



SAPIENZA
UNIVERSITÀ DI ROMA

Sapienza University of Rome

Physics department
PhD in Physics

THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Two essays concerning complexity, language and machine learning

Thesis Advisor
Prof. Luciano Pietronero

Candidate
Marco Miccheli
1129043

XXXIV cycle

Preface

The research project presented in this thesis was born from a collaboration between Sapienza University of Rome, Consiglio Nazionale delle Ricerche (CNR) and Translated SRL, a professional translation service provider, i.e. a company that provides professional linguistic translations.

When I started the PhD program that led to this research project, I still did not know Translated and I was working with Professor Luciano Pietronero and his group, just master graduated with a thesis entitled *Volleyball as dynamical system*, in which I created a method to forecast some specific situations of the volleyball match, starting from an encoded transcription of the match through the application of dynamical systems modelling and Natural Language Processing techniques, overcoming the current state of the art in the field.

Since I did not have yet a project to work and nor a fellowship, I attended the school of AI, a two-months program hosted by Pi School, a company founded by Marco Trombetti, CEO of Translated Srl. In the meanwhile, I attended the courses for the PhD students at Sapienza, including the Bayesian Inference course held by Prof. Andrea Messina.

At the end of Pi School, I had earned a specific knowledge about Natural Language Processing and this led me to join Dr Andrea Tacchella, Dr Andrea Zaccaria and Prof. Luciano Pietronero in the research that led to the second part of this PhD thesis: *Relatedness estimate through Natural Language Processing*.

In September 2019 the collaboration with Translated Srl officially started, and so it was that I started the work related to the first part of the research project presented here.

That is why this PhD thesis is composed of two distinct essays: the first concerning the application of complex systems methods to Translation Quality Assessment (TQA) ([Miccheli et al., 2022](#)), and the second one related to economic complexity, which led to the paper *Relatedness in the Era of Machine Learning* ([Tacchella et al., 2021](#)), currently under revision.

Contents

List of Figures	iv
List of Tables	vi
Nomenclature	viii
Acronyms	ix
List of terms	xi
I Complexity of Translation Quality Assessment	1
1 Introduction	4
1.1 Complex systems	9
1.1.1 What is a complex system?	9
1.1.2 Some examples of complex systems and relative works . . .	10
1.1.3 Common methods for modelling complex systems	11
1.1.4 Linguistic translation: a complex system case	13
1.2 Translation Quality Assessment - state of the art	16
1.2.1 Evaluation methods	18
1.2.2 Improving Translation Quality Assessment reliability - re- lated works	26
1.2.3 Translated Srl	28
2 Dataset and methods	30
2.1 Translation Quality Assessment dataset	30
2.1.1 Text corpora	38
2.1.2 Errors Per Thousands consistency: chain of consecutive re- views	42
2.1.3 Glut of perfect translations	53
2.2 Modelling Translation Quality Assessment	55

2.2.1	Models	58
3	Results	66
3.1	Validation in unsupervised setting	66
3.1.1	Posterior predictive check	68
3.1.2	Skills retrieval	73
3.1.3	Quality assessment prediction	89
4	Conclusions and future works	94
II	Relatedness estimate through Natural Language Processing	100
5	Relatedness in Economic Complexity	102
6	Forecasting paths for Countries' development	107
6.1	Prediction results	111
6.1.1	Evaluation metrics	113
7	Prediction methods and Relatedness assessment	116
7.1	Density-based models	116
7.2	Tree-based models	118
7.3	Textual descriptions' similarity	119
7.3.1	Fuzzy Bag of Words	120
	Bibliography	125
A	Revision Guidelines	146
B	Additional figures	148
C	Quality assessment predictions - models application	150

List of Figures

1.1	Example of language network	13
1.2	Examples of different error-analysis frameworks, with different choices of categories (error types) and severities. Table take from Martínez (2014).	20
1.3	Error annotation-setting examples	24
2.1	Translated workflow	31
2.2	MateCat screenshots	34
2.3	1PR dataset	40
2.4	1PR data distributions	41
2.5	Numerical 2PR Dataset description	43
2.6	2PR dataset distributions	43
2.7	Errors' comparison for 1PR and 2PR - categories	45
2.8	Errors' comparison for 1PR and 2PR - severities	46
2.9	$\epsilon_2 - \epsilon_1$ per category and severity	50
2.10	Categories and severities correlation	51
2.11	Glut of perfect translations	54
2.12	Gaussian model - Bayesian network	61
2.13	Gaussian fit of the standardised log EPT values. Skewness=-0.01, Kurtosis=-0.03.	62
2.14	Hurdle model - bayesian network	65
3.1	Posterior predictive check: example of distributions	69
3.2	Posterior predictive check: ratio of zeros	71
3.3	Posterior predictive check: KL divergence	73
3.4	Translators assessment - CI per skill	76
3.5	Translators assessment - skills retrieval	80
3.6	Reviewers' assessment - CI per skill	82
3.7	Reviewers' assessment - skills retrieval	83
3.8	Reviewers assessment - skills retrieval (aggregated)	85
3.9	Linguists assessment - cross-evaluation as reviewers and translators	88

3.10	Double review experiment - absence of correlation between EPT . .	91
3.11	RMSE on EPT prediction	92
6.1	Comparison of prediction performance	111
A.1	Complete guidelines for revision process	147
B.1	$\epsilon_2 - \epsilon_1$ per category and severity (complete dataset)	149
C.1	EPT prediction examples	151
C.2	Posterior Predictive Check on test set	152
C.3	RMSE predictions for double-review experiment	153

List of Tables

2.1	Error categories - guidelines	36
2.2	Error severities - guidelines	37
2.3	$\epsilon_2 - \epsilon_1$ per category and severity quantiles	48
2.4	Perfect translations statistics	55
3.1	Dataset details for the percentage of linguists per role (translator/reviewer) and skill level	74
3.2	Correlation between mutual relations of reviewers	93
C.1	RMSE predictions for double-review experiment	153

Nomenclature

Translation Quality Assessment

β_r	Bias of reviewer r , page 56
ϵ_t	Error propensity of translator t , page 56
μ_β	Parameter μ of $p(\beta_r)$ [Hurdle model], page 65
μ_ϵ	Parameter μ of $p(\epsilon_t)$ [Hurdle model], page 65
π_β	Parameter π of $p(\beta_r)$ [Hurdle model], page 65
π_ϵ	Parameter π of $p(\epsilon_t)$ [Hurdle model], page 65
σ_β	Parameter σ of $p(\beta_r)$ [Hurdle model], page 65
σ_ϵ	Parameter σ of $p(\epsilon_t)$ [Hurdle model], page 65
d_j	Difficulty of the document (job) j to be translated, page 56
q_j	Real quality of the translated document (job) j , page 56

Acronyms

1PR 1st Pass Review

2PR 2nd Pass Review

BLEU BiLingual Evaluation Understudy

CAT Computer-Assisted Translation

CDF Cumulative Density Function

CI Confidence Interval

CSCS Complex Systems in Cognitive Science

EPT Error Per Thousands

HT Human Translation

IAA Inter Annotator Agreement

KDE Kernel Density Estimation

LL Language Lead

MAP Maximum A Posteriori

MCMC Markov Chain Monte Carlo

ML Machine Learning

MT Machine Translation

NUTS No U-Turn Sampler

PDF Probability Density Function

ppc Posterior Predictive Check

RCA Revealed Comparative Advantage

RMSE Root Mean Squared Error

TM Translation Memory

TQA Translation Quality Assessment

TS Translation Studies

WER Word Error Rate

WMT Workshop on Machine Translation

Special Terms

ground-truth the real data that represents the true value with which predictions can be compared

test set Subset of the dataset used to assess the performance of the model

training set Subset of the dataset that the model use to optimise its parameters

Part I

Complexity of Translation

Quality Assessment

Acknowledgements

This first essay concerning Translation Quality Assessment is published as a preprint paper ([Miccheli et al., 2022](#)) and it is going to be submitted in a journal through a peer review process. I want to thank all the co-authors of the paper, namely Andrej Leban, Andrea Tacchella, Andrea Zaccaria, Dario Mazzilli and Sébastien Bratières for their valuable contribution.

Abstract

In this first essay, we present our research project concerning the extraction of information from the Translation Quality Assessment (TQA) process, in which the quality of a translation conducted by a human translator from one language to another, is evaluated by another human translator. We take advantage of the dataset provided by the professional translation service provider Translated SRL, consisting of thousand of translations, produced by human translators and edited (with error annotations) by human reviewers. We deal with subjectivity that raises from the linguists involved in the process and we aim to understand which are the features able to catch translators' behaviour. We applied Bayesian Networks methods to build a probabilistic framework that helps us to understand the patterns of the translation process, assessing the difficulty of the source texts, the skill of the translators, and the strictness of the reviewers, together with the consistency of both the last two. We run three validation methods in which we test the two Bayesian models created, comparing them, and showing that they can reasonably fit the data and retrieve significant patterns in the behaviour of the linguists involved. We designed an experiment to add new data to our dataset to check the predictability of the quality of the individual translated texts, and despite the single estimate of the quality of the specific translation has been shown

to be poorly predictable, the best of our models has proven to be able to predict the mutual relations between reviewers, showing the possibility to represent a useful tool in assessing linguists behaviours and therefore in establishing their reliability.

Chapter 1

Introduction

Translation Quality Assessment (TQA) is the field of Translation Studies (TS) (the academic interdisciplinary dealing with the systematic study of the theory, description and application of translation ([Holmes, 1988](#))) that focuses on the evaluation of a translated text quality, both in the case where the translator is a human (Human Translation (HT)) and where it is artificial intelligence (Machine Translation (MT)). The translation is clearly a complex process, involving several linguistic and extra-linguistic variables, associated with texts, translators, languages and the environment, and such complexity is reflected in the quality estimation of the translation's final result ([Castilho et al., 2018](#)). For this reason, translation quality has been a central research topic in the field of translation studies since decades ([Holmes, 1988](#); [Nida, 1964](#)). One of the main issues raised by translation scholars is the lack of an objective way to estimate translation quality. In 1997 House stated that "evaluating the quality of a translation presupposes a theory of translation. Thus different views of translation lead to different concepts of translational quality, and hence different ways of assessing it" ([House, 1997](#)).

Together with the complexity of TQA, the need of having a reliable way to assess translation quality attracted the attention of many researchers. Such

necessity lies both in academic research and in industry interests, which need a reliable way to assess the quality of the produced translations. As for the former, in addition to what has been said so far, it has to be remarked that TQA is crucial for the comparison of MT systems, and thus in a scenario in which MT usage is increasingly high (Lagarda et al., 2015; Way, 2018), the growing effort put in MT systems improvements brings along the need of reliable metrics through which the changes in performance can be evaluated and compared between different engines. As a concrete example of issues that may arise from a lack of reliability in TQA, we recall that in 2018 Hassan et al. (2018) claimed that Microsoft’s MT engine reached human parity in terms of quality. The result was obtained using datasets from a conference on Machine Translation¹. It was obviously a surprising claim, but it was immediately refuted by Toral et al. (2018). Due to their lack of expertise, the evaluators generated unreliable data on quality assessment and induced Hassan et al. (2018) to claim for the reaching of MT human parity that did not actually apply, as Toral et al. (2018) proved with the same exercise carried on by expert translators. It is clear, therefore, that is crucial to have a reliable TQA in academic research concerning MT, and, although its importance, it is not a common practice to investigate the reliability of the methods used in the TQA process (Han, 2020) and of the translators involved. Most of the difficulty in estimating such reliability is guided by the lack of standard validation methods, which pushed us to investigate new methods able to trace translators’ behaviour so that to give insights about their reliability.

