DataLit^{MT} – Teaching Data Literacy in the Context of Machine Translation Literacy

Janiça Hackenbuchner
Institute of Translation and
Multilingual Communication
TH Köln – University of
Applied Sciences Cologne, Germany
janica.hackenbuchner
@th-koeln.de

Ralph Krüger
Institute of Translation and
Multilingual Communication
TH Köln – University of
Applied Sciences Cologne, Germany
ralph.krueger@th-koeln.de

Abstract

This paper presents the DataLit^{MT} project conducted at TH Köln – University of Applied Sciences. The project develops learning resources for teaching data literacy in its translation-specific form of professional machine translation (MT) literacy to students of translation and specialised communication programmes at BA and MA levels. We discuss the need for data literacy teaching in a translation/specialised communication context, present the three theoretical pillars of the project (consisting of a Professional MT Literacy Framework, an MT-specific data literacy framework and a competence matrix derived from these frameworks) and give an overview of the learning resources developed as part of the project.

1 Introduction

In recent years, the professional translation industry has seen accelerating processes of digitalisation – in the form of powerful new artificial intelligence algorithms in the field of natural language processing (NLP) and beyond (most recently, the transformer neural network architecture by Vaswani et al., 2017) – and datafication – through accumulating large volumes of translation data for training translation-specific NLP applications such as neural machine translation (NMT) systems. This has led to a considerable increase in translation automation, mostly through the integration of NMT systems in translation production workflows (e.g., ELIS Research, 2022). Accordingly, an adequate degree of *machine translation literacy* (Bowker and Buitrago Ciro, 2019) is becoming more and more relevant for professional translators. In translation studies, the concept of MT literacy has been applied both to professional translators working in MT-assisted translation production networks, and to layperson audiences (Kenny, 2022), who can use powerful MT technology as cloud-based "everyware" (Cronin, 2010) in their daily lives and should thus have a basic understanding of this technology. In order to delineate layperson MT literacy from MT literacy geared towards professional translators, Krüger and Hackenbuchner (2022) define *professional MT literacy* as "the full range of MT-related competences professional translators (and other language professionals) may require in order to participate successfully in the various phases of the MT-assisted professional translation process". The concept of professional MT literacy was then further expanded in a *Professional MT Literacy Framework*, which we discuss in more detail in section 3.1.

Parallel to the increasing relevance of MT literacy for professional and layperson audiences, adequate *data literacy* is also becoming more and more important, both at the overall level of modern datafied societies and at the level of specific professional fields (such as translation), where management and production processes have also become increasingly datafied in recent years (Misra, 2021). Against this background, data literacy is seen as a key prerequisite for enabling people to "navigate the complexity of modern data ecosystems" (ibid.). Ridsdale et al. (2015) define data literacy in a rather general way as "the ability to collect, manage, evaluate, and apply data, in a critical manner". Other authors attempt more context-bound conceptualisations of data literacy, situating it, for example, within the process of knowledge creation (Schüller, 2020) or within the overall data lifecycle (Misra, 2021). A common thread running through these different approaches is that they not only highlight the technical dimension of this concept but also stress that adequate data literacy involves critical awareness of the impact of using data in various application contexts. There is an immediate link between (professional) MT literacy and data literacy, since modern corpus-based MT systems have to be trained on large volumes of high-quality translation data in order to produce high-quality translations (Koehn, 2020). From this perspective, data literacy can be seen as an important building block of (professional) MT literacy. We expand upon the interface between MT literacy and data literacy in section 3.3.

2 DataLit$^{\mathrm{MT}}$

The DataLit$^{\mathrm{MT}}$ project is based at the Institute of Translation and Multilingual Communication at TH Köln – University of Applied Sciences, Cologne, Germany. DataLit$^{\mathrm{MT}}$ starts from the following premise: Although data literacy education is becoming increasingly relevant for students of translation and specialised communication programmes due to the increasing datafication of the respective professional fields and the growing societal relevance of data literacy as discussed above, there is less of a ‘natural fit’ between data literacy and these study programmes – which have traditionally focused more on linguistic, communicative and (inter)cultural aspects – than, for example, between data literacy and more technology-focused programmes such as computer science, data science or computational linguistics. For data literacy education in a translation/specialised communication context to be feasible, we must therefore first establish suitable points of contact between data literacy and topics which are more central to translation and specialised communication. DataLit$^{\mathrm{MT}}$ assumes that machine translation is well-suited to serve as a conceptual bridge between data literacy on the one hand and translation and specialised communication on the other.

In the preparatory stage of DataLit$^{\mathrm{MT}}$, we conducted a small survey among the students of the BA and MA programmes at the Institute of Translation and Multilingual Communication at TH Köln. As part of the survey, we asked students for their free associations with the term “data literacy”. Figure 1 illustrates the answers given (n=24).

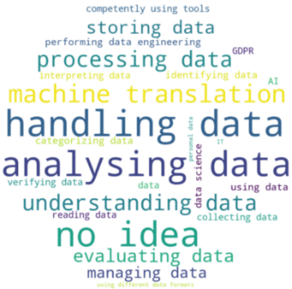


Figure 1: Students’ free associations with the term *data literacy*.

