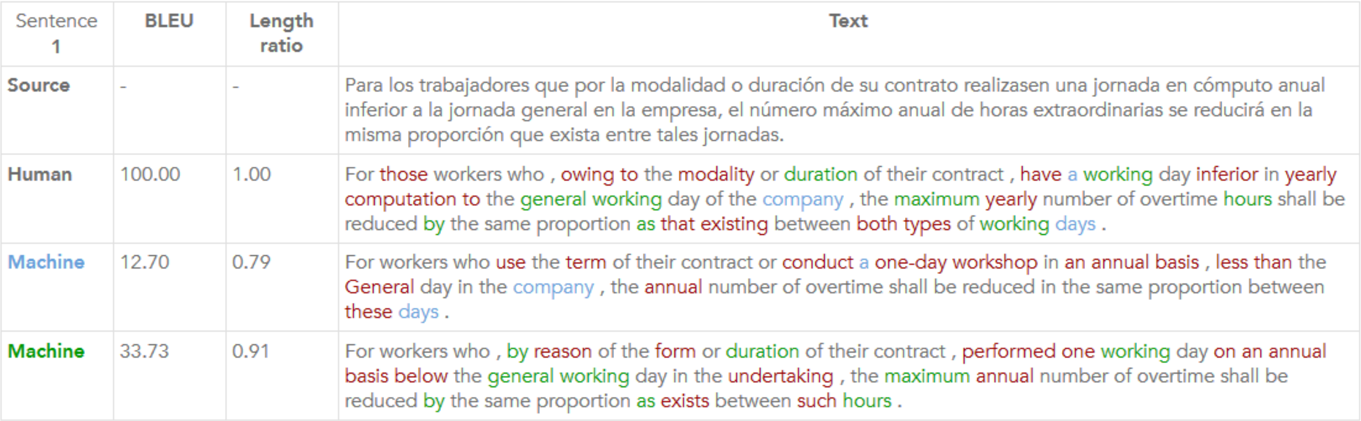
Example of Terminology extraction result (in RDF format)



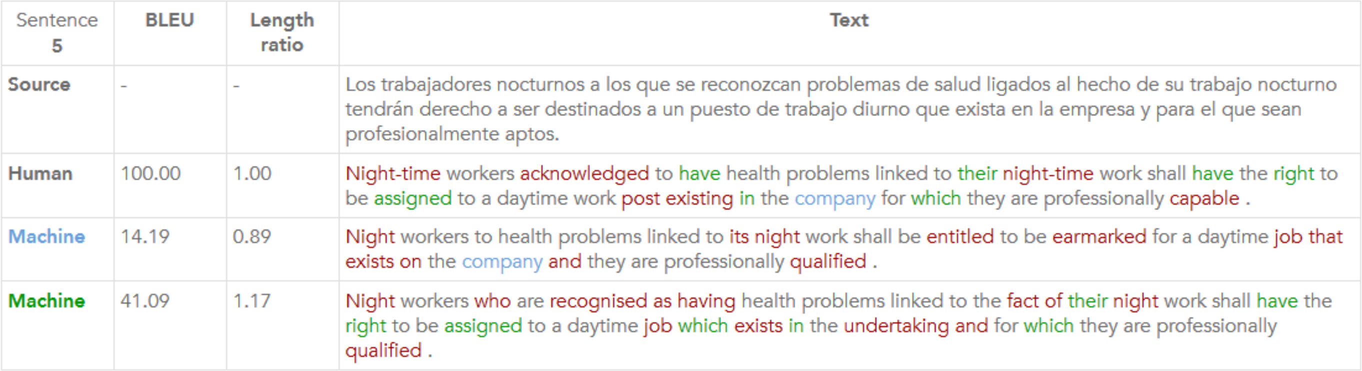
```
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix ns0: <http://www.w3.org/2005/11/its/rdf#> .

<https://term.tilde.com/lynx/termExtraction/#1>
  a skos:Concept ;
  skos:prefLabel "storage tank"@en ;
  ns0:taConfidence 0.32 .

<https://term.tilde.com/lynx/termExtraction/#2>
  a skos:Concept ;
  skos:prefLabel "Secret Intelligence"@en ;
  ns0:taConfidence 0.26 .

<https://term.tilde.com/lynx/termExtraction/#3>
  a skos:Concept ;
  skos:prefLabel "OCCUPATIONAL Requirement"@en ;
  ns0:taConfidence 0.68 .

<https://term.tilde.com/lynx/termExtraction/#4>
  a skos:Concept ;
  skos:prefLabel "Police service"@en ;
  ns0:taConfidence 0.12 .

<https://term.tilde.com/lynx/termExtraction/#5>
  a skos:Concept ;
  skos:prefLabel "securities field"@en ;
  ns0:taConfidence 0.13 .

<https://term.tilde.com/lynx/termExtraction/#6>
  a skos:Concept ;
  skos:prefLabel "paidup share"@en ;
  ns0:taConfidence 0.51 .

<https://term.tilde.com/lynx/termExtraction/#8>
  a skos:Concept ;
  skos:prefLabel "Certified translation"@en ;
  ns0:taConfidence 0.42 .

<https://term.tilde.com/lynx/termExtraction/#9>
  a skos:Concept ;
  skos:prefLabel "annual value"@en ;
  ns0:taConfidence 0.11 .

<https://term.tilde.com/lynx/termExtraction/#10>
  a skos:Concept ;
  skos:prefLabel "contractual pay"@en ;
  ns0:taConfidence 0.28 .
```