From the translation industry point of view, it is clear that Professional Translation Service Providers (companies that provide professional linguistic translations) need a reliable method to evaluate the translators engaged and to be more

¹WMT-2017, Bojar et al. (2017)

and more competitive, being able to communicate with transparency the quality of the service provided to their customers ([Martínez, 2014](#)).

In the process of evaluating MT systems' performance, the benchmark for the quality of produced translation is the correlation with human reference ([Castilho et al., 2018](#); [Han, 2016](#)): the MT are asked to translate sentences for which a human translation is already available (references). The closer an automatic translation output is to the relevant human reference, the higher is the "performance score" of the MT engine that produced it. This process has two main shortcomings:

1. The **lack of standard metrics**: the MT output can be directly judged by a human translator (human assessment) or automatic metrics can be used to find the distance of the MT output from the human reference. Furthermore, in both cases there exist several kinds of automatic metrics or human annotation frameworks, and the choice of one of them is far from being a uniform practice (see [Han, 2020](#); [Rivera-Trigueros, 2021](#)).
2. The **human-reference translations** are commonly used as a gold standard to compare with, but often their **quality is not verified** ([Castilho et al., 2018](#)) and this inevitably introduces noise in the evaluation.

One may argue that the translations used as references to assess outputs are provided by expert translators and so it is legitimate to ask why the human references' quality should be questioned. The main concern about the quality of human translation rises from its intrinsic subjectivity ([Zehnalová, 2013](#)), as revealed by some analyses of the agreement between translation annotators ([Daems et al., 2013](#)). Such subjectivity naturally recalls the concept of variance, in a physicist's mind. We could ask ourselves: How many different translations we will have if we would ask a group of linguists (even experts) to complete independently the

task of translating a sentence? How many times the same translator will produce the exact same translation, if he did it several times (assuming that he could not remember the previous time he worked on that)? Speaking of translation quality assessment, for instance, in their exercise of identifying translation problems in HT and human-edited MT, [Daems et al. \(2013\)](#) made two expert translators annotate translation errors in the same set of texts and they found that in evaluating the acceptability of the translations, the expert annotated the same errors with the same labels in only 38% of the cases, meaning that even among experts, the agreement in this kind of evaluation is not to be taken for granted.

Of course, also keeping in mind that human translation subjectivity is somehow its intrinsic property, the need is to go toward the maximum degree of objectivity ([White et al., 1994](#); [Yildiz, 2020](#)), by identifying and minimising the sources of variability caused by subjectivity, so that to have a quality assessment with the highest possible agreement. The claim for the necessity of subjectivity "removal" ([Han, 2020](#); [Rivera-Trigueros, 2021](#); [Turchi et al., 2014](#)) opened a research area that, to the best of our knowledge, is still quite young.

Given the complexity of the TQA task, there are few attempts to overcome such a subjectivity: [Turchi et al. \(2014\)](#) replace human annotations on translation quality with an automatic system that annotates the quality of the translated sentences by looking directly at the similarity between MT output and its human-edited version, discarding the human error annotations. [Daems et al. \(2013\)](#), instead, let the translators find the consensus by talking to each other, to handle their low agreement in independent evaluations.

The research project described in this thesis is an attempt to provide a reliable quality estimate through the application of complex systems methods, in particular in a Bayesian probabilistic framework, by modelling translation subjectivity sources and separating their contribution to the final quality score, gaining

insights about the parameters that affected the result and maximising the objectivity of the judgement. Such methods are expected to catch more complex patterns by looking at the same time at all the agents involved in the process of quality assessment, like the linguist who produced the translation, the reviewer who annotated its quality, the source text from which the translation started, comparing all the available co-occurrences in the dataset.

It is worth mentioning that we would like to propose a general framework, a model which is not only applicable to Translation Quality Assessment but could be used in other fields, wherever the aggregation of subjective quantitative judgements is required, e.g. in medicine, where there could be a lack of consensus in physician's diagnosis ([Jauniaux et al., 2018](#); [Rhee et al., 2016](#); [Tacchella et al., 2018b](#)).

To clarify the background of this project, we will briefly describe the following concepts in the next sections of this chapter:

- **Complex systems.** Which are the characteristics of a system to be considered a complex system and which are the fields where notable examples of complex systems arise, with some examples of applicability of the methods to model such systems;
- **TQA - state of the art.** The methods currently used to evaluate translation quality, both produced by a computer (MT) and by a human (HT). We will describe some of the methods through which the quality score is created, with a special focus on error analysis, the one adopted to create our dataset. Finally, the translation company Translated Srl will be introduced, to explain the nature of the collaboration that gave rise to this project and the workflow that the company uses, as well as their business questions, useful for a better comprehension of our contribution in the business viewpoint

and not only for research purposes;

1.1 Complex systems

Since few years scholars started to view translation as a complex system ([Marais, 2014](#); [Shreve, 2020](#)), but *what do we mean when we say that translation is a complex process?* And *what is exactly a complex system?* In the following we will try to answer these questions, starting this section by going deeper into the concept of complex systems and finishing by framing linguistic translation in the context of complex systems.

1.1.1 What is a complex system?

While there is not a universally accepted definition of complexity, we can follow the recent Nobel Laureate Giorgio Parisi in saying that *a system is complex if its behaviour crucially depends on the details of the system* [Parisi \(2002\)](#). This dependence could be very hard to understand so it is impossible to predict the exact state of the system in a deterministic way. At the beginning of the 19th century, Laplace wrote that the present state of the universe is just the effect of its past and the cause of its future and that an ideal intelligence who knew the current state of a system and the laws of motion could predict exactly its future states. Actually, we know, now, that very few are the systems for which Laplace's statement is true. The study of complex systems, together with statistical mechanics and the discovery of quantum physics, made us reformulate the notion of *prediction*. Indeed complex systems are unpredictable from the deterministic point of view, but it is still possible to have information about future states if we accept such predictions to be probabilistic.

1.1.2 Some examples of complex systems and relative works

It is possible to find examples of complex systems in several fields, and within each of them, it is likely to encounter research works aimed to describe and model such complexity. One property shared by many complex systems is the so-called "emergent behaviour", which is the collective behaviour that we can observe as a result of the interaction between the simple elements of which the system is made up. Everyone has seen the fascinating dance to which the flocks of birds sometimes give rise in the sky: that is a perfect example of self-organisation, created by the interactions between the individual birds that constitute the flock. In ([Ballerini et al., 2008](#)) the authors found out that topological distance between individuals is an important parameter for understanding the formation of collective behaviour. Moving to social science, [Sirbu et al. \(2017\)](#) model the mechanisms that compete to opinion formation, considering the interaction of different elements, like individual predisposition, peer interactions, available information and so on. [Tacchella et al. \(2012\)](#) use an algorithmic approach to describe the economic competitiveness of a country, creating new economic parameters such as the Fitness of a country and the complexity of a traded product. A current application of agent-based models (the agents are the simple elements constituting the system) is the one concerning pandemic spread. It should not be surprising the amount of uncertainty around Covid-19 predictions and the difficulties in modelling it ([Vespignani et al., 2020](#)), given the big number of parameters that guide the evolution and the spread of the infection, like social contacts, virus recombination, its capacity to affect both human beings and animals, and so on. Despite that, it is possible to retrieve valuable information thanks to the application of complex systems' methods, like, for instance, Bayesian inference to estimate the efficacy of non-pharmaceutical interventions ([Flaxman et al., 2020](#)) or to assess

the reliability of data concerning imported infected cases detection ([Niehus et al., 2020](#)), as well as Machine Learning methods used to forecast the disease outbreak dynamics ([Radev et al., 2021](#)). It's obviously beyond the scope of this chapter to give an exhaustive review of the literature concerning complex systems, but it is worth mentioning that such modelling attempts cover many different fields, like, as said, social science ([Del Vicario et al., 2016](#); [Sîrbu et al., 2017](#)), economy ([Tacchella et al., 2012, 2018a](#); [Zaccaria et al., 2016](#)), physics ([Mézard et al., 1987](#)), language science ([Amancio et al., 2011](#); [Antiqueira et al., 2007](#); [Rêgo et al., 2014](#)), biology ([Gorochowski et al., 2012](#)) and so on.

1.1.3 Common methods for modelling complex systems

Given the property of complex systems to exhibit collective behaviour as a result of interactions between simple agents, it's useful to represent such systems as networks (or graphs), composed of nodes, i.e. the constituting elements of the systems, and by edges, that are the interactions that connect the elements. With network representation, we are allowed to consider interactions between elements without limitation in number or distance, and through that, it is possible to gain interesting insights about the system, as in the example shown in figure 1.1, where [Samoilenko et al. \(2016\)](#) showed similarities in the network of languages (nodes) connected through shared topics of interest, extrapolated from Wikipedia co-editings dataset. The methods used to get insights into such networks are usually common to the different fields of application. It is standard practice, for instance, to make use of Machine Learning (ML), that branch of computer science where a machine learns how to face a given task. The great advantage of ML is that such algorithms are capable to extract from a big set of features information that would be too complicated to find otherwise. We will give some examples of ML algorithms used in this framework, without of course covering all the possibilities

offered by this computer science field. An example of ML is neural networks, algorithms that receive as input a set of features and appropriately combine them, learning parameters of such combination, to return an output, usually minimising a loss function calculated with respect to the ground-truth (supervised learning). Neural networks are widely used both when the output is categorical (classification problems, like in the case one wants to know if a patient is sick or not, given a set of symptoms) and in the ones in which it is a continuous score (regression). Decision trees are particularly used in decision making and data mining, usually in a supervised learning setup (see for example [Tacchella et al. \(2018b\)](#)), with the advantage of great interpretability. Another kind of algorithm considerably useful in understanding complex systems is clustering techniques, like hierarchical clustering, in which the algorithm learns how to group entities based on a well-defined distance between them and makes it possible to get useful insights even in unsupervised learning, that is the context in which the researcher cannot directly compare the results of the algorithm with actual ground-truth.

In the study of complex systems, researchers aim usually to find probability distributions over the possible outcomes, having theoretical prior knowledge of the context. That's why Bayesian inference has taken an important place in the study of such systems. Bayesian inference is a statistical method that uses Bayes' rules to update some prior hypotheses according to the available evidences (observed data).

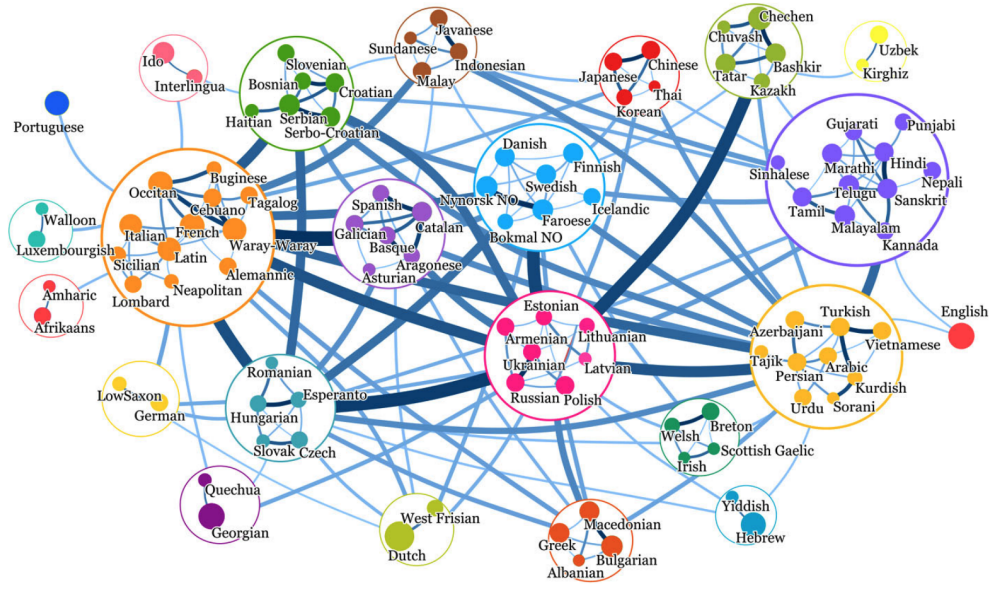


Figure 1.1: Example of a network. Languages (nodes) are connected through shared interests, evaluated on the Wikipedia dataset. It is possible to recognise readable clusters among languages. Taken from [Samoilenko et al. \(2016\)](#).