As can be seen, the most common associations are “handling data” (n=5), “analysing data” (n=5), and “no idea” (n=5) followed by “machine translation” (n=4), “processing data” (n=3) and “understanding data” (n=3). We interpret these results as follows. Despite its high societal and professional relevance as discussed in section 1, the term *data literacy* does not seem to be universally known to students (“no idea”=5). Several students already link data literacy to machine translation, which highlights the interface between the two concepts to be exploited by DataLit$^{\mathrm{MT}}$. Finally and perhaps not surprisingly, most of the students’ associations are related to ‘hands-on’ steps of working with data (handling/analysing/processing data). More abstract and higher-level aspects of data literacy, such as critical thinking and data ethics, as well as strategic aspects, such as data requirement analyses or data-driven decision making (see the discussion of our data literacy framework in section 3.2), seem less immediately obvious to students. Although based on a small and non-randomised sample, these results can be taken to indicate both the general need of comprehensive data literacy education and the feasibility of our basic didactic idea of teaching data literacy in an application context which will already be familiar to students of translation/specialised communication programmes.

Against this backdrop, the DataLit$^{\mathrm{MT}}$ project develops learning resources for teaching data literacy in its translation-specific form of professional MT literacy aimed at translation and specialised communication programmes at BA and MA levels. The learning resources are made publicly available on the DataLit$^{\mathrm{MT}}$ website[1] and GitHub repository[2]. The project also comprises a YouTube channel with tutorial videos for individual learning resources[3].

3 Theoretical Pillars of DataLit$^{\mathrm{MT}}$

In the preparatory stage of DataLit$^{\mathrm{MT}}$, we developed a *Professional MT Literacy Framework* and an MT-specific data literacy framework (*DataLit^{MT} Framework*) (Krüger, 2022a; Krüger and Hackenbuchner, 2022) in order to provide internal structure to the two frames of reference relevant to the project and to identify points of contact between them. Based on the interface between the two concepts, we then developed a competence matrix (*DataLit^{MT} Competence Matrix*) (Krüger and Hackenbuchner, forthcoming) comprising MT-specific competence descriptors for the individual (sub)dimensions of the DataLit$^{\mathrm{MT}}$ Framework.

3.1 Professional MT Literacy Framework

The Professional MT Literacy Framework depicted in figure 2 consists of five dimensions, which are divided further into individual subdimensions. The framework attempts to capture a comprehensive set of MT-related competences relevant to translators and other language professionals working in professional MT-assisted translation production networks. We discuss this framework in a concise form here. A more exhaustive discussion can be found in (Krüger, 2022a) and (Krüger and Hackenbuchner, 2022).

Technical MT literacy, as the name implies, covers the technical side of (mostly neural) machine translation. This is probably the dimension of professional MT literacy which is the most controversial in a translation/specialised communication context, since the technical side of MT is usually considered to be the area of

[1] https://itmk.github.io/The-DataLitMT-Project/

[2] https://github.com/ITMK/DataLitMT

[3] https://www.youtube.com/channel/UCnLNzT55g2X0_7emt45e0xg/

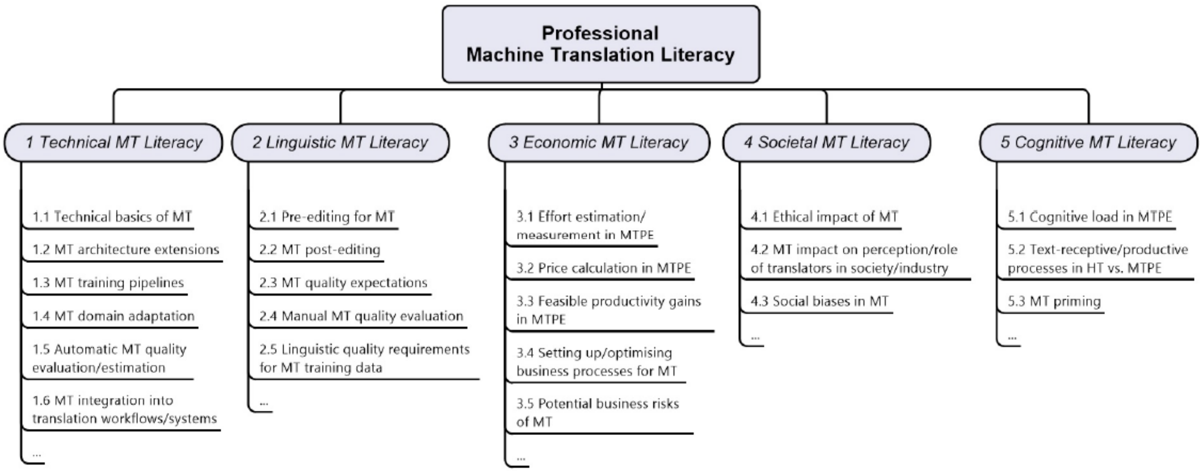

Professional Machine Translation Literacy
1 Technical MT Literacy
1.1 Technical basics of MT
1.2 MT architecture extensions
1.3 MT training pipelines
1.4 MT domain adaptation
1.5 Automatic MT quality evaluation/estimation
1.6 MT integration into translation workflows/systems
...
2 Linguistic MT Literacy
2.1 Pre-editing for MT
2.2 MT post-editing
2.3 MT quality expectations
2.4 Manual MT quality evaluation
2.5 Linguistic quality requirements for MT training data
...
3 Economic MT Literacy
3.1 Effort estimation/measurement in MTPE
3.2 Price calculation in MTPE
3.3 Feasible productivity gains in MTPE
3.4 Setting up/optimising business processes for MT
3.5 Potential business risks of MT
...
4 Societal MT Literacy
4.1 Ethical impact of MT
4.2 MT impact on perception/role of translators in society/industry
4.3 Social biases in MT
...
5 Cognitive MT Literacy
5.1 Cognitive load in MTPE
5.2 Text-receptive/productive processes in HT vs. MTPE
5.3 MT priming
...