Examples of MT system training results

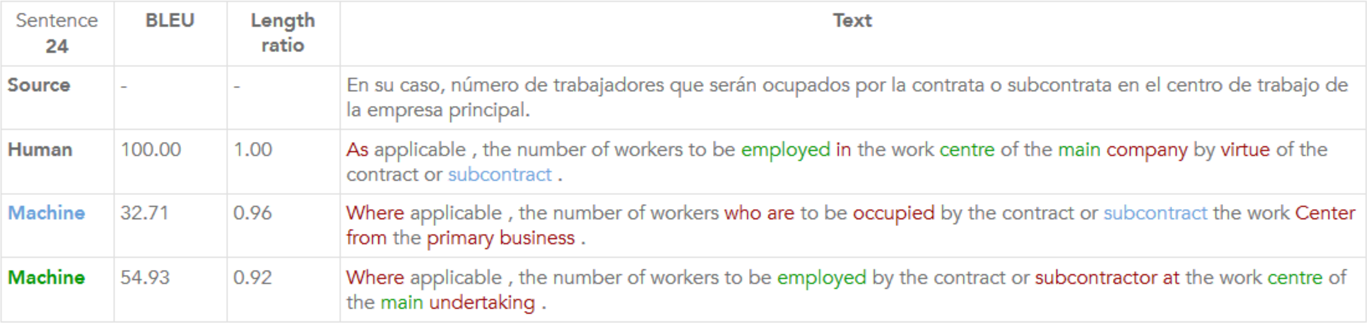
Spanish → English



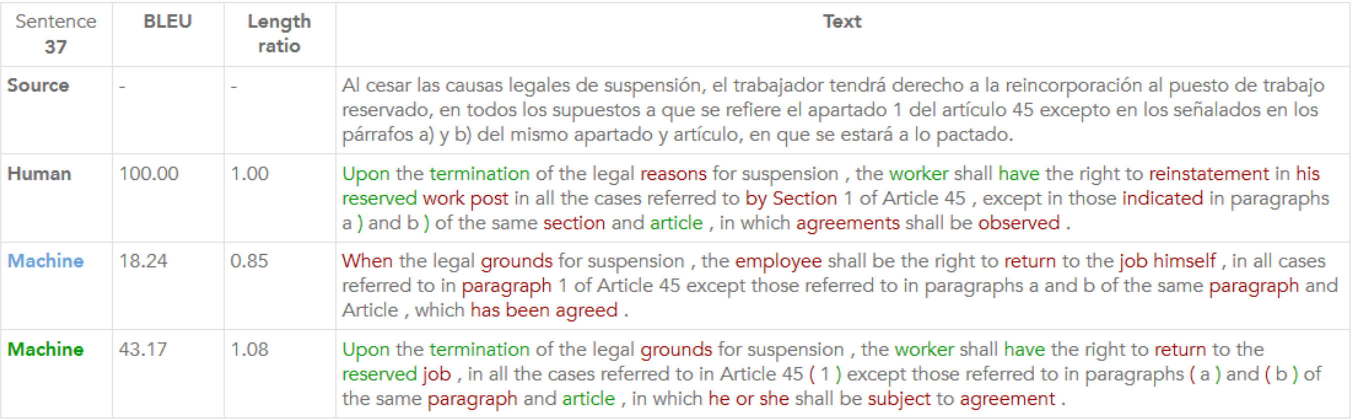
Sentence 1	BLEU	Length ratio	Text
Source	-	-	Para los trabajadores que por la modalidad o duración de su contrato realizasen una jornada en cómputo anual inferior a la jornada general en la empresa, el número máximo anual de horas extraordinarias se reducirá en la misma proporción que exista entre tales jornadas.
Human	100.00	1.00	For those workers who , owing to the modality or duration of their contract , have a working day inferior in yearly computation to the general working day of the company , the maximum yearly number of overtime hours shall be reduced by the same proportion as that existing between both types of working days .
Machine	12.70	0.79	For workers who use the term of their contract or conduct a one-day workshop in an annual basis , less than the General day in the company , the annual number of overtime shall be reduced in the same proportion between these days .
Machine	33.73	0.91	For workers who , by reason of the form or duration of their contract , performed one working day on an annual basis below the general working day in the undertaking , the maximum annual number of overtime shall be reduced by the same proportion as exists between such hours .