1.1.4 Linguistic translation: a complex system case

In a position paper published in 2009 [Beckner et al. \(2009\)](#) invited the scholars to join their sight of language as a complex system, highlighting the advantages of such a vision. Such a view was not completely new, especially to physicists, that were starting to lend methods of statistical physics for studying "the emergence of a shared lexicon (i.e., a communication system) in a population of agents" ([Dalla'Asta et al., 2006](#)). In the following years, the interest of scholar from multiple disciplines has been growing ([Kretzschmar and Kretzschmar Jr, 2015](#); [Loreto et al., 2011](#); [Steels, 2011](#)), calling for cross-pollination between different fields.

The same adaptation of complex systems framework came later for translation theory. Following [Shreve \(2020\)](#), one of the first to talk about translation as an emergent behaviour was [Marais \(2014\)](#). Shreve specifically refers to *cognitive translatology*, that field of TS that focuses on the cognitive nature of translation

and considers translation as a complex adaptive system² "with a pattern that emerges from interaction of relevant systems and element collectives, rather than a process with specific control mechanism". In doing so he traces convincing analogies between the characteristics of a complex system (the ones that we illustrated at the beginning of this section) and the postulates formulated by [Muñoz Martín \(2015\)](#) to define cognitive translatology. Among Munoz's postulates concerning translation, it is particularly useful to mention the following:

- It is interpretations, not texts or discourses, which are translated or interpreted.
- Translating is an interpersonal activity.
- Translation expertise implies the continuous development of natural cognitive skills.
- Cognitive translatology should focus on the interaction between translators and their environment.

It is evident that these statements about cognitive translatology make us think about complex systems. Following Shreve's arguments we can notice that translation is an example of cognitive systems, as Munoz states, that are widely accepted to be considered a complex system (see special issue with satellite on CSCS [Jordan et al. \(2015\)](#)). In the time course of translation process, the translator is influenced by some constraints (like stakeholder guidelines) and at the same time he will be guided by his creativity too. This inevitably creates a certain degree

²"Adaptive" term refers to the existence of a final goal shared across the agents of the complex system. Think of the the flocks of starlings that organise themselves in order to defend against predators. We won't specify the difference between adaptive or non adaptive complex system, that is not necessary for the understanding of our work.

of unpredictability in the process that ends with the translation output, resulting from variable interaction of its agents.

In the specific case of TQA, moreover, it has to be considered that also the evaluator (or reviewer) adds variability to the process, by applying his/her style and revealing his/her agreement with the translator, in evaluating the fluency of the translation and its adequacy, that is the quantity of information that correctly passes from source to target language in the translation process.

For the purpose of illustrating how a translation can be practically considered as complex system, we mention the works that make use of the notion of *entropy of translation*: borrowing the Shannon definition of entropy, i.e. degree of uncertainty inherent to a random variable possible outcomes ([Shannon, 1948](#)), the entropy of translation can be defined as

$$H(T) = - \sum_{t \in T} p(t) \log p(t)$$

Where T is the set of all possible t outcomes of translation. Here $p(t)$ represents the probability of the translation outcome t to be suitable for the respective source text. The more predictable is the observed translation in the target language of a given word of the source, the smaller will be its entropy. This allows to gain information about the difficulty of the text to be translated and so the amount of cognitive effort required to accomplish the task ([Dimitrova, 2005](#); [Hale, 2006](#); [Krings, 2001](#)). Using the entropy of translation, [Zhao et al. \(2019\)](#) improve the quality of the output of MT systems by decreasing the high-entropy words, identified to be the causes of some under-translation³ effects. [Martínez Martínez and Teich \(2017\)](#) use the translation entropy to distinguish learner from expert

³Under-translation is a negative behaviour of MT engines that miss some words in the source text.

behaviour in translation and [Carl and Schaeffer \(2017\)](#) highlight the differences between literal and non literal translations.

Few attempts were made also in the direction of using networks topology to face translation's complexity, as in the case of [Amancio et al. \(2008, 2011\)](#) where the authors make use of metrics of the complex networks to distinguish between manual (HT) and automatic (MT) translation, clustering different systems on the basis of such metrics. [Wang et al. \(2018\)](#) build a complex network of the core factors identified by 10 contemporary translation theories and verify their relevance in the actual performance of translations.

Besides few other attempts similar to the ones mentioned here, to the best of our knowledge there is still scarcity of literature concerning the study of translation under the light of its complexity.

1.2 Translation Quality Assessment - state of the art

As we stated at the beginning of this chapter, the lack of standard metrics for evaluating translation quality is one of the highest obstacles for TQA improvement. There are many ways to evaluate the quality of a translation and it is a common opinion among scholars that the choice of one method rather than another strongly depends on the personal view of translation ([Castilho et al., 2018](#); [House, 1997](#)). Furthermore, TQA has got a big relevance both in the academic and in the industry points of view, thing that created even more availability of different evaluation settings, since usually industry and academic research have different needs. While in the TQA academic research the focus is on the assessment of performance improvements and comparison between different sources of translation (e.g. the correct evaluation of a MT engine's output quality), in the translation industry the companies are more interested in having a reliable

and transparent quality evaluation to share with the customers. In this regard, [Koby et al. \(2014\)](#) suggest that "an agreement on an objective way to measure translation quality seems desirable", especially for the advantage that the whole community would have in having the possibility to compare translations using few universally accepted metrics. While at the beginning of MT, it made sense to have different settings for evaluating translations produced by a computer or by a human linguist, in the last years MT quality considerably enhanced and, although it is still possible to distinguish machine translated texts from human translated ones ([Bhardwaj et al., 2020](#)), the two of them are now pretty close in terms of quality. Moreover they are increasingly interpenetrating, also because of the usage of tools assisting humans in the process by providing them a MT version of the translated text, that need to be just post-edited by the human translator. So, given that currently MT and HT are so close to each other, we agree with [Castilho et al. \(2018\)](#) and [Doherty \(2016\)](#) in saying that the still existing differences in evaluating HT and MT should be less marked.

Finally, especially in the case of MT, further variability is created by the possibility to use either humans or automatic metrics to evaluate the translation outputs.

So, to summarise, the lack of standardisation is mostly due to three main different factors:

1. **The purpose of TQA** - either to provide a product with a quality assessment in case of industry, or to judge improvements in case of Translation Studies;
2. **The source of the translation** - either humans, usually evaluated by human reviewers, or machines, for which the TQA is more oriented in assessing the similarity with human translations in specific standard sentences,

sometimes with automatic metrics;

3. **The evaluator** - automatic metrics score or human linguists

1.2.1 Evaluation methods

Without going deep into the big variety of methods used for TQA (for which we refer to [House \(2014\)](#)), we will just cover some practices, making a distinction between the ones that directly involve humans in the evaluation and those that rely on automatic assessment.

Human evaluation

We can find different theoretical models for Translation Quality Assessment in literature, that [House \(2014\)](#) accurately reviewed. Historically, every time that a quality assessment method showed a weakness, a new one came out to fill the gaps left by the previous method. At the beginning TQA was rather a translation criticism ([Nida, 1964](#)) and then it slowly moved in the direction of subjectivity removal. Here we mention some of the changes happened in the theoretical setting of TQA.

Theoretical background One of the first approaches proposed to minimise the subjectivity in TQA was the *equivalence approach*, i.e. the approach in which the translation quality is determined assessing how much the translated text is "equivalent" to the source text. Translation has to be considered as a reproduction of the source text in the target-language and so it is "the attempt to reproduce the source text as closely as possible" ([Lauscher, 2000](#)). As can be easily guess, this approach created a long debate about the practical meaning of an "optimal equivalence" ([House, 2014](#)) and resulted in a lack of consensus on how to establish this equivalence practically. It brought to several different approaches like the

functionalist approach (Reiss and Vermeer, 1984) where the quality of translation is measured with its level of *adequacy*⁴, so it is very "target-oriented" and "it is the purpose of a translation that determines the translation strategy and the shape it takes in the host culture" (Williams, 2013). The "purpose" of the translation, though, is not always easy to define and, above all, it is not clear "how one can determine whether a given translation fulfils its skopos" (purpose) (House, 2014).

Applied settings. The different theories on TQA of course have led to a plethora of practical methods, born usually to pursue different individual interests, ranging from the need of a qualitative judgement about the whole translated document, to quantitative analysis of the translation errors. In his review, Han (2020) provided useful insights about some of these methods, highlighting their strengths and weaknesses. One of the most criticised was the *intuitive assessment* (Eyckmans et al., 2012; Martínez Melis and Hurtado Albir, 2001), a global judgement of the overall quality of the translation, that provides just a qualitative assessment of it, obviously suffering from high degree of subjectivity. *Error analysis* (Martínez, 2014), instead, attracted a lot of attention among scholars and industries because of ease of understanding of its results. With this method, widely used since the end of 20th century (Williams, 1989), translation quality can be set by a quantitative score that can be calculated with a weighted sum of the errors, identified within both categories and severities. Both the weights assigned to the error severities and the choice of categories and severities to use, though, are usually adapted to the specific needs (Martínez, 2014): in the case of a translation company, for instance, they will be decided together with the translation requester. In figure 1.2, we report four examples of settings for error-analysis,

⁴Adequacy is the the measurement of how much the information in the source text is actually transferred to the translated text.

METRICS	LISA QA Model	SAE J 2450	TAUS Benchmarking	QAT
ERROR TYPES		Miscellaneous		
	Accuracy	Omission	Accuracy	Omission
	Terminology	Wrong Term	Terminology	Terminology
	Language	Syntactic Punctuation Misspelling	Language	Grammar Punctuation Spelling
		Word structure or agreement error		
	Country		Country standards	
	Consistency		Consistency	Reference documents
	Style		Style	Clarity
	Mistranslation		Mistranslation	Sense
SEVERITY LEVELS	minor, major, critical	minor, major	minor, major, critical	minor, major

Figure 1.2: Examples of different error-analysis frameworks, with different choices of categories (error types) and severities. Table take from [Martínez \(2014\)](#).

taken from [Martínez \(2014\)](#). For professional translation service providers, error analysis turned out to be the best way to be transparent with customers (*ibid.*), as it could be set an agreement on the threshold of acceptability for a translation job, and on categories and severities to consider, in order produce a final result as suitable as possible for the customer needs. As [Han \(2020\)](#) pointed out, error analysis has been criticised because of an "over-hyped objectivity claim", in agreement with the assumption that "it is logical to assume that the assessors would not always agree with one another on error type and error severity. As such, there is no guarantee that error analysis is objective and reliable.", that is in fact the central point of our work. Another approach on practical application of TQA is the one where references are provided to the assessors, so to have a benchmark for judging the lexicon and the choices adopted for the translation. It is the case of *corpus-based evaluation* method ([Bowker, 2001](#)), where it is possible to assess the quality of a translation job through the quantification of difference with a big reference-corpus, also with statistical analysis ([De Sutter et al., 2017](#)),

which "attempt to quantify the overall acceptability of translation, by statistically analysing a set of linguistic features (e.g., type-token ratio, lexical density, word frequency) of original and translated languages in order to identify possible differences" (Han, 2020). We agree with Han when he concludes that corpus-based evaluation would be a "helpful tool providing complementary information" rather than a valid replacement of error analysis, especially because of the energy, time and expertise required to compile such reference-corpora, that limit the wide usage of the corpus-based methods.