Figure 2: Professional MT Literacy Framework (Krüger and Hackenbuchner, 2022).

expertise of computer scientists or computational linguists. However, an adequate degree of technical MT literacy in professional translators may further agendas of translator empowerment by demystifying the operating principle of this powerful translation technology and enabling translators to better intervene in MT-assisted translation workflows (Moorkens, 2018; Kenny, 2019).

Linguistic MT Literacy covers those aspects of MT literacy which have traditionally been associated with translation. It should be pointed out that post-editing is included as just one subdimension of linguistic MT literacy, which includes other aspects such as manual MT quality evaluation, an awareness of feasible MT quality in different application scenarios (including the ability to communicate feasible MT quality to other relevant actors in translation production networks), etc. Integrating post-editing into an expanded linguistic MT literacy which is in turn only one of five dimensions of overall professional MT literacy serves to illustrate the MT-induced "upskilling of translators" (Olohan, 2017), who, in current and future MT-assisted translation workflows, will have to master an expanded set of MT-related competences going beyond traditional post-editing.

Economic MT Literacy covers the management side of MT-assisted translation projects and involves aspects of translation process analysis and organisation with a view to integrating MT into these projects. This subdimension of professional MT literacy is therefore particularly relevant to translation project managers but may also contribute to translators' MT-related "consulting competence" (Nitzke et al., 2019) vis-à-vis relevant actors in translation production networks.

Societal MT Literacy covers competences associated with the overall societal and translation industry-internal impact of NMT, including its ethical dimension (Moorkens, 2022). Adequate societal MT literacy enables translators to engage in overall societal discourses about the status and role of professional translators in the context of powerful MT technologies, but also in translation industry-internal discourses about the intellectual added value of human/expert-in-the-loop translation production workflows, particular in the context of recent claims concerning superhuman MT performance (e.g., Popel, 2020).

Cognitive MT Literacy is concerned with awareness of the cognitive impact of NMT on translators working in MT-assisted translation production workflows. Cognitive MT literacy may, in particular, serve to develop translators' metacognitive monitoring competence (Göpferich, 2008), which may contribute, e.g., to an awareness of potential MT-induced priming effects (Carl and Schaeffer, 2017).

3.2 DataLit^{MT} Framework

The DataLit^{MT} Framework depicted in figure 3 is derived from the data literacy frameworks proposed by Ridsdale et al. (2015), Schüller (2020) and Misra (2021) and adjusted slightly to fit the overall data lifecycle in MT-assisted translation scenarios. Similar to the Professional MT Literacy Framework, the DataLit^{MT} Framework comprises five dimensions, each consisting of several subdimensions. Again, we discuss this framework in a concise way below and refer to the more detailed discussion in Krüger (2022a) and Krüger and Hackenbuchner (2022).

The first dimension is the *data context*, which is primarily theoretical in nature. It covers general knowledge and a critical awareness of how to use and apply data and potential ethical implications of working with data, as well as the ability to identify and specify individual tasks within a workflow that could be supported or optimised with the help of data.

Data planning serves as a bridge between the theoretical data context and the more practical sections of the framework. Data planning involves performing a data requirement analysis in order to identify which specific data is required to support/optimise individual tasks, developing a data strategy which guides the ac-

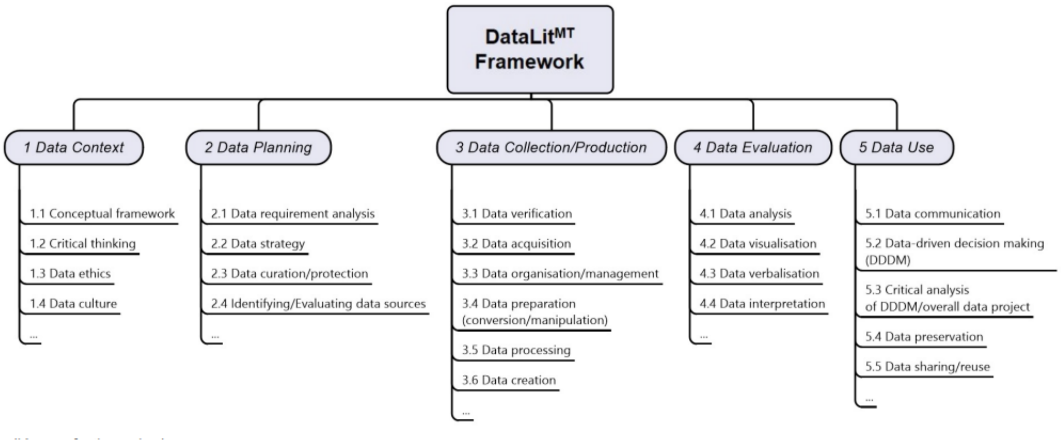


Figure 3: DataLit$^{\mathrm{MT}}$ Framework (Krüger and Hackenbuchner, 2022).

quisition of this data, practical aspects of data curation and protection, and identifying and evaluating potential data sources.

Data collection and production is the first 'hands-on' dimension of the framework. It basically describes the process of collecting relevant data as identified in the data planning step, applying tools to work with this data (organisation, metadata creation, conversion, cleaning and filtering, etc.), and using this data to create new data (e.g., collecting and preparing data to train an MT engine and using this MT engine to produce new translation data). This stage is critical for any practical data project (such as training NMT engines) and will be further expanded upon in our discussion of the DataLit$^{\mathrm{MT}}$ Competence Matrix in section 3.4.