Sentence 5	BLEU	Length ratio	Text
Source	-	-	Los trabajadores nocturnos a los que se reconozcan problemas de salud ligados al hecho de su trabajo nocturno tendrán derecho a ser destinados a un puesto de trabajo diurno que exista en la empresa y para el que sean profesionalmente aptos.
Human	100.00	1.00	Night-time workers acknowledged to have health problems linked to their night-time work shall have the right to be assigned to a daytime work post existing in the company for which they are professionally capable .
Machine	14.19	0.89	Night workers to health problems linked to its night work shall be entitled to be earmarked for a daytime job that exists on the company and they are professionally qualified .
Machine	41.09	1.17	Night workers who are recognised as having health problems linked to the fact of their night work shall have the right to be assigned to a daytime job which exists in the undertaking and for which they are professionally qualified .

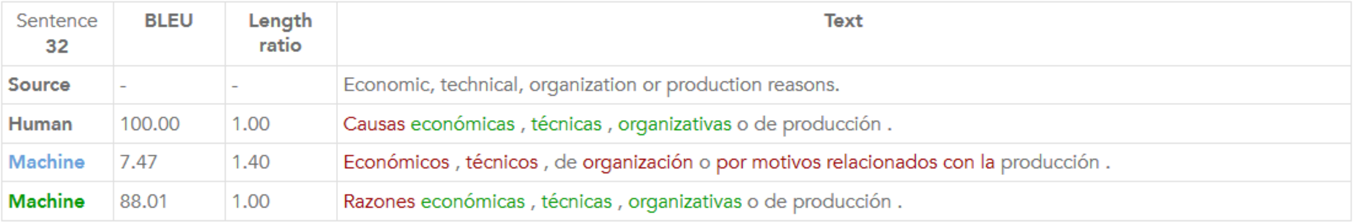


Sentence 24	BLEU	Length ratio	Text
Source	-	-	En su caso, número de trabajadores que serán ocupados por la contrata o subcontrata en el centro de trabajo de la empresa principal.
Human	100.00	1.00	As applicable , the number of workers to be employed in the work centre of the main company by virtue of the contract or subcontract .
Machine	32.71	0.96	Where applicable , the number of workers who are to be occupied by the contract or subcontract the work Center from the primary business .
Machine	54.93	0.92	Where applicable , the number of workers to be employed by the contract or subcontractor at the work centre of the main undertaking .