As the opposite of the bottom-up approach of error analysis, *rubric scoring evaluation* (Waddington, 2003) is a top-down approach aimed to catch both textual and extralinguistic aspects of translation quality. In this setting assessment criteria are established and the assessor assign scores, usually in the form of rating scale, after reading the whole translation, according to such criteria. Rubric-scoring method showed to have a good correlation with error-analysis one (Lai, 2011), but requires more competence in building the assessment criteria.

Han (2020) goes on with other different types of methodologies for TQA, and even if it is out of the scope of this work to provide a comprehensive list of such methods, it is useful to remark that the author concludes the review highlighting the paucity of literature concerning reliability of human evaluation methods and call for cross-pollination from other fields. Our work aims to partially fill this gap, as a contribution to the reliability assessment concerning human translators and reviewers, especially in the case in which the linguists involved are all expert.

Some other evaluation methods have been developed to be specifically suited for translations produced a MT system. When comparing MT systems, they are usually asked to translate sentences, and the evaluation of the translation quality is therefore run at sentence-level, rather than in a global setting, like instead it is done in the rubric-scoring method described above, where the evaluator judges

the translation at document-level.

Most of the evolution of the MT outputs evaluation systems can be followed by looking at the Workshop on Machine Translation (WMT)⁵, one of the reference conferences on Machine Translation, in which the participants propose a MT system that competes with the others in different translation tasks (translating a corpus of news texts, translating between languages other than English and so on). Since 2006 in WMT the translation tasks were evaluated by the participants themselves, applying different evaluation frameworks that changed through the years, setting a benchmark for human evaluations of MT systems. The evaluation methods adopted range from direct evaluations of the translated sentence quality through the assignment of a score for every established criterion (such as fluency and adequacy⁶), to sentence ranking, where the evaluators are asked to rank the translations by quality, providing a comparison between different translation systems.

In general, the issue when assessing the quality of the translations produced by a MT systems, is the big volume of translations produced (both for research purposes and for industry interests) and the availability of expert evaluators has become inadequate to meet the demand. To overcome this issue two methods have been employed:

- **Crowdsourcing.** It is the model through which the assessments are obtained by non-expert's crowds, with the advantage of low cost and speed in response. We will briefly introduce it in 1.2.2. WMT added crowdsourced evaluations to the usual expert evaluations between 2010 and 2013, but then they stopped to make use of crowdsourcing because of the very low agree-

⁵<https://www.statmt.org/wmt21/>

⁶Fluency takes into account only the translated text, while adequacy measures how much the information in the source text are transferred in the translated one.

ment rates between evaluators, that created poor reliability of the results (Bojar et al., 2013)

- **Automatic metrics** In this case human evaluation is totally replaced by automatic metrics. These methods obviously bring gain in term of speed, since they do not involve humans, but still lack of reliability and flexibility. In the following we will briefly discuss automatic metrics' employment and limits.

Automatic metrics

The use of automatic metrics emerged to fulfil the big demand of inexpensive and quick evaluation tools, able to compare MT systems performance on a large-scale framework, task that was impossible to accomplish by relying entirely on human

Judge Sentence

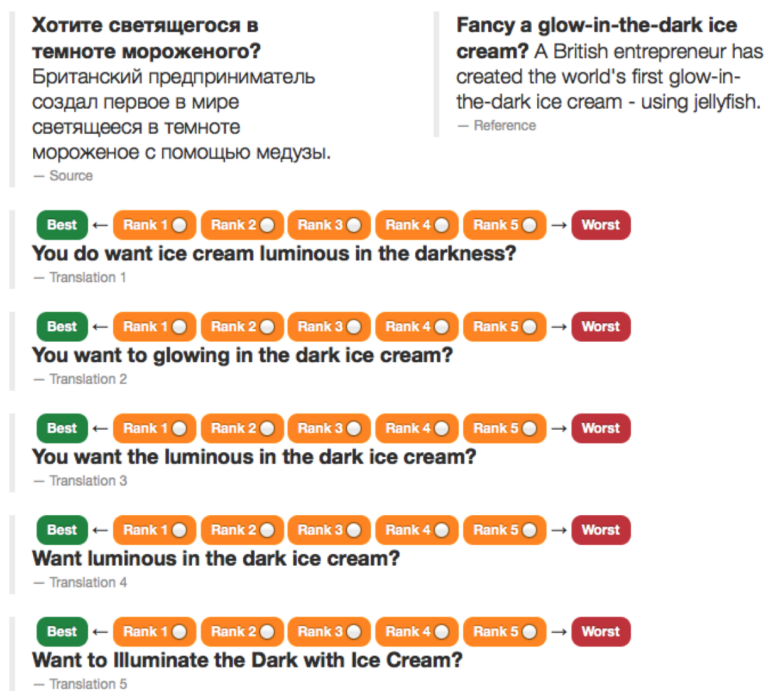
You have already judged 14 of 3064 sentences, taking 86.4 seconds per sentence.

Source: les deux pays constituent plutôt un laboratoire nécessaire au fonctionnement interne de l'ue .

Reference: rather , the two countries form a laboratory needed for the internal working of the eu .

Translation	Adequacy	Fluency
both countries are rather a necessary laboratory the internal operation of the eu .	☹ ☹ ☹ ☹ ☹ 1 2 3 4 5	☹ ☹ ☹ ☹ ☹ 1 2 3 4 5
both countries are a necessary laboratory at internal functioning of the eu .	☹ ☹ ☹ ☹ ☹ 1 2 3 4 5	☹ ☹ ☹ ☹ ☹ 1 2 3 4 5
the two countries are rather a laboratory necessary for the internal workings of the eu .	☹ ☹ ☹ ☹ ☹ 1 2 3 4 5	☹ ☹ ☹ ☹ ☹ 1 2 3 4 5
the two countries are rather a laboratory for the internal workings of the eu .	☹ ☹ ☹ ☹ ☹ 1 2 3 4 5	☹ ☹ ☹ ☹ ☹ 1 2 3 4 5
the two countries are rather a necessary laboratory internal workings of the eu .	☹ ☹ ☹ ☹ ☹ 1 2 3 4 5	☹ ☹ ☹ ☹ ☹ 1 2 3 4 5
Annotator: Philipp Koehn Task: WMT06 French-English	Annotate	
Instructions	5= All Meaning 4= Most Meaning 3= Much Meaning 2= Little Meaning 1= None	5= Flawless English 4= Good English 3= Non-native English 2= Disfluent English 1= Incomprehensible

(a) Fluency and adequacy evaluation, from Koehn and Monz (2006).



(b) Sentence ranking evaluation method example, from [Bojar et al. \(2014\)](#).

Figure 1.3: Two different evaluation settings used in WMT.

evaluation, given the scarcity of professional human translators with respect to the big volume of MT engines to be evaluated on different tasks. Notwithstanding its subjectivity, that makes it not always totally reliable, human translation is, at present, the closest thing we have to a gold standard. Hence, whenever one wants to compare two MT systems, or adjust the parameters of a specific MT system ([Och, 2003](#); [Watanabe et al., 2007](#)) he has to establish how close a MT output is to its respective human translated reference.

In practice, given a translation t and a human reference r , all the automatic metrics aim to provide a distance $d(t, r)$ between them. It is not straightforward at all to find the best way for calculating such distance. Because of the intrinsic complexity of the translation that we illustrated in 1.1.4, it is not a trivial task to automatically assess if the differences between a MT output and the respective

human translation reference are acceptable or not. We could have two different translations, that use a different set of words and they could be both totally acceptable, for instance. Several metrics have been developed, trying to deal with this challenging task of establishing distances between sentences. Usually, whenever a distance definition showed a weakness, a new one emerged to strengthen that specific aspect. The first sentence-distance developed is Word Error Rate (WER)([Olive et al., 2011](#)), based on Levenshtein distance([Levenshtein, 1966](#)), that takes into consideration substitutions, insertions and deletions of words needed to go from t to r , but it is absolutely inflexible since it does not allow for words' reordering and use of synonyms. Modifications on WER came out in order to face this inflexibility, that brought to BiLingual Evaluation Understudy (BLEU) ([Papineni et al., 2002](#)), still today the most used metrics to evaluate MT outputs ([Rivera-Trigueros, 2021](#)), because it has been shown to correlate well with human judgements ([Castilho et al., 2018](#)). BLEU takes into account similarities between n -grams of the two sentences, allowing reordering, and it is designed to consider all the reference-sentences available for a given MT output. As for the other metrics, also BLEU is not totally reliable and there are concerns about its limitation among translation scholars ([Song et al., 2013](#)). So, other metrics emerged, such as the ones that created the branch of the reference-free automatic metrics ([Agrawal et al., 2021](#)), that are models trained to produce a score for MT output just given the source sentence, without the need of a human reference. Although these models are very interesting because they do not need human intervention at all, they have been proved to be unable to punish the literally word-to-word translation, called "translationese" ([Gellerstam, 1986](#)), often semantically wrong ([Zhao et al., 2020](#)).

In the end, automatic metrics represent a powerful tool for MT evaluation, constituting a valid alternative when human evaluation is not feasible, but still

not a proper replacement in terms of reliability, so that the combination of human and automatic evaluation appears to be the best solution ([Chatzikoumi, 2020](#)).

In addition, most of the metrics need to be run on a specific testset, and therefore they are useless when we want to evaluate a new translation, without reference. In our work, therefore, we are aware of the need of human assessment and we propose a method to trace the behaviour of the linguists, both when translating and when evaluating translations, so that to provide a tool to improve the reliability of human assessments, by tracking the sources of disagreement between the evaluators.

1.2.2 Improving Translation Quality Assessment reliability - related works

Even though most of research concerning applied settings for TQA is focused on evaluating MT systems and their outputs, usually by finding the best way to assess differences with human references, some effort has been invested also in facing translation assessors lack of reliability. Here we mention some of the proposed solutions. One of the reasons that pushed scholars to question the quality of human evaluations was that crowdsourcing became quite popular ([Ambati, 2012](#); [Bloodgood and Callison-Burch, 2014](#); [Denkowski et al., 2010](#); [Zaidan and Callison-Burch, 2011](#)). Crowdsourcing is "the act of taking a job traditionally performed by a designated agent (usually an employee) and outsourcing it to an undefined, generally large group of people in the form of an open call" ([Howe, 2008](#)). In the specific case of translation, with this method it is possible to obtain a quite big amount of evaluations at low cost. Obviously, there is no guarantee that the evaluators have expertise in the field of translation, thus could happen that the requesters receive also poor quality assessments, sometimes from spammers too. This brought along the need for a reliability-check of the crowdworkers. It has

been proved, indeed, that crowdsourced TQA data absolutely need to be filtered, in order to overcome, for instance, the risk of claiming wrong statements ([Freitag et al., 2021](#); [Graham et al., 2019](#); [Toral et al., 2018](#)).