Data evaluation is another hands-on dimension, focused on working with the data collected and/or produced in the previous step of a data project. The focus here lies on applying methods and tools for data analysis and evaluation, creating graphical or textual representations of data analysis results and understanding these results by identifying key insights.

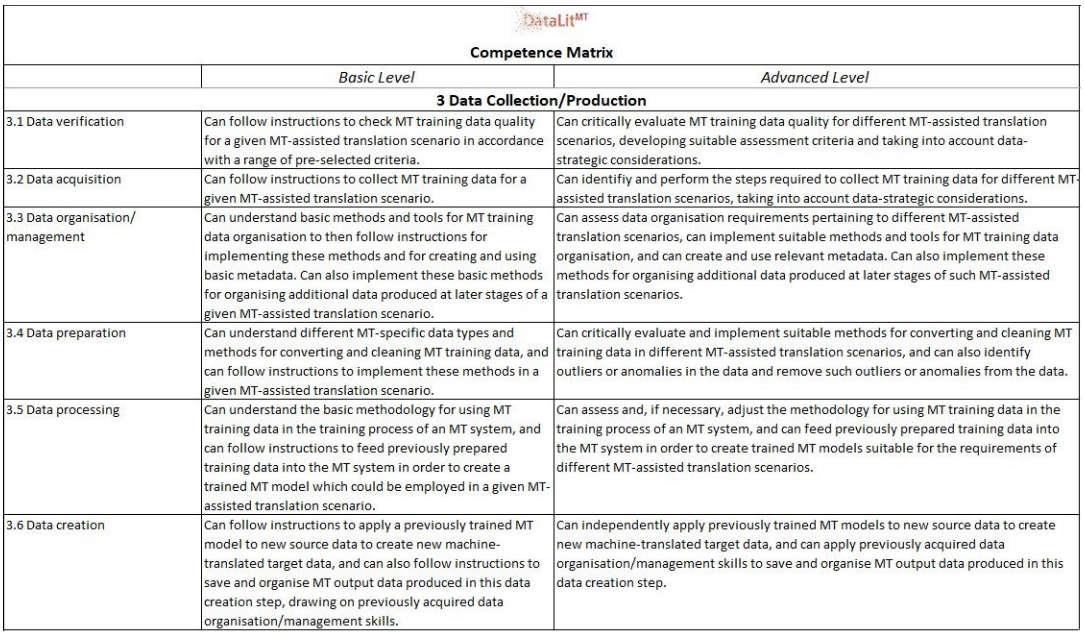
The *Data use* dimension completes a typical data project. The subdimensions of data use focus on communicating data analysis results to relevant stakeholders within an organisation, making data-driven decisions informed by the analysis results, critically evaluating the impact of these decisions and the overall data project, and taking practical measures such as preserving data and sharing them for future reuse.

In any data project in the context of machine translation, some if not all of these (sub)dimensions of the DataLit$^{\mathrm{MT}}$ Framework will likely play an important role, as discussed in the following section.

3.3 Interface between the Professional MT Literacy Framework and the DataLit$^{\mathrm{MT}}$ Framework

In order to lay the groundwork for the competence matrix guiding the development of specific learning resources, we first established relevant points of contact between the Professional MT Literacy Framework and the DataLit$^{\mathrm{MT}}$ Framework. For example, the *data context* subdimensions of critical thinking and data ethics can be readily linked to *societal MT literacy*, which is concerned with the wider ethical and societal impact of MT and requires critical thinking and ethical awareness, as stipulated by the data context. The *data planning* subdimensions can be linked in particular to *technical MT literacy* (and here specifically to MT training pipelines and MT domain adaptation) and to *linguistic MT literacy* (and here specifically to linguistic quality requirements for MT training data), since aspects such as volume, domain, language combination and quality of MT training data will be established in the data planning phase and will in turn guide individual data planning steps such as identifying suitable data sources. *Data collection and production* links primarily to MT training pipelines as part of *technical MT literacy*, with data acquisition, organisation, preparation and processing describing the central steps of such a training pipeline. *Data evaluation* can also be linked to *technical MT literacy* (and here particularly to automatic MT quality evaluation/estimation) and to *linguistic MT literacy* (particularly manual MT quality evaluation), since data evaluation in an MT context will usually be concerned with data produced by a previously trained MT engine. *Data use*, lastly, can be linked primarily to *economic MT literacy*, which is concerned with the management/business side of MT-assisted translation projects, such as effort estimation/measurement in machine translation post-editing (MTPE), price calculation in MTPE, setting up or optimising business processes with a view to MT integration, etc. Ideally, these decisions are data driven and informed by the results of respective data analyses (e.g. results of automatic/manual MT quality evaluation or MTPE productivity measurements).