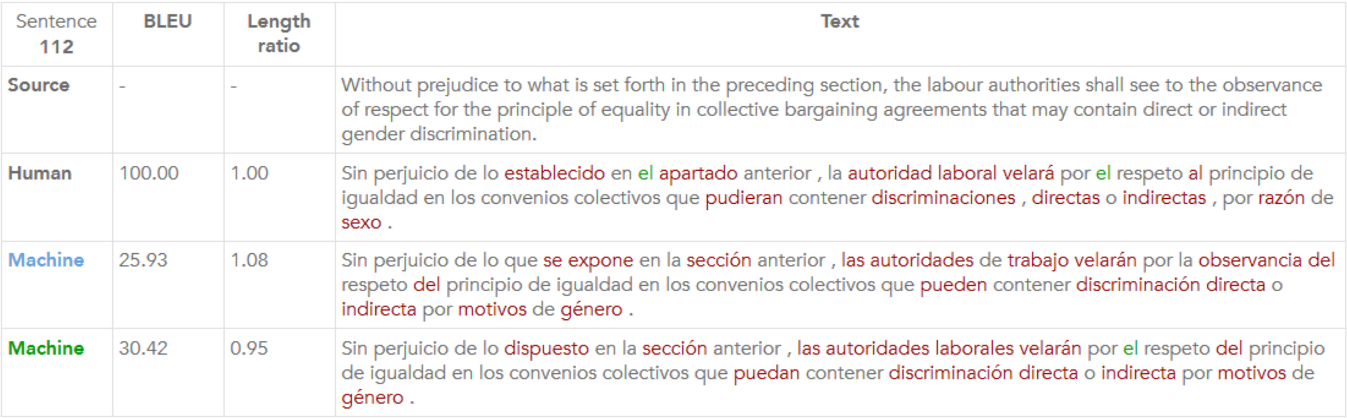


Sentence 37	BLEU	Length ratio	Text
Source	-	-	Al cesar las causas legales de suspensión, el trabajador tendrá derecho a la reincorporación al puesto de trabajo reservado, en todos los supuestos a que se refiere el apartado 1 del artículo 45 excepto en los señalados en los párrafos a) y b) del mismo apartado y artículo, en que se estará a lo pactado.
Human	100.00	1.00	Upon the termination of the legal reasons for suspension , the worker shall have the right to reinstatement in his reserved work post in all the cases referred to by Section 1 of Article 45 , except in those indicated in paragraphs a) and b) of the same section and article , in which agreements shall be observed .
Machine	18.24	0.85	When the legal grounds for suspension , the employee shall be the right to return to the job himself , in all cases referred to in paragraph 1 of Article 45 except those referred to in paragraphs a and b of the same paragraph and Article , which has been agreed .
Machine	43.17	1.08	Upon the termination of the legal grounds for suspension , the worker shall have the right to return to the reserved job , in all the cases referred to in Article 45 (1) except those referred to in paragraphs (a) and (b) of the same paragraph and article , in which he or she shall be subject to agreement .