In order to minimise this risk, some precautions have been occasionally taken. In their work, [Daems et al. \(2013\)](#) wanted to test a new TQA approach and needed a gold standard with which to correlate the different assessments of the new approach. Instead of using just one expert annotator as benchmark, they created the gold standard asking two expert annotators to evaluate the translations and added a second phase in which the two of them could compare each one's own annotations with those of the other, so that the authors reached a higher agreement.

[Graham et al. \(2013\)](#) propose a method to increase the agreement between translation evaluators (Inter Annotator Agreement (IAA)) based on the usage of a continuous scale that the evaluators could use to judge a translated sentence. The method implies the insertion of quality-controlled data, such as human-post-edited sentences as good quality benchmarks, worsened sentences as bad quality benchmarks and repeated sentences, useful to estimate the consistency of the evaluators (10 quality-controlled out of 100 total evaluated sentences for each category), that were used to establish the reliability of the annotator, so that one could keep just the annotations from the most reliable ones. In [Graham et al. \(2015\)](#), then, the authors showed that with $N=15$ annotators per sentence, the average score for a single sentence could be considered accurate enough. In [Mathur et al. \(2018\)](#) the authors propose a probabilistic method to aggregate the annotations created by Graham method described previously, by modelling annotators' reliability. They show that with Graham's multiple-annotations method of cutting off the annotators that don't meet the consistency criteria established on the basis of the quality-controlled sentences check, it could happen that some

reliable annotators can be removed together with the unreliable ones. The model proposed by Mathur is almost the same than our first bayesian model (that we call Gaussian model), independently developed. We think that the presence of another similar contributions in assessing evaluators' reliability is the sign that our work is going in toward the right direction, where are concentrated the endeavours of translation scholars.

In contrast with Mathur's work, we use expert translators and we question their claimed objectivity, using an errors' ratio calculation as quality score and considering it as a noisy measurement of a true quality of the translation job. Furthermore, we work on document-level and we have just one human judgement per document available, without, of course, a ground-truth to compare our results with. Finally, as opposite to [Mathur et al. \(2018\)](#), we do not constrain our model on the expected evaluation outcomes for quality-controlled samples, since we do not have such available data in our dataset. The details will be further explained in the chapter 2.

1.2.3 Translated Srl

This project was born from a collaboration with Translated Srl, a translation company founded in 1999 that is currently one of the top professional translation service providers in the world ⁷. Translated provides high quality translation to customers all over the world, taking advantage of a big network of professional freelance translators available (since 2014 Translated hired more than 20 thousand translators selected among over 300 thousand candidates).

To make the translator's job faster, Translated makes use of a Computer-

⁷<https://csa-research.com/Featured-Content/For-LSPs/Global-Market-Study-2021/Top-100-LSPs>

Assisted Translation (CAT) tool, that splits the source text into segments and provides the translators with the source segment in the original language, along with a pre-translated version of that segment in the target language (see figure 2.2a), that could come from a Translation Memory (if the segment has been already translated) or it can be a MT system output. In 2015 Translated coordinated a three-years Horizon 2020 innovation action⁸ with which they developed ModernMT, the statistical Machine Translation system now integrated into MateCat, the CAT tool coming from another European project in which Translated participated. For these reasons most of the translated text provided by Translated Srl comes from a post-editing work, run by professional translators, on the texts already translated by ModernMT. Some customers can request a review process, chained with the post-editing translation, and so Translated Srl hires another human translator (that we refer to as reviewer) to use MateCat in the revision-mode (see figure 2.2b), the same user-interface, in which reviewers can also annotate errors when editing the translated segments. It is a common practice for the biggest customers to request a revision phase.

This revision process is the most interesting for us, because it generates a quality-score, that is the data we use as our observable. The way in which this quality-score is obtained will be further discussed in 2.1.

⁸<https://cordis.europa.eu/project/id/645487>

Chapter 2

Dataset and methods

2.1 Translation Quality Assessment dataset

As mentioned in 1.2.3, our dataset consists in one quality assessment per job, for which we have translation-id, translator who produced it, reviewer who evaluated the translation, and EPT quality-score, the result of the reviewer’s evaluation. All the data are entirely retrieved by Translated’s database, so here we will explain how these scores are collected.

Translated workflow When customers ask for a translation, they provide a source text and they indicate the language (or languages) they want the text to be translated to. Translated recruits an expert translator¹, and the translation’s workflow starts. The text is uploaded on MateCat, the CAT tool made in-house by Translated that we mentioned in 1.2.3 that splits the text into segments. The length of a segment ranges from a single word to an entire paragraph², but usu-

¹Our dataset has been generated by only translators who are native speaker in the target language.

²Usually the main criteria to decide how to create a segment are punctuation and source-text file’s layout: MateCat is able to recognise figures’ captions, for instance, and it creates a single segment for it.

ally one segment consists of a single sentence. At this point, the first human intervention take part: the translator in charge for the job logs into MateCat and there he can see the segments and several MateCat suggestions about the possible translations for that segment. For each segment, MateCat proposes to the human translator a version of its translation, that is either retrieved from the translator's TM (if the segment matches previously translated segments in the TM used by that translator - see 1.2.3) or produced by ModernMT (MT system, see 1.2.3). Here the translator edits the proposed translations, segment by segment, until the entire document is completely translated and approved (see figure 2.2a).

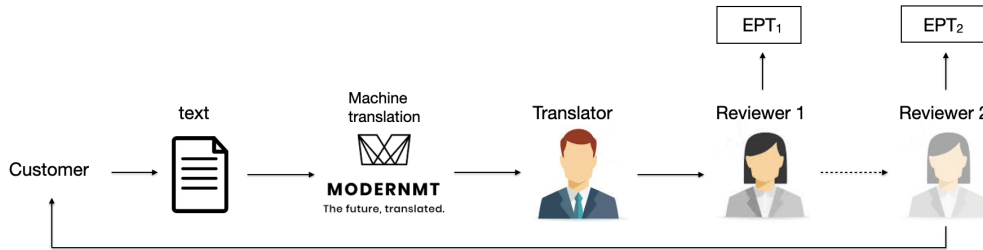


Figure 2.1: *Translated's workflow and creation of EPT dataset: the text is first processed by MT system ModernMT, then a human translator edits its outputs. At the next step, a reviewer edits the human translation and annotates the mistakes, producing the first quality-score (1st Pass Review phase produces EPT₁). Finally, the text passes through another review process, conducted by the most expert translator for the specific language-pair, who produces the second quality score (2nd Pass Review phase produces EPT₂). This last phase occurs just for a random sample of the entire volume of translations delivered in a month for a single customer.*

In order to deliver high quality products, Translated goes further in this process and they hire another translator (to which we will refer as "reviewer") who will review the human translation of the text. This action is taken not only to improve even more the quality of the translation (since the reviewer edits the mistake that he finds), but also to assess the quality of the translated text (that will be shared with the customer) and to evaluate the skills of the translator who

produced it, before the reviewer’s intervention in the chain of translations. This is the TQA process conducted at Translated and we will refer to this first revision process as 1st Pass Review (1PR). In the 1PR will be generated the quality score that we will use as our observed variable: the **Error Per Thousands (EPT)**. It is a ratio between the number of errors found in a translation and the number of words in the respective source text. We will define this quantity in details in the next paragraph.

As a predetermined agreement with the customer, a threshold is set on EPT, so that the customer knows, as quality assurance, that Translated will deliver only jobs that stay within that threshold, and so that do not contain too many errors. Eventually, just for regulars customers, Translated selects the best translator³ for a given language pair and commend to her/him the review of the last translation version, the one provided by the previous reviewer, who edited the translation in the 1PR. This last phase, which we will call 2nd Pass Review (2PR), happens monthly, not for all the translations produced for the requester, but for only a sample of them. To be precise, since we used all the translations created for a specific customer, in our specific case the sample size of the text reviewed twice (consecutively and not in parallel, thus the two reviews are not two different evaluations of the same translation) is around 20% of the entire monthly volume per language pair⁴.

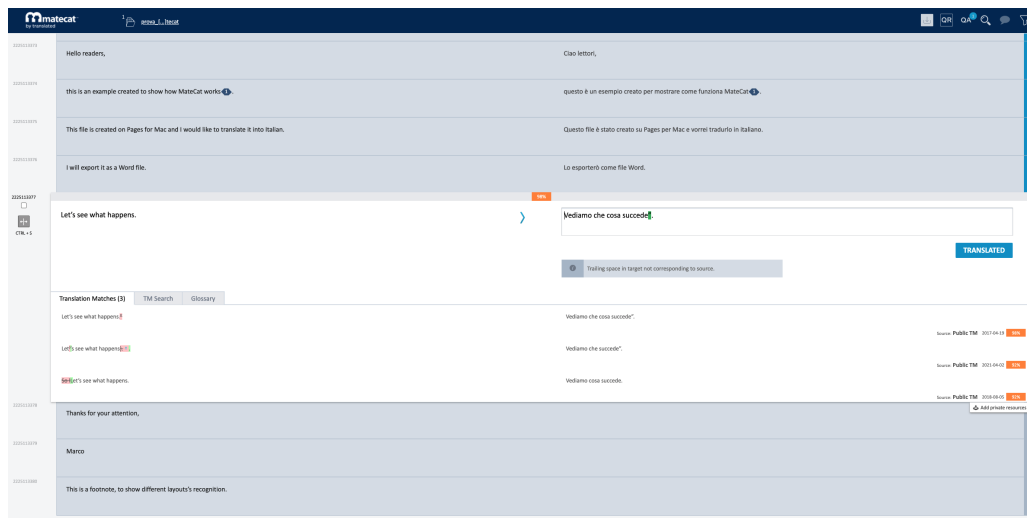
So, at every step of the process, the translation version changes, and only in the review phases the reviewers annotates translation errors while editing the text. The entire workflow is summarised in figure 2.1

³The selection is the result of a combination between certified expertise in the text-subject and that language pair, and customer’s feedback on the quality of translations provided by the linguist.

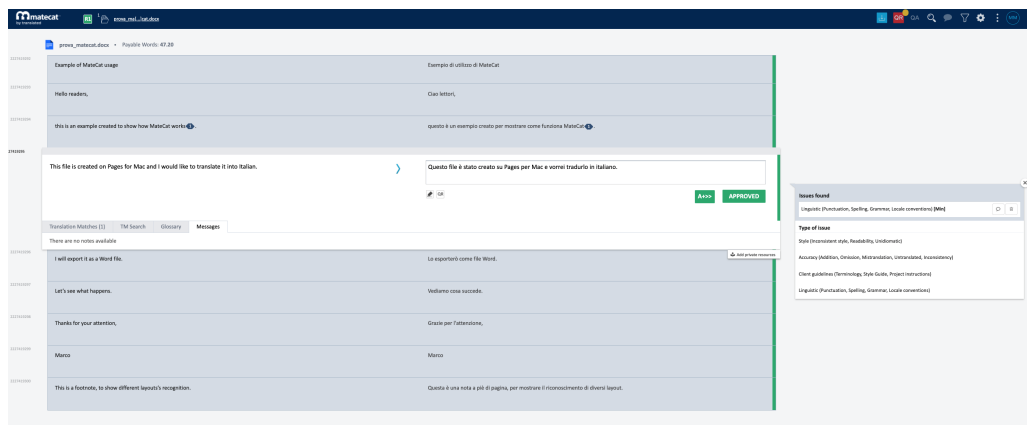
⁴The 20% is calculated on the total amount of words (about 100 thousands), and not documents, so, for this customer, Translated provides a 2nd Pass Review by randomly picking documents in order to reach approximately 20 thousands words.