These are merely a few examples of how the DataLit$^{\mathrm{MT}}$ Framework can be mapped onto the Professional MT Literacy Framework, illustrating the need for



DataLit$^{\text{MT}}$

Competence Matrix

	Basic Level	*Advanced Level*
3 Data Collection/Production		
3.1 Data verification	Can follow instructions to check MT training data quality for a given MT-assisted translation scenario in accordance with a range of pre-selected criteria.	Can critically evaluate MT training data quality for different MT-assisted translation scenarios, developing suitable assessment criteria and taking into account data-strategic considerations.
3.2 Data acquisition	Can follow instructions to collect MT training data for a given MT-assisted translation scenario.	Can identifiy and perform the steps required to collect MT training data for different MT-assisted translation scenarios, taking into account data-strategic considerations.
3.3 Data organisation/ management	Can understand basic methods and tools for MT training data organisation to then follow instructions for implementing these methods and for creating and using basic metadata. Can also implement these basic methods for organising additional data produced at later stages of a given MT-assisted translation scenario.	Can assess data organisation requirements pertaining to different MT-assisted translation scenarios, can implement suitable methods and tools for MT training data organisation, and can create and use relevant metadata. Can also implement these methods for organising additional data produced at later stages of such MT-assisted translation scenarios.
3.4 Data preparation	Can understand different MT-specific data types and methods for converting and cleaning MT training data, and can follow instructions to implement these methods in a given MT-assisted translation scenario.	Can critically evaluate and implement suitable methods for converting and cleaning MT training data in different MT-assisted translation scenarios, and can also identify outliers or anomalies in the data and remove such outliers or anomalies from the data.
3.5 Data processing	Can understand the basic methodology for using MT training data in the training process of an MT system, and can follow instructions to feed previously prepared training data into the MT system in order to create a trained MT model which could be employed in a given MT-assisted translation scenario.	Can assess and, if necessary, adjust the methodology for using MT training data in the training process of an MT system, and can feed previously prepared training data into the MT system in order to create trained MT models suitable for the requirements of different MT-assisted translation scenarios.
3.6 Data creation	Can follow instructions to apply a previously trained MT model to new source data to create new machine-translated target data, and can also follow instructions to save and organise MT output data produced in this data creation step, drawing on previously acquired data organisation/management skills.	Can independently apply previously trained MT models to new source data to create new machine-translated target data, and can apply previously acquired data organisation/management skills to save and organise MT output data produced in this data creation step.

Figure 4: *Data Collection/Production* section of the DataLit$^{\text{MT}}$ Competence Matrix.

data literacy competences in the context of MT-assisted translation workflows.

3.4 DataLit$^{\text{MT}}$ Competence Matrix

Based on the interface between the Professional MT Literacy Framework and the DataLit$^{\text{MT}}$ Framework, as discussed in the previous section, we developed a competence matrix of MT-specific data literacy competence descriptors. Here, the subdimensions of the DataLit$^{\text{MT}}$ Framework provide the descriptive categories of the individual matrix sections and the Professional MT Literacy Framework provides the application contexts to which the individual competence descriptors refer. The competence matrix was inspired by PACTE's work on establishing competence levels in translation competence acquisition (PACTE Group, 2018) and describes MT-specific data literacy competences at Basic and Advanced Levels. The Basic Level descriptors refer to lower-level cognitive tasks such as memorising and recalling facts and demonstrating a basic understanding of specific concepts, and the Advanced Level descriptors address higher-level cognitive tasks such as applying concepts to new situations, analysing complex contexts into individual components or relating and integrating information from different sources. Accordingly, Basic Level competence descriptors generally require students to "follow instructions" or to "understand" certain concepts, whereas Advanced Level requirements are generally to "assess", "critically evaluate", "implement" or "independently apply" certain concepts. Specifically, the Basic Level addresses less complex knowledge of data literacy and MT literacy and requires a lower degree of IT skills (particularly programming skills) for understanding and following the concepts discussed in the respective learning resources. The Advanced Level, on the other hand, aims at more complex knowledge and skills related to data literacy and MT literacy and presupposes a higher degree of IT competence in order to comprehend and follow the concepts discussed in the respective learning resources.

Figure 4 illustrates the section of the DataLit$^{\text{MT}}$ Competence Matrix comprising competence descriptors for data collection/production. The full matrix provides a detailed description of the MT-oriented knowledge and skills required for each data literacy subdimension at both Basic and Advanced Levels and is described in more detail in Krüger and Hackenbuchner (forthcoming). The full matrix is also available on the DataLit$^{\text{MT}}$ project website[4].

As discussed in sections 3.2 and 3.3, the data collection/production dimension of MT-related data literacy basically describes the individual steps of an MT training pipeline, from checking the adequacy of a particular set of MT training data for an MT-assisted translation scenario, to collecting this training data, organising the data (e.g., using adequate folder structures and/or metadata) preparing the data for MT training (e.g., by converting and cleaning them), processing the data in the actual training stage in order to train an MT model and finally creating new translation data (e.g., by translating a test set for evaluating the quality of the final MT model), which again may have to be organised/managed in a specific way. The wording of the individual competence descriptors at Basic and Advanced Levels reflects the distinction between lower and higher-level cognitive tasks as discussed previously. Since this section of the competence

[4] `https://itmk.github.io/The-DataLitMT-Project/matrix/`

Learning Resource Topic	Level	Format
Conceptual data overview & resources	Basic Level	Paper
Data Ethics and MT	Basic Level	Paper
Social Bias in MT	Basic Level	Paper
		Tutorial Video
	Advanced Level	Paper
MT Training Data Preparation	Basic Level	Jupyter Notebook
		Tutorial Video
	Advanced Level	Jupyter Notebook
		Tutorial Video
Training an NMT Model	Advanced Level	Jupyter Notebook
		Tutorial Video
Terminology Integration into MT Models	Basic Level	Paper
		Tutorial Video
	Advanced Level	Paper (as above)
		Jupyter Notebook
		Tutorial Video
Automatic MT Quality Evaluation	Basic Level	Jupyter Notebook
		Tutorial Video
	Advanced Level	Jupyter Notebook
		Tutorial Video
Companion Notebooks:		
String Matching-based Metrics	Basic Level	Jupyter Notebook
Embedding-based Metrics	Basic Level	Jupyter Notebook
Evaluation at Document Level	Advanced Level	Jupyter Notebook
		Tutorial Video
Pre- and Post-Editing	Basic Level	Paper
Machine Translationese & Post-Editese	Basic Level	Paper
		Tutorial Video
	Advanced Level	Paper (as above)
		Jupyter Notebook
		Tutorial Video

Table 1: Overview of DataLit$^{\text{MT}}$ learning resources as of April 2023.

matrix and the following section concerned with data evaluation may require information-technological skills which exceed the skills that, on average, can be expected from students of translation or specialised communication programmes, the learning resources developed for the respective competence descriptors require an adequate didactic scaffolding in order to bridge this skill gap. In section 4.2, we discuss an example of one of our learning resources and illustrate how students can use this resource to perform the technical steps involved in MT-specific data collection/production without any advanced IT skills.