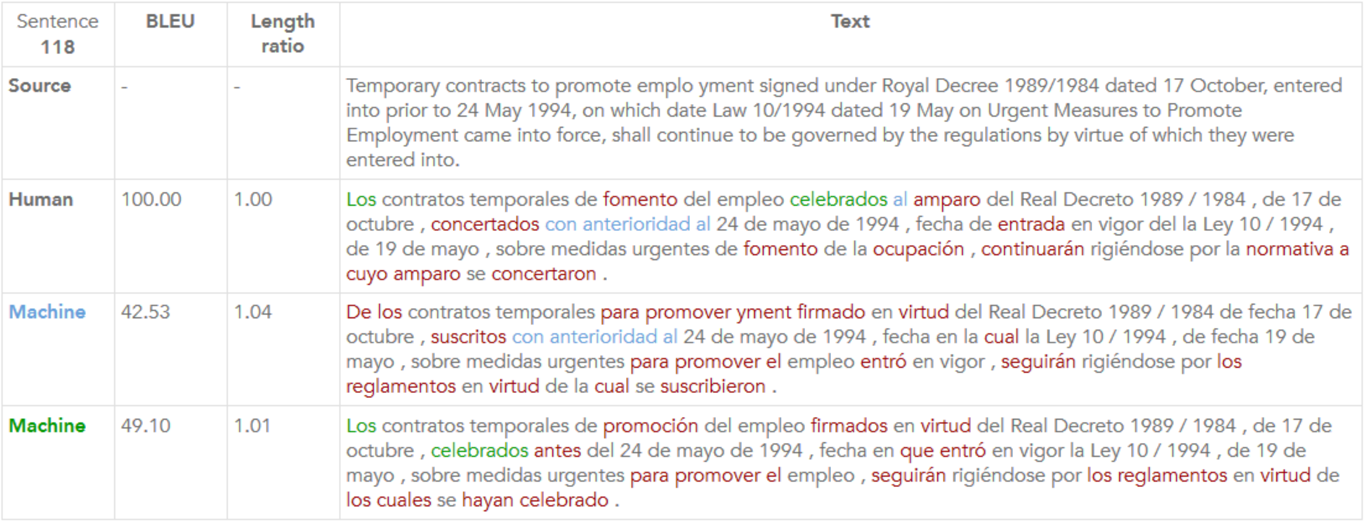
English → Spanish



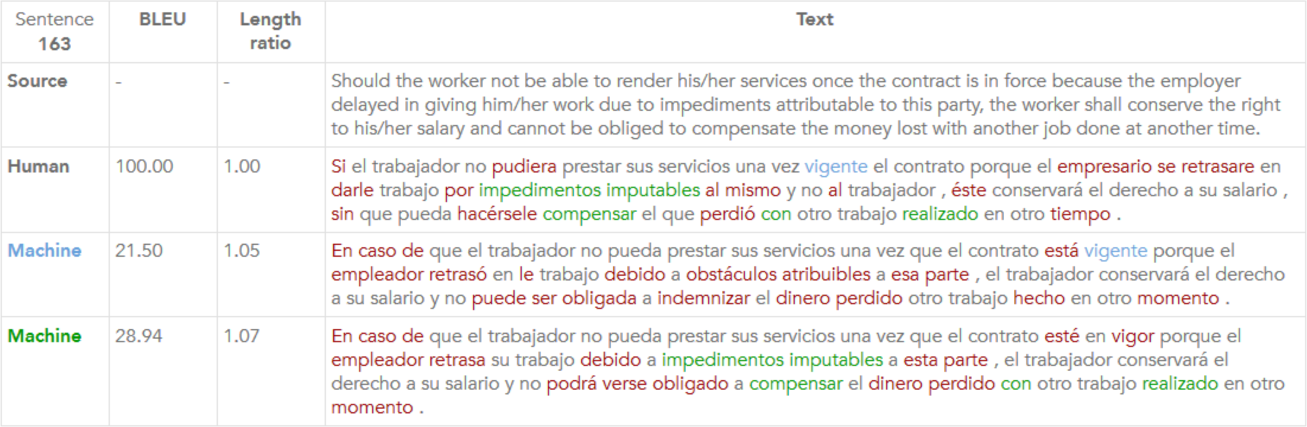
Sentence 32	BLEU	Length ratio	Text
Source	-	-	Economic, technical, organization or production reasons.
Human	100.00	1.00	Causas económicas , técnicas , organizativas o de producción .
Machine	7.47	1.40	Económicos , técnicos , de organización o por motivos relacionados con la producción .
Machine	88.01	1.00	Razones económicas , técnicas , organizativas o de producción .



Sentence 112	BLEU	Length ratio	Text
Source	-	-	Without prejudice to what is set forth in the preceding section, the labour authorities shall see to the observance of respect for the principle of equality in collective bargaining agreements that may contain direct or indirect gender discrimination.
Human	100.00	1.00	Sin perjuicio de lo establecido en el apartado anterior , la autoridad laboral velará por el respeto al principio de igualdad en los convenios colectivos que pudieran contener discriminaciones , directas o indirectas , por razón de sexo .
Machine	25.93	1.08	Sin perjuicio de lo que se expone en la sección anterior , las autoridades de trabajo velarán por la observancia del respeto del principio de igualdad en los convenios colectivos que pueden contener discriminación directa o indirecta por motivos de género .
Machine	30.42	0.95	Sin perjuicio de lo dispuesto en la sección anterior , las autoridades laborales velarán por el respeto del principio de igualdad en los convenios colectivos que puedan contener discriminación directa o indirecta por motivos de género .