Error annotation process - EPT As showed in figure 2.1, the revision phase leads to an assessment of the translated document's quality. EPT is the quantitative measurement of such quality, a real number that we take as quality-score of the translation. In the following we will describe the process that produces the EPT.

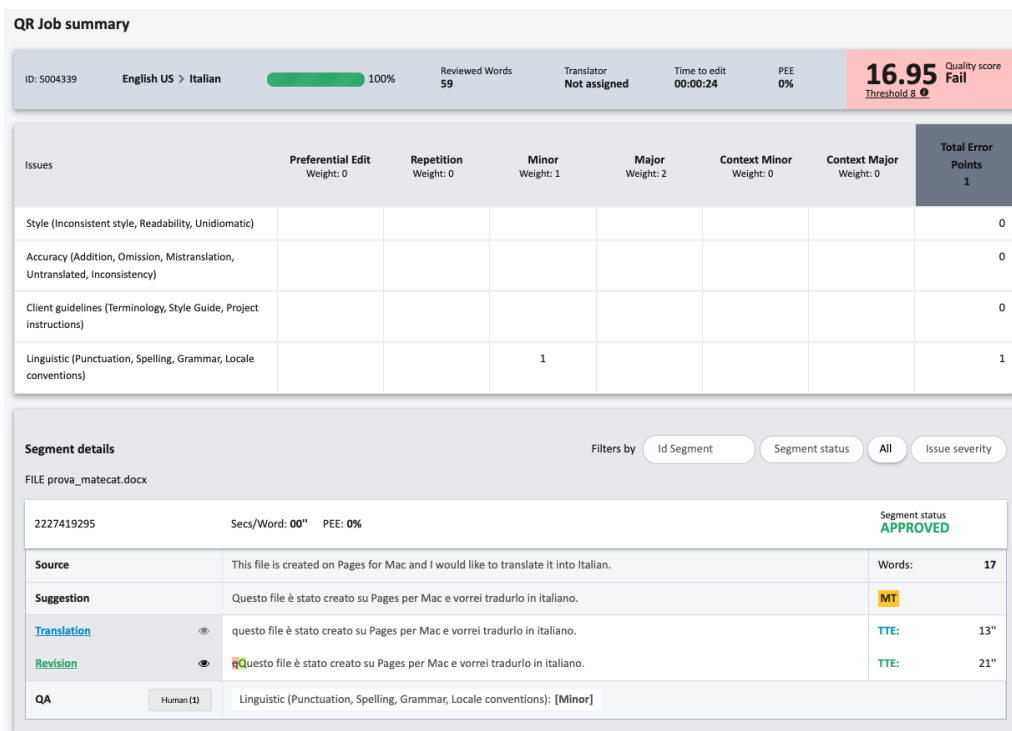
First of all, it is necessary to define the annotation framework used by Translated, i.e. the set of categories and severities within the reviewer can classify the



(a) *Translation phase: on the left the translator can see the source segments of the document in the original language, and on the right he can type his translation, modifying the proposed translation (coming either from MT or TM)*



(b) *Review phase: on the left the reviewer can see the source segments of the document in the original language, and on the right he can type his translation, modifying the proposed translation (coming either from the human translator who produced it in the translation phase. In this phase the reviewer can also annotate the errors found in the translation)*



(c) Quality report: summary of errors annotated in the whole document, with category and severity information (below) and quality-score (EPT) of the document. Here the box in the top-right corner is red because the quality-score is below the set threshold.

Figure 2.2: Screenshots of the MateCat user interface: translation phase (a), review phase (b) and quality report (c)

translation's errors he find. It is not a "one-size-fits-all" solution adopted for each customer and, in the specific case of the data we selected for our project, the annotations are classified into four *categories* and four *severities*. The guidelines received by reviewers are summarised in tables 2.1 and 2.2 and are detailed in appendix A. With reference to what has been said in chapter 1 regarding the evaluation criteria for the translation quality assessment, we could say that Accuracy category refers to the concept of "adequacy" (i.e. how much the information in the source text are transferred to through the translation) and that Linguistic and Style ones concern the "fluency" (i.e. the quality of the translated text, regardless of the source).

The figure 2.2b shows a screenshot of the MateCat user interface in the review

stage: when the reviewer finds an error in a segment, he makes his edits and he annotates the error classifying it with one category and one severity, then he approves the segment and moves to the next one. Once the reviewer has edited and approved all the segments of the document, MateCat generates the quality report (see figure 2.2c), that summarises all the errors annotated by the reviewer in the whole document. On the top-right of the screenshot in figure 2.2c we can see a red box with a "quality score" (16.95 EPT). This score is exactly our EPT and the box is green if the document passed the quality-check, that means that the EPT score is below the set threshold, and red otherwise. In the example in figure 2.2c, with one minor error in a 59 words text, the translation did not meet the quality requirements: it scored 16.95 EPT, that is above the threshold of 8 EPT.

Table 2.2 indicates a weight below every severity label; that weight is used to calculate the final sum of the errors, that will be normalised with the total number of words in the source text of the translation under examination. If we call m and M the total number of minor and major errors respectively, and W the number of words of the document in the source-language, then

$$EPT = 1000 \frac{m + 2M}{w} \quad (2.1)$$

So, the preferential and repetition severities won't contribute to the final quality assessment in the framework used (they have 0 weight) and the major errors count double. The result of the weighted sum of the errors, normalised with the number of words, is multiplied by 1000 for readability reasons and that is the reason why the quality-score is called Error Per Thousands.

Category	Guidelines
Accuracy	Any issue that may impair the conveying of the original message (addition, omission, mistranslation, untranslated, inconsistency)
Linguistic	Any issue that can be proven by official linguistic reference sources (punctuation, spelling, grammar, locale conventions/formats)
Style	The style of the translation is not suitable for the content type or the target audience, translation does not sound and feel native; literal or awkward translation (inconsistent style or tone, readability, unidiomatic/convoluted)
Client Guidelines	Any violation of company guidelines, instructions or ignoring the information provided in the reference resources, including the violation of customer term base (terminology, style guide, project instruction)

Table 2.1: *error categories*

Severity	Guidelines
Preferential (weight=0)	A preferential choice of words/syntactic structures by the reviewer, when they want to improve the translation quality, but the edit is not caused by an error in accordance with customer Quality Framework
Repetition (weight=0)	The result of a single misunderstanding; an error for identical source that is repeated in the same project. Typos should not be seen as repeated errors
Minor (weight=1)	The error that does not confuse the user. The intention of message is clear even with an error
Major (weight=2)	The message meaning is changed or it is hard to comprehend; the message confuses the user; might lead to an unintended action, sending a query to customer support or even stopping using the product

Table 2.2: *error severities*

Pages 38-53 hidden for data protection reasons

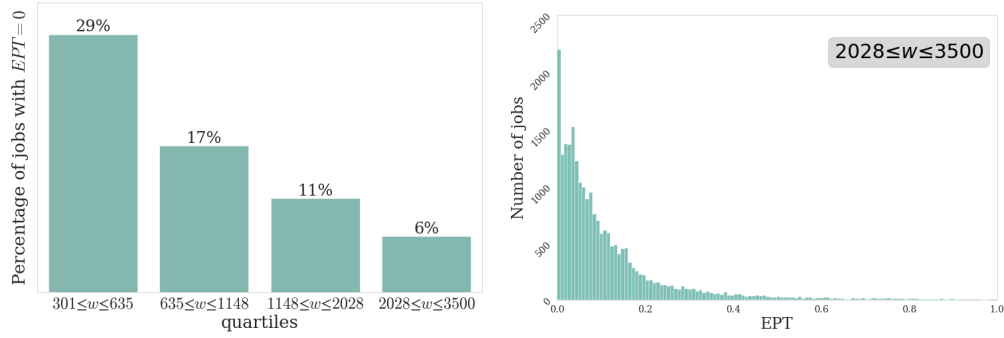
Note

In the following, all absolute EPT values shown in plots have been rescaled by dividing them by their maximum value and languages are anonymised for data protection reasons.

2.1.3 Glut of perfect translations

The last feature of our EPT dataset that is worth describing is the not negligible presence of translation jobs for which the reviewers did not annotate any weighted error: we will refer to these jobs as "perfect translations" or "zeros". In the EPT distribution showed in figures 2.6 and 2.4d, a high peak can be noticed in correspondence of $EPT = 0$, that seems to be detached from the rest of the distribution.

You may argue that such abundance of zeros can be caused by the presence of short documents, for which there is a smaller probability to find an error (and so it is due to a finite-size effect in the size of the documents). To check if it was the only reason for this glut of zeros, we checked the distribution of EPT for longer documents: we divided our dataset into four subgroups, using the quartiles of the distribution of documents per number of words (see fig.2.4a) as boundaries for the separation. We therefore drew again the EPT distribution for the last quartile (documents with a number of words $2028 \leq w \leq 3500$) and we found a similar peak for $EPT = 0$. At the same time, the percentage of perfect translations actually decreases with increasing document size (see figure 2.11a), meaning that the finite-size effect becomes less evident for longer jobs (and so it is true that part of this effect can be explained with the document lengths). On the other hand, the presence of the peak at $EPT = 0$ also for longest documents (see figure 2.11b) drives us to avoid explaining the glut of zeros with just the finite-size effect.



(a) Percentage of jobs with $EPT = 0$ (1PR) within quartiles of jobs distribution divided by number of words w (b) Number of jobs per scored EPT (1PR) for the last quartile of jobs distribution divided by number of words w

Figure 2.11: Glut of perfect translations. Even for longest documents there is a not negligible amount of perfect translations ($EPT = 0$).

It can be suspected, thus, that the reason behind the presence of this glut of zeros can be found both in the limitations of the dataset (i.e. the EPT size-effect introduced by limited number of words) and in physical evidences of some mechanism of the TQA process. With the latter we are referring to the following possibilities

- easy job - if the source-text's translation is quite straightforward it is more likely to find no errors in the review phase;
- good translator - if the human translator who produces the translation to be evaluated is very skilled, it is more likely that the TQA process will produce a zero;
- lenient reviewer - it could be that one reviewer tends to easily accept the work of some translators, because he is just lenient, or maybe because there is a strong agreement between translator and reviewer. Moreover, the reviewer could lack of training and he could use always not weighted severities for the errors. Lastly, it could happen that the reviewer does not pay attention at all at his work.

field	$\epsilon_1 = 0$ (%)	$\epsilon_{1,2} = 0$ (%)
acc	44	73
lng	38	70
gui	59	79
sty	30	57
rep	57	80
pre	25	56
min	23	58
maj	90	96

Table 2.4: *Statistics of translations that scored $\epsilon_1 = 0$ (1PR phase) and $\epsilon_2 = 0$ (2PR phase). In most of 1PR revision jobs (90%), no major errors are annotated and almost each of these (96%) scored no major errors also in the 2PR phase.*

These situations could all lead to get a perfect translation.

2.2 Modelling Translation Quality Assessment

In the light of what we learnt about the dataset and the TQA variability due to its subjectivity, we finally set up our modelling of the quality assessment process. Under the idea that the EPT score is a measurement of a real latent quality of a translation job, produced by a noisy and possibly biased source (the reviewer), we managed to parameterise the variables that affected such measurement. We have taken into consideration the possible sources of the error affecting the quality measurement, so to have insights about the process through the estimation of the latent variables behind the measurements (e.g. the reviewer’s reliability and the translator’s skill). We considered as sources of error the three main agents involved in the TQA process, namely the document itself, the human translator and the reviewer, who finally produces the EPT, our observed measurement.