4 DataLit$^{\text{MT}}$ Learning Resources

In this section, we discuss the open educational learning resources that we developed based on the competence matrix illustrated in the previous section. The resources are not set up as comprehensive course syllabi, but are intended to complement translation technology/NLP courses or courses with other foci (e.g. on ethical aspects of the professional translation industry) in translation and/or specialised communication programmes. The resources can be used as extensive lecture materials in the classroom (or for self-study purposes) to theoretically explain and practically exemplify various aspects of MT-specific data literacy. For example, at TH Köln, several of the DataLit$^{\text{MT}}$ learning resources will complement introductory courses on translation technology in our BA in Multilingual Communication programme and advanced translation technology and MT-specific courses in our MA in Specialised Translation and MA in Terminology and Translation Technology programmes. All learning resources are written in English to expand the international reach of this project. They are published under a Creative Commons BY-SA-4.0 license and made publicly available on the DataLit$^{\text{MT}}$ website[5].

Section 4.1 provides an overview of the full range of learning resources developed for DataLit$^{\text{MT}}$ and section 4.2 zooms in on one particular learning resource concerned with MT-specific data collection/production.

4.1 Overview of DataLit$^{\text{MT}}$ Learning Resources

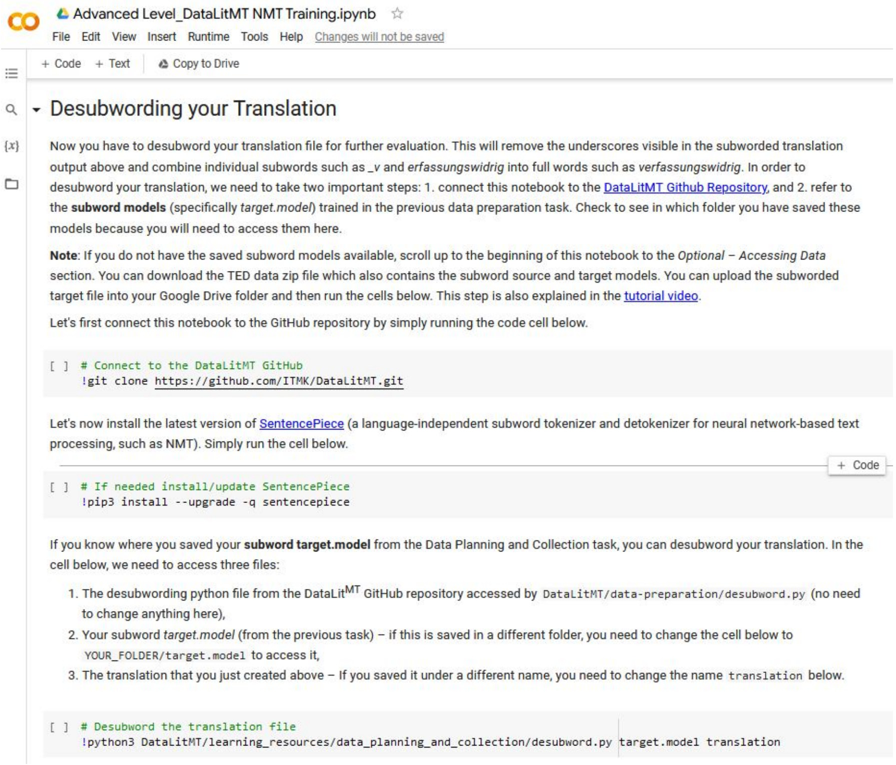
Table 1 presents an overview of the DataLit$^{\text{MT}}$ learning resources available as of April 2023.

Depending on the topics covered, the learning resources are available in different formats, i.e., as papers, web-based Jupyter notebooks[6] hosted in a Google Colab environment[7], or as tutorial videos. Several learning resources combine these different formats and are therefore available as a combination of paper + video, notebook + video or paper + notebook + video. Where

[5] https://itmk.github.io/The-DataLitMT-Project/

[6] https://jupyter.org/

[7] https://colab.research.google.com/


Advanced Level_DataLitMT NMT Training.ipynb
File
Edit
View
Insert
Runtime
Tools
Help
Changes will not be saved
+ Code
+ Text
Copy to Drive
+ Code

Desubwording your Translation

Now you have to desubword your translation file for further evaluation. This will remove the underscores visible in the subworded translation output above and combine individual subwords such as *_v* and *erfassungswidrig* into full words such as *verfassungswidrig*. In order to desubword your translation, we need to take two important steps: 1. connect this notebook to the DataLitMT Github Repository, and 2. refer to the **subword models** (specifically *target.model*) trained in the previous data preparation task. Check to see in which folder you have saved these models because you will need to access them here.