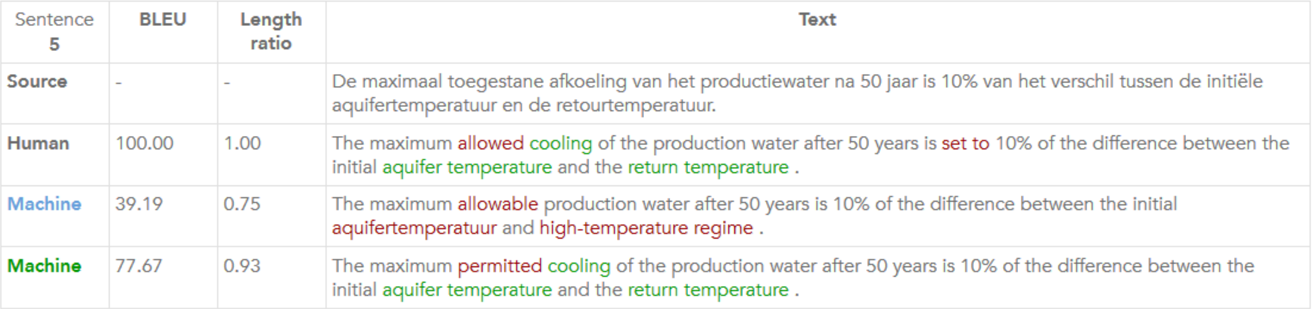


Sentence 118	BLEU	Length ratio	Text
Source	-	-	Temporary contracts to promote emplo yment signed under Royal Decree 1989/1984 dated 17 October, entered into prior to 24 May 1994, on which date Law 10/1994 dated 19 May on Urgent Measures to Promote Employment came into force, shall continue to be governed by the regulations by virtue of which they were entered into.
Human	100.00	1.00	Los contratos temporales de fomento del empleo celebrados al amparo del Real Decreto 1989 / 1984 , de 17 de octubre , concertados con anterioridad al 24 de mayo de 1994 , fecha de entrada en vigor del la Ley 10 / 1994 , de 19 de mayo , sobre medidas urgentes de fomento de la ocupación , continuarán rigiéndose por la normativa a cuyo amparo se concertaron .
Machine	42.53	1.04	De los contratos temporales para promover yment firmado en virtud del Real Decreto 1989 / 1984 de fecha 17 de octubre , suscritos con anterioridad al 24 de mayo de 1994 , fecha en la cual la Ley 10 / 1994 , de fecha 19 de mayo , sobre medidas urgentes para promover el empleo entró en vigor , seguirán rigiéndose por los reglamentos en virtud de la cual se suscribieron .
Machine	49.10	1.01	Los contratos temporales de promoción del empleo firmados en virtud del Real Decreto 1989 / 1984 , de 17 de octubre , celebrados antes del 24 de mayo de 1994 , fecha en que entró en vigor la Ley 10 / 1994 , de 19 de mayo , sobre medidas urgentes para promover el empleo , seguirán rigiéndose por los reglamentos en virtud de los cuales se hayan celebrado .

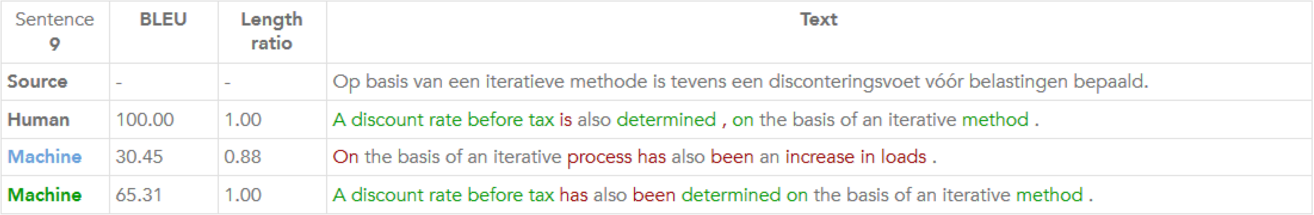


Sentence 163	BLEU	Length ratio	Text
Source	-	-	Should the worker not be able to render his/her services once the contract is in force because the employer delayed in giving him/her work due to impediments attributable to this party, the worker shall conserve the right to his/her salary and cannot be obliged to compensate the money lost with another job done at another time.
Human	100.00	1.00	Si el trabajador no pudiera prestar sus servicios una vez vigente el contrato porque el empresario se retrasare en darle trabajo por impedimentos imputables al mismo y no al trabajador , éste conservará el derecho a su salario , sin que pueda hacérsele compensar el que perdió con otro trabajo realizado en otro tiempo .
Machine	21.50	1.05	En caso de que el trabajador no pueda prestar sus servicios una vez que el contrato está vigente porque el empleador retrasó en le trabajo debido a obstáculos atribuibles a esa parte , el trabajador conservará el derecho a su salario y no puede ser obligada a indemnizar el dinero perdido otro trabajo hecho en otro momento .
Machine	28.94	1.07	En caso de que el trabajador no pueda prestar sus servicios una vez que el contrato esté en vigor porque el empleador retrasa su trabajo debido a impedimentos imputables a esta parte , el trabajador conservará el derecho a su salario y no podrá verse obligado a compensar el dinero perdido con otro trabajo realizado en otro momento .