The parameters that we created in order to model such measurement are the following:

- d_j - job difficulty. The intrinsic source of variability for the quality of one translation, that here we assume being language-dependent
- ϵ_t - translator's error propensity. The error introduced by the translator that will affect the EPT score measured
- q_j - real quality-score of the translated document. It is the latent parameter measured by the the reviewer with the EPT (so the higher is q_j , the worse is the quality of the translation)
- β_r - reviewer's bias. It is the contribution given by the reviewer to the error in the measurement of the real quality q_j

So, we are saying that the final EPT score is the result of a combination between the intrinsic difficulty of the job, the intervention of the translator and the one of the reviewer. Both the human involved can theoretically contribute to an overestimation or an underestimation of the difficulty of the translation job, when looking at the final measurement represented by the EPT, and the way we combine (some of) these parameters define the specific model that we created.

To compound the estimate of these latent parameters, there is the fact that we have only one measurement per translation job: just one review (and so one EPT score) that creates a unique triplet job-translator-reviewer. This makes hard the estimate of the real quality for a single job with high confidence, but we could take advantage of the network of translators and reviewers, linked by the EPT assigned by the reviewer to translator's work, to retrieve valuable information about the linguists.

We faced this problem from a probabilistic point of view, building a bayesian network in which we set our prior knowledge about the parameters of interest, and let the network update such priors according to the set of EPT data observed.

If we call Θ our set of parameters and $\mathcal{E}$ the set of EPT, then we know, from Bayes theorem, that

$$\underbrace{p(\Theta|\mathcal{E})}_{\text{posterior}} \propto \underbrace{p(\mathcal{E}|\Theta)}_{\text{likelihood}} \underbrace{p(\Theta)}_{\text{prior}} \quad (2.2)$$

that gives us the probability distribution of the parameters given the observed data (posterior), calculated as the product of the prior probability on the parameters (prior) and the likelihood probability of observing our set of data $\mathcal{E}$, with our available set of parameters Θ (likelihood). The optimal set of parameters is found by training the models via Markov Chain Monte Carlo (MCMC) (Gelman et al., 1995). With MCMC we infer the shape of the posterior $p(\Theta|\mathcal{E})$ i.e. the probability that Θ is the set of parameters that best fit the observed EPT data, according to our prior knowledge and the model chosen.

Given the amount of parameters (thousands for the jobs, and hundreds for the linguists) we did not approach analytically the MAP estimate, but we made use of PyMC3⁶, a probabilistic-programming framework implemented in Python, that uses the No U-Turn Sampler (NUTS) algorithm (Hoffman and Gelman, 2014), an adaptation of the Markov Chain Monte Carlo (MCMC) technique (van Ravenzwaaij et al., 2018).

Going deeper into the details of our approach, we have to distinguish between the three models that we created:

- **Baseline** - non probabilistic model through which we make point-estimates (and so we do not infer probability distributions) of the parameters
- **Gaussian** - the least complex among our two bayesian models. We called it Gaussian because it takes the assumption that the EPT are normally distributed, so we have a Gaussian likelihood for our data. In the theoretical

⁶<https://docs.pymc.io/en/v3/>

framework of this model, we assume that the reviewer introduces an additive error in estimating the quality of a translation job, while the translator is not parameterised

- **Hurdle** - this model takes into account the "glut of perfect translations", the feature of our dataset that we described in 2.1.3. The likelihood is a mixture of two probability distributions, that belongs to the class of the Hurdle distributions (Cragg, 1971) . Here we assume a multiplicative error, both for the one introduced by the translator and reviewer.

The general bayesian framework that we built has got the advantage that it can be gradually made more complex and comprehensive, just by adding parameters and/or new information. This is exactly what we did, starting from the Gaussian model, in which we consider only the contribution of the reviewer as parameter, together with the job difficulty, to estimate the real quality-score of the translation, and then we added more information, e.g. the translator's contribution and the excess of zeros in the dataset, building the Hurdle model.

In order compare the performance of our model and set a benchmark we first created a baseline model

2.2.1 Models

Baseline

The first model that we present here is the one that we use as baseline. Here we consider the EPT quality-score e_{jtr} for the job j , translated by the translator t and produced by the reviewer r , as obtained through

$$e_{jtr} = q_{jt} + \beta_r \quad (2.3)$$

where q_{jt} is the latent quality score "unbiased": the reviewer is supposed to introduce an additive bias on the measurement (e_{jtr}) that is calculated as

$$\beta_{\hat{r}} = \langle e_{jt\hat{r}} \rangle - \langle e_{jtr} \rangle \quad (2.4)$$

where $\langle e_{jtr} \rangle$ is the mean EPT produced by all the reviewers for all the translations in the language-pair dataset of the specific reviewer $\hat{r}$. So, the bias $\beta_{\hat{r}}$ of the reviewer $\hat{r}$ is given by the difference between his average EPT $\langle e_{jt\hat{r}} \rangle$ and the overall average EPT $\langle e_{jtr} \rangle$.

If we want to find the real latent quality-score of a translation job q_{jt} with the baseline model, we have only to "debias" the scored EPT, by subtracting the estimated reviewer's bias. Then, if we have one quality-score for a job j produced by a reviewer r_1 and we want to predict the quality-score for the same job produced by a second reviewer r_2 , we just have to remove β_1 and apply β_2 .

Gaussian model

The simplest Bayesian model considers the EPT scores as a random variable obtained by the measurement of the real latent quality score q_j . In this model, the errors are introduced solely by the reviewer who evaluates the translation. We have taken into account two kinds of stochastic errors introduced by the reviewer:

- β_r : Offset - a systematic additive error that shifts the distribution of the EPT with respect to the real job quality q_j . Its interpretation is the level of *strictness* of the reviewer: the higher the estimated β_r , the stricter the reviewer, giving EPT scores that are higher (i.e. the texts supposedly contain more errors) than the real quality.
- σ_r^2 : Variance - the variability introduced by the reviewer into the values of EPT observed. For this reason, we consider σ_r to be the reviewer's

consistency.

Our model for the posterior distribution of EPT is thus:

$$p(e_j|\beta_r, \sigma_r) = \mathcal{N}(q_j + \beta_r, \sigma_r^2) \quad (2.5)$$

Since the linguists under consideration are all experts, we give the bias β_r a zero-centred prior, which is additionally weakly informative ($\sigma_\beta = 10$). For σ_r , we set as prior a half-Normal distribution with $\sigma_\sigma = 1$. The scalar q_j is modelled by a Gamma distribution $\Gamma(k, \theta)$ with $k = 1$ and $\theta = 3$, since we wanted to have a considerable tail reflecting potentially high values of the real quality score (i.e., bad translations).

Recapitulating, our Gaussian model is as follows

$$\begin{aligned} q_j &\sim \Gamma(1, 3) \\ \beta_r &\sim \mathcal{N}(0, 10^2) \\ \sigma_r &\sim \overline{\mathcal{N}}(0, 1) \\ e_j &\sim \mathcal{N}(\mu_j = q_j + \beta_r, \sigma_r^2) \end{aligned} \quad (2.6)$$

where $\overline{\mathcal{N}}(0, 1)$ is the truncated standard Normal. This model is illustrated in 2.12 in plate-diagram notation. Notice that we have an independent prior for each reviewer and for each job, so we need to maintain N_j parameters q_j and N_r parameters β_r and σ_r , where N_j and N_r are the numbers of jobs and reviewers respectively. Furthermore, we have just one review per translation job, and so we have one unique random variable e_{jr} for the couple (j, r) of job-reviewer, and this is the reason why the two rectangles in fig. 2.12 do not intersect.

This model allows $e_j < 0$, which is inconsistent with the definition Eq. 2.1. We address this shortcoming in 2.2.1.

The model described here is similar to the model of [Mathur et al. \(2018\)](#), which estimates the reliability of crowdsourcing workers, asking them to assess

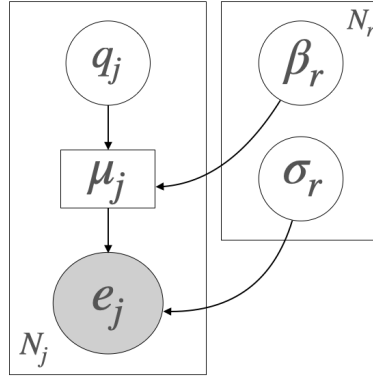


Figure 2.12: *Gaussian model Bayesian network. In this plate diagram circles denote random variables, squares deterministic variables, and grey background indicates an observed random variable. Arrows indicate the direction of influence between the variables. The rectangles that enclose the variables indicate their multiplicity: N_j is the total number of jobs and N_r is the number of reviewers.*

the quality of translated sentences against some quality-controlled sentences. As a consequence, [Mathur et al. \(2018\)](#) focuses on modelling quality estimates, which we do not need here.

Hurdle model

The model described above has three main modelling shortcomings: (i) it allows negative EPT values (positive by definition - see Eq.2.1); (ii) it does not take into account the excess of zeros, described in 2.1.3; (iii) it uses a symmetric likelihood (Gaussian) for the EPT values, while the data shows an asymmetric distribution with a significant tail (see Figure 2.4d).

To address these, we introduce a more realistic model, where we replaced the Gaussian distribution with a hurdle probability distribution ([Cragg, 1971](#)). The class of hurdle distributions has two separate parts, one modelling the excess of zeros, while the other one is the distribution of non-zeros. They are similar to the Zero-inflated distributions ([Hilbe, 2014](#)), with the difference that the latter considers two sources of zeros, the first coming from the same distribution of non-

zeros (which therefore must be defined for zero values too) and the second from another source that creates the "inflation" of zeros. Hurdle distributions take their name from the discontinuity which the variable has to overcome in order to attain a non-zero value. A zero vs. non-zero mixture distribution is governed by a Bernoulli distribution with success parameter π .

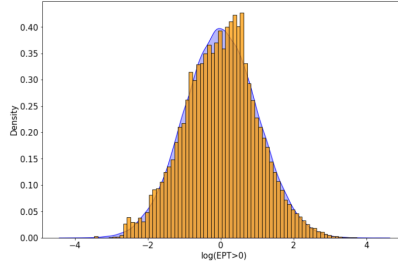


Figure 2.13: *Gaussian fit of the standardised log EPT values. Skewness=-0.01, Kurtosis=-0.03.*

In modelling the glut of zeros described in 2.1.3, we have considered the perfect translations (our zeros) as the result of a separated phenomenon that does not emerge from the usual TQA process and so, we assumed that a particular situation must occur for the translation job to be judged as perfect. More specifically, we considered that zeros can be produced by situations in which, for example, the reviewer skips a segment out of distraction, or the translator is so skilled that he is likely to produce perfect translations. Thus, we considered the possibility that the zeros belong to a different probability distribution than the one for the jobs with $EPT > 0$: the model decides whether it is more likely for a specific combination of job-translator-reviewer to produce a zero, or a positive value, and in the last case the EPT value is modelled by a lognormal distribution.