Note: If you do not have the saved subword models available, scroll up to the beginning of this notebook to the *Optional – Accessing Data* section. You can download the TED data zip file which also contains the subword source and target models. You can upload the subworded target file into your Google Drive folder and then run the cells below. This step is also explained in the tutorial video.

Let's first connect this notebook to the GitHub repository by simply running the code cell below.

```
[ ]  # Connect to the DataLitMT GitHub
     !git clone https://github.com/ITMK/DataLitMT.git
```

Let's now install the latest version of SentencePiece (a language-independent subword tokenizer and detokenizer for neural network-based text processing, such as NMT). Simply run the cell below.

```
[ ]  # If needed install/update SentencePiece
     !pip3 install --upgrade -q sentencepiece
```

If you know where you saved your **subword target.model** from the Data Planning and Collection task, you can desubword your translation. In the cell below, we need to access three files:

1. The desubwording python file from the DataLit$^{\text{MT}}$ GitHub repository accessed by `DataLitMT/data-preparation/desubword.py` (no need to change anything here),
2. Your subword *target.model* (from the previous task) – if this is saved in a different folder, you need to change the cell below to `YOUR_FOLDER/target.model` to access it,
3. The translation that you just created above – If you saved it under a different name, you need to change the name `translation` below.

```
[ ]  # Desubword the translation file
     !python3 DataLitMT/learning_resources/data_planning_and_collection/desubword.py target.model translation
```

Figure 5: Example section of a Jupyter notebook on NMT model training.

data files are required to work through a learning resource (e.g., in the case of Machine Translationese and Post-Editese), these files are made available for download in the GitHub repository[8]. If specific libraries are required to work through individual notebooks (such as the Natural Language Toolkit[9], SpaCy[10] or SentencePiece[11]), the sources of these libraries are linked in the respective notebook and predefined code cells can be run to automatically install them in the notebook environment. To exemplify the structure in which the learning resources are presented: The learning resources section of the DataLit$^{\text{MT}}$ website[12]is structured according to the individual (sub)dimensions of the DataLit$^{\text{MT}}$ Framework. From there, we link to the corresponding folder of the GitHub repository, where all materials for that learning resource are made available. For Jupyter notebooks, we also link directly to the Colab implementation of these notebooks from the DataLit$^{\text{MT}}$ website so that users can start working with these notebooks directly in a Colab environment. The website, and the respective notebooks, also links directly to the YouTube tutorial videos for individual learning resources.

4.2 Example of a Jupyter Notebook-Based DataLit$^{\text{MT}}$ Learning Resource

Figure 5 depicts an example section of a Colab-hosted Jupyter notebook on training an NMT model from scratch based on the OpenNMT-py toolkit (Klein et al., 2017). This resource is concerned with the subdimensions of data processing and data creation of the DataLit$^{\text{MT}}$ Competence Matrix. The notebook covers all steps from accessing NMT training data (for example, those prepared in the learning resource on NMT training data preparation), defining the parameters of the model to be trained, training the actual model, and then using this model to translate the test dataset.

As discussed above, hands-on data steps such as preparing training data or training NMT models are quite technical in nature and require an adequate degree of didactic scaffolding if these steps are to be performed by users with low to moderate information-technological skills. Therefore, we implemented these workflows using Jupyter notebooks, which have recently been proposed as suitable didactic instruments for translation technology teaching to non-technical translation audiences (Krüger, 2022b). The notebook section depicted in figure 5 is concerned with desubwording the translated test set for further evaluation. The documentation section in the upper half of the figure explains the individual steps that are necessary for desubwording the translation. The following two code cells connect the notebook to the Google Drive folder where the required subword models are stored and install the SentencePiece subword tokenizer. The follow-

[8] https://github.com/ITMK/DataLitMT

[9] https://www.nltk.org

[10] https://spacy.io/

[11] https://github.com/google/sentencepiece

[12] https://itmk.github.io/The-DataLitMT-Project/resources/

ing documentation section then explains the structure of the third code cell at the bottom of the figure, which accesses an external python script for (de)subwording, the target subword model created in the previous data preparation resource and the translated test dataset to be desubworded. The first documentation section also links to the tutorial video for this resource, in which users are guided explicitly through the individual steps of the NMT model training notebook. The Python code in the notebook is set up in such a way that only a minimum of user intervention is required (i.e., most of the code cells can be simply run by users ‘as is’). Wherever code needs to be changed (e.g., to refer to individual folders or files), this is explained in detail both in the corresponding documentation sections and in the tutorial video. This extensive didactic scaffolding supports non-technical users in working through the technical steps of an MT workflow which would usually require an adequate degree of programming skills or which would have to be implemented in a graphical user interface that non-technical users are familiar with. Further technical MT workflow aspects covered by Jupyter notebook-based learning resources developed by DataLit$^{\mathrm{MT}}$ are, in particular, training data preparation and calculating a range of string matching- and embedding-based MT quality metrics (see table 1).