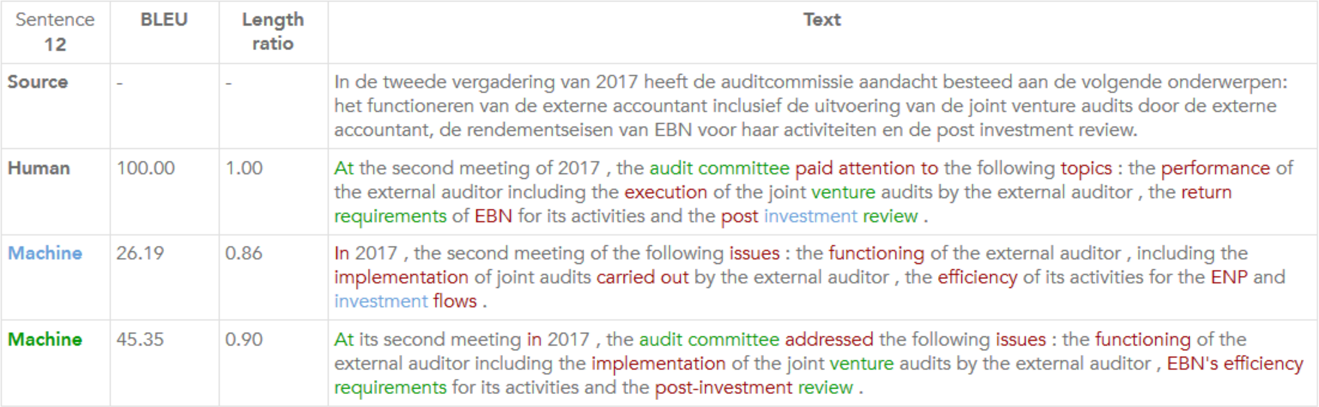
Dutch → English



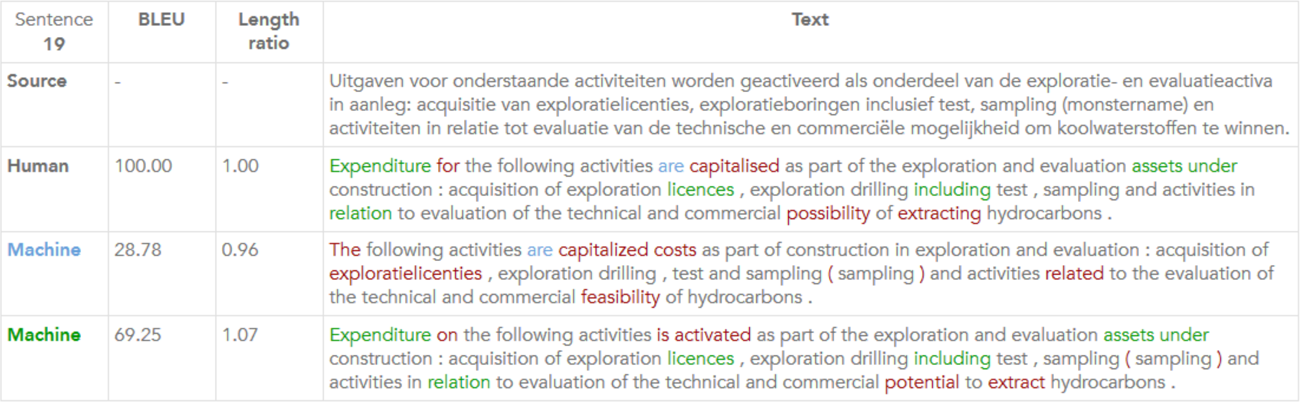
Sentence 5	BLEU	Length ratio	Text
Source	-	-	De maximaal toegestane afkoeling van het productiewater na 50 jaar is 10% van het verschil tussen de initiële aquifertemperatuur en de retourtemperatuur.
Human	100.00	1.00	The maximum allowed cooling of the production water after 50 years is set to 10% of the difference between the initial aquifer temperature and the return temperature .
Machine	39.19	0.75	The maximum allowable production water after 50 years is 10% of the difference between the initial aquifertemperatuur and high-temperature regime .
Machine	77.67	0.93	The maximum permitted cooling of the production water after 50 years is 10% of the difference between the initial aquifer temperature and the return temperature .