5 Conclusion & Outlook

This paper presented the DataLit$^{\mathrm{MT}}$ project, which develops learning resources for teaching data literacy in its translation-specific form of professional MT literacy to students of translation and specialised communication programmes at BA and MA levels. We hope that these resources help students develop an adequate degree of data literacy *cum* MT literacy both for their later professional careers in the translation/specialised communication sector or beyond and for their role as citizens in modern digitalised and datafied societies. Since the project was completed only recently (February 2023), we do not yet have any data on the didactic effectiveness of the learning resources in actual teaching scenarios. We intend to investigate this in a follow-up study at TH Köln. In the future, we also aim to expand our work on transversal digital literacies relevant to the fields of translation/specialised communication and at societal level to include *artificial intelligence literacy*, which Long and Magerko (2020) define as “a set of competencies that enables individuals to critically evaluate AI technologies; communicate and collaborate effectively with AI; and use AI as a tool online, at home, and in the workplace”. Professional translation has been a forerunner in AI-based automation in recent years, mostly due to the implementation of NMT in production systems since 2016. More recently, powerful large language models based on Vaswani et al.'s transformer architecture have emerged, perhaps most notably in the form of ChatGPT[13]. The powerful generative capabilities of ChatGPT and related models have extended AI-based automation far beyond its previous scope of application, making an adequate degree of AI literacy of the citizens whose societies are about to be transformed by AI a pressing matter. Since modern AI technologies such as NMT or the GPT language models rely on large volumes of high-quality training data, there is an immediate link between data literacy, MT literacy and AI literacy. It can therefore be assumed that a solid data literacy/MT literacy education as discussed in this paper may act as a stepping stone for a more extensive AI literacy education.

6 Acknowledgements

The DataLit$^{\mathrm{MT}}$ project reported on in this paper was funded by the Stifterverband and the Ministry of Culture and Science of North Rhine-Westphalia, Germany, under the funding line Curriculum 4.0 (`https://www.stifterverband.org/curriculum-4-0`).

7

References

Bowker, Lynne, and Jairo Buitrago Ciro. 2019. *Machine Translation and Global Research: Towards Improved Machine Translation Literacy in the Scholarly Community*. Bingley: Emerald Publishing.

Carl, Michael, and Moritz Schaeffer. 2017. Sketch of a Noisy Channel Model for the Translation Process. In *Empirical Modelling of Translation and Interpreting*, 71–116. Berlin: Language Science Press.

Krüger, Ralph. 2022a. Integrating Professional Machine Translation Literacy and Data Literacy. *Lebende Sprachen*, 67(2):247–282.

Krüger, Ralph. 2022b. Using Jupyter Notebooks as Didactic Instruments in Translation Technology Teaching. *The Interpreter and Translator Trainer*, 16(4):503-523.

Krüger, Ralph, and Janiça Hackenbuchner. 2022. Outline of a Didactic Framework for Combined Data Literacy and Machine Translation Literacy Teaching. *Current Trends in Translation Teaching and Learning E.*, 375–432.

Krüger, Ralph, and Janiça Hackenbuchner. Forthcoming. A Competence Matrix for Machine Translation-Oriented Data Literacy Teaching.

Cronin, Michael. 2010. The Translation Crowd. *Revista Tradumàtica* 8, 1–7.

Göpferich, Susanne. 2008. *Translationsprozessforschung. Stand – Methoden – Perspektiven*. Tübingen: Narr.

[13] `https://openai.com/blog/chatgpt`

ELIS Research 2022. European Language Industry Survey 2022. https://elis-survey.org/

Kenny, Dorothy. 2019. Machine Translation. In *The Routledge Handbook of Translation and Philosophy*, pages 428–445. London: Routledge.

Kenny, Dorothy. 2022. Introduction. In *Machine Translation for Everyone. Empowering Users in the Age of Artificial Intelligence*, pages v–viii. Berlin: Language Science Press.

Klein, Guillaume, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. *Proceedings of ACL 2017, System Demonstrations*, pages 67–72. Association for Computational Linguistics.

Koehn, Philipp. 2020. *Neural Machine Translation*. Cambridge: University Press.

Long, Duri, and Brian Magerko. 2020. What is AI Literacy? Competencies and Design Considerations. *CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems,* pages 1–16. New York: Association for Computing Machinery.

Misra, Archita. 2021. Advancing Data Literacy in the Post-Pandemic World. A Primer to Catalyse Dialogue and Action. *PARIS21.*

Moorkens, Joss. 2018. What to Expect from Neural Machine Translation: A Practical In-Class Translation Evaluation Exercise. *The Interpreter and Translator Trainer*, 12(4):375–387.

Moorkens, Joss. 2022. Ethics and Machine Translation. In *Machine Translation for Everyone. Empowering Users in the Age of Artificial Intelligence*, pages 121–140. Berlin: Language Science Press.

Nitzke, Jean, Silvia Hansen-Schirra, and Carmen Canfora. 2019. Risk Management and Post-Editing Competence. *Journal of Specialised Translation* 31, 239–259.

Olohan, Maeve. 2017. Technology, Translation and Society. *Target. International Journal of Translation Studies*, 29(2):264–283.

PACTE Group, Amparo Hurtado Albir (principal investigator), Anabel Galán-Mañas, Anna Kuznik, Christian Olalla-Soler, Patricia Rodríguez-Inés, and Lupe Romero. 2018. Competence Levels in Translation: Working Towards a European Framework. *The Interpreter and Translator Trainer*, 12(2):111–131.

Popel, Martin, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtský. 2020. Transforming Machine Translation: a Deep Learning System Reaches News Translation Quality Comparable to Human Professionals. *Nature Communications* 11, 1–15.

Ridsdale, Chantel, James Rothwell, Michael Smit, Hossam Ali-Hassan, Michael Bliemel, Dean Irvine, Daniel Kelley, Stan Matwin, and Bradley Wuetherick. 2015. *Strategies and Best Practices for Data Literacy Education. Knowledge Synthesis Report.* Dalhousie University.

Schüller, Katharina. 2020. *Future Skills: A Framework for Data Literacy*. Hochschulforum Digitalisierung.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Lilion Jones, Adrian N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010.