Sentence 9	BLEU	Length ratio	Text
Source	-	-	Op basis van een iteratieve methode is tevens een disconteringsvoet vóór belastingen bepaald.
Human	100.00	1.00	A discount rate before tax is also determined , on the basis of an iterative method .
Machine	30.45	0.88	On the basis of an iterative process has also been an increase in loads .
Machine	65.31	1.00	A discount rate before tax has also been determined on the basis of an iterative method .



Sentence 12	BLEU	Length ratio	Text
Source	-	-	In de tweede vergadering van 2017 heeft de auditcommissie aandacht besteed aan de volgende onderwerpen: het functioneren van de externe accountant inclusief de uitvoering van de joint venture audits door de externe accountant, de rendementseisen van EBN voor haar activiteiten en de post investment review.
Human	100.00	1.00	At the second meeting of 2017 , the audit committee paid attention to the following topics : the performance of the external auditor including the execution of the joint venture audits by the external auditor , the return requirements of EBN for its activities and the post investment review .
Machine	26.19	0.86	In 2017 , the second meeting of the following issues : the functioning of the external auditor , including the implementation of joint audits carried out by the external auditor , the efficiency of its activities for the ENP and investment flows .
Machine	45.35	0.90	At its second meeting in 2017 , the audit committee addressed the following issues : the functioning of the external auditor including the implementation of the joint venture audits by the external auditor , EBN's efficiency requirements for its activities and the post-investment review .



Sentence 19	BLEU	Length ratio	Text
Source	-	-	Uitgaven voor onderstaande activiteiten worden geactiveerd als onderdeel van de exploratie- en evaluatieactiva in aanleg: acquisitie van exploratielicenties, exploratieboringen inclusief test, sampling (monstername) en activiteiten in relatie tot evaluatie van de technische en commerciële mogelijkheid om koolwaterstoffen te winnen.
Human	100.00	1.00	Expenditure for the following activities are capitalised as part of the exploration and evaluation assets under construction : acquisition of exploration licences , exploration drilling including test , sampling and activities in relation to evaluation of the technical and commercial possibility of extracting hydrocarbons .
Machine	28.78	0.96	The following activities are capitalized costs as part of construction in exploration and evaluation : acquisition of exploratielicenties , exploration drilling , test and sampling (sampling) and activities related to the evaluation of the technical and commercial feasibility of hydrocarbons .
Machine	69.25	1.07	Expenditure on the following activities is activated as part of the exploration and evaluation assets under construction : acquisition of exploration licences , exploration drilling including test , sampling (sampling) and activities in relation to evaluation of the technical and commercial potential to extract hydrocarbons .

REFERENCES

Bertoldi, N., Haddow, B., & Fouet, J.-B. (2009). Improved Minimum Error Rate Training in Moses. The Prague Bulletin of Mathematical Linguistics, 91(1), 7–16.

Heafield, K. (2011). KenLM: Faster and Smaller Language Model Queries. In Proceedings of the Sixth Workshop on Statistical Machine Translation (pp. 187–197). Association for Computational Linguistics.

Junczys-Dowmunt, M., Grundkiewicz, R., Grundkiewicz, T., Hoang, H., Heafield, K., Neckermann, T., Martins, A. (2018). Marian: Fast neural machine translation in C++. arXiv preprint arXiv:1804.00344.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions (pp. 177–180). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=1557769.1557821

Koehn, P., Och, F. J., & Marcu, D. (2003). Statistical Phrase-Based Translation. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1 (pp. 48--54). Association for Computational Linguistics.

Krause, B., Lu, L., Murray, I., and Renals, S. (2017). Multiplicative LSTM for sequence modelling. In 5th International Conference on Learning Representations, page 9, Toulon, France, feb.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th annual meeting on association for computational linguistics (pp. 311–318).

Sennrich, R., Haddow, B., & Birch, A. (2016). Neural Machine Translation of Rare Words with Subword Units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Vol. 1, pp. 1715-1725).