In addition to the reviewer latent variables β_r, σ_r considered in the Gaussian model, we introduce the *translator error propensity* ϵ_t , which represents the contribution of the translator to the final EPT score, and a *difficulty parameter* d_l specific to the language direction l (e.g. English to German). Here, the mecha-

nistic model underlying our interpretation of contributions, is

$$e_j = d_l \epsilon_t \beta_r \quad (2.7)$$

In this model, therefore, the real translation job's quality, which we called q_j for the Gaussian model (see 2.2.1) is the result of the two multiplicative contributions of language direction and translator's skill: $q_j = d_l \epsilon_t$. We model each of d_l, ϵ_t, β_r as a hurdle lognormal. This is suggested by two factors: (1) the shape of the observed EPT distribution for non-zero values, which can be reasonably fit by a lognormal distribution, as seen in Figure 2.13, where we compare a standard normal distribution with the standardised values of $\log(EPT)$, (2) the error introduced by the reviewer is better modelled as multiplicative, and not additive as stated for the Gaussian model. Therefore, if we use θ in place of any of the three random variables d_l, ϵ_t, β_r ,

$$p(\theta) = \mathcal{HL}(\pi, \mu, \sigma^2) = \begin{cases} \pi & \theta = 0 \\ (1 - \pi)\mathcal{L}(\mu, \sigma^2) & \theta > 0 \end{cases} \quad (2.8)$$

Throughout this article, random variables are indexed by j, l, r, t and have the corresponding dimensionality. **We emphasise that our model needs only a single observation for each triplet (j, r, t) to condition random variables attached to various t, r .** Moreover, we train one model for every language direction l separately.

Using the facts that (i) products of independent Bernoulli random variables are Bernoulli, (ii) products of independent log-normal random variables are log-normal, we can **collapse** the model and write

$$\begin{cases} p(e_j) = \mathcal{HL}(\pi_j, \mu_j, \sigma_j^2) \\ \pi_j = 1 - (1 - \pi_l)(1 - \pi_t)(1 - \pi_r) \\ \mu_j = \mu_l + \mu_t + \mu_r \\ \sigma_j^2 = \sigma_l^2 + \sigma_t^2 + \sigma_r^2 \end{cases} \quad (2.9)$$

where μ_l is identical for jobs from the same language.

We now give suitable priors to all l, t, r -related variables. To avoid too long tails, we give $\sigma_l, \sigma_t, \sigma_r$ informative priors: a truncated (only positive values) normal distribution $\mathcal{N}(0.5, 0.25)$. We give π beta-distributed priors, so as to constrain π in the interval $[0, 1]$. We choose $\mathcal{B}(2, 5)$ for π_l and π_t , and $\mathcal{B}(1.5, 5)$ for π_r , to model a lower mode, expecting that a perfect translation should come out rather due to job difficulty (in the language direction l) or translator skill, than due to the reviewer, who should probably not work at all on the quality assessment to judge a translation as perfect.

We end up with the following probability model:

$$\begin{array}{ll} \pi_{v \in \{l, t\}} \sim \mathcal{B}(1.5, 5) & \pi_j = 1 - (1 - \pi_l)(1 - \pi_t)(1 - \pi_r) \\ \pi_r \sim \mathcal{B}(2, 5) & \mu_j = \mu_l + \mu_t + \mu_r \\ \mu_{v \in \{l, t, r\}} \sim \mathcal{N}(0, 1) & \sigma_j^2 = \sigma_l^2 + \sigma_t^2 + \sigma_r^2 \\ \sigma_{v \in \{l, t, r\}} \sim \overline{\mathcal{N}}(0.5, 0.25) & e_j \sim \mathcal{HL}(\pi_j, \mu_j, \sigma_j^2) \end{array} \quad (2.10)$$

is illustrated as plate diagram in Figure 2.14.

We conclude the illustration of the hurdle model by giving our interpretations of the parameters in view of the next section:

- $\pi_{v \in \{l, t, r\}}$ - probability contributions to the final π_j , the probability to have a perfect translation. π_l here is the overall difficulty of translation in the language pair l , while π_t and π_r are the analogue probabilities for the individual translator and reviewer respectively. The higher π , the more likely the outcome of the TQA process will be a perfect translation ($\text{EPT} = 0$).
- $\mu_{v \in \{l, t, r\}}$ - contributions to the mean μ_j of the underlying normal distribu-

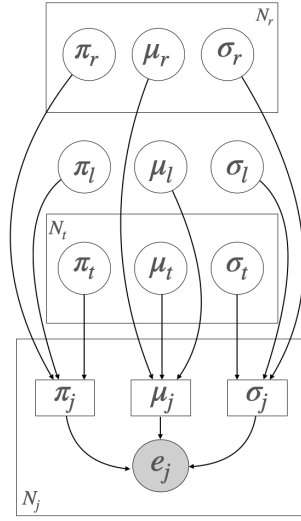


Figure 2.14: *Hurdle model bayesian network. Circles enclose random variables and squares the deterministic ones. The arrows indicate the direction of the influence between variables and the grey-filled circle indicates the observed variable. Here we have N_j for the number of jobs, N_t for the number of translators, and N_r for the number of reviewers. The rectangles enclose the variables with the same multiplicity. The reader can imagine the entire figure as contained in a plate indexed by l , since each models are independent for each language.*

tion for the $\log(e_j)$. The higher μ_t , the worse the skill of the translator it represents; the higher μ_r , the more biased the reviewer; the higher μ_l , the more difficult it is to produce a high-quality translation in the language pair l .

- $\sigma_{v \in \{l,t,r\}}$ - contributions to the standard deviation σ_j of the underlying normal distribution for the $\log(e_j)$. The higher σ_t , the more inconsistent the quality of the translations by translator t . The higher σ_r , the noisier the contribution of the reviewer to the EPT, regardless of the translator's skill who produces the translation and the difficulty of the job (we will refer to the reviewer with high variance as an "inconsistent").

Chapter 3

Results

3.1 Validation in unsupervised setting

The final goal of our modelling is to use the inferred parameters of the model (see 2.2) to extract reliable insights on the features affecting the TQA process: job’s difficulty, translator’s skill and reviewer’s behaviour. The hardest challenge that we face is how to validate our results. Obviously we are inferring some latent parameters for jobs, reviewers, translators because we want to gain insights about the physical meaning of such parameters, so that we could state that the real quality of the translation jobs provided by Translated Srl, for instance, is higher (or lower) than the actually measured one, or, again, that one reviewer shows a high level of inconsistency, or strictness, compared to the others.

We cannot directly measure how close our inferred parameters are to the true values of the quantities they represent, since there is not an actual external (to the model) evidence of them. This measurable evidence we are lacking of, is called *ground-truth* in Machine Learning domain, and the modelling setting in which the ground-truth is missing is called *unsupervised*, since it is not possible to supervise (comparing predictions to real outcomes) any model performance improvements.

In order to address the challenge posed by the lack of ground-truth, we built a framework in which we can *indirectly* compare our results with an external available source of information. The reason behind the term *indirectly* lies in the idea that we can use latent parameters inferred by the models (for which we have no direct evidence) in order to produce data whose we actually have a ground-truth to compare with. For example, we are not aware of whether a reviewer has actually a strict and consistent behaviour, but if we are able to reproduce his performance by extracting new data from the probability distributions shaped by the inferred parameters, then we could be reasonably confident that we are correctly fitting the data with the right parameters. The better the model outcomes will approximate the truth, the more meaningful will be the inferred parameters, and so the better will be the performance of the model. Under this idea, we managed to validate our results with three experiments:

1. Posterior predictive check - once we have optimised the parameters, we measured how well the models could reproduce the same raw EPT data observed in the optimisation phase, just drawing samples from the inferred probability distributions;
2. Double review experiment - we created a new dataset of translations reviewed independently by two different reviewers, in order to predict new unseen EPT;
3. Skilled translators recognition - we checked the ability of the model to assign higher skill-levels to the translators who have certified expertise;

In the following we will describe in detail the validation methods and the respective results.

3.1.1 Posterior predictive check

Based on the observation that "the observed data should look plausible under the posterior predictive distribution" (Gelman et al., 1995), we draw new EPT samples $\mathcal{E}^{rep}$ from the *posterior predictive distribution*

$$p(\mathcal{E}^{rep}|\mathcal{E}) = \int p(\mathcal{E}^{rep}|\Theta) p(\Theta|\mathcal{E}) d\Theta \quad (3.1)$$

where $p(\mathcal{E}^{rep}|\Theta)$ is the model likelihood, while $p(\Theta|\mathcal{E})$ is the posterior. The integral is not usually calculated analytically, but instead approximated through a Monte Carlo estimate based on samples from the posterior.

In order to compare simulated against observed data, we have taken the following features of the distribution as a benchmark:

- number of zeros - as we pointed out in 2.1.3, a distinctive feature of our dataset is the one we called the "glut of zeros". Since we are interested in the interpretation of the source of this phenomenon, we want to be sure that the model is able to reproduce such inflation of jobs with $EPT = 0$;
- Kullback-Leibler divergence (KL) - we compare the overall shape of the generated distribution with the shape of the observed EPT through KL.

Results

First, we split the dataset by language and run the inference, obtaining per-language posteriors of the parameters for both the Gaussian and Hurdle models. Then, using the same combinations of job-translator-reviewer, we draw new samples from the posterior predictive distribution defined in Eq. 3.1, i.e. we generate a simulated EPT value for each job of the language pair. For each language pair, we produce 1000 replications of the language pair subsets of EPT dataset, both

with the Gaussian model and with the Hurdle one, and we obtained a set of distributions, shown in Figure 3.1.

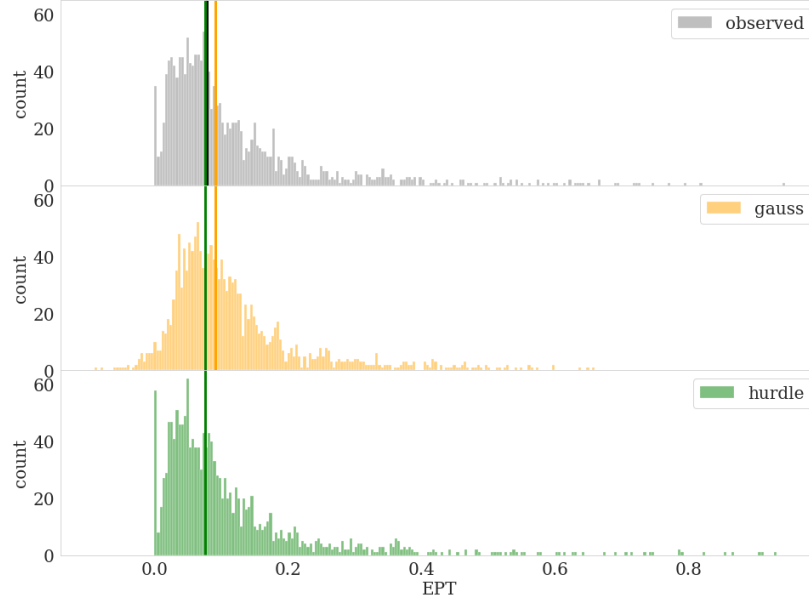


Figure 3.1: *Posterior predictive check: examples of replicated distributions for a single language-pair, created with Gaussian (orange) and Hurdle (green) models compared to the observed distribution (grey). The vertical line is the median of the distribution, with the respective colour. All the bins in the histogram have the same width which makes it easier to visualise the differences. EPT values have been rescaled for data protection reasons.*

Upon inspection, we notice that, as suspected, the Gaussian model suffers from some modelling weaknesses: its likelihood distribution (and so its posterior predictive one) is not bounded by zero, so it allows for negative EPT, which is wrong by definition (see Eq. 2.1), and it actually exhibits this behaviour. In spite of this, for some generated data, the Gaussian model shows a median (the orange vertical line in Figure 3.1) above that of the observed data.

On the other hand, the Hurdle model seems to replicate the presence of perfect translations, producing a peak at $EPT = 0$ quite similar to the one in the observed distribution. Additionally, it accounts for the presence of extremely large values of EPT in the observed data, while the Gaussian model generates essentially no observations above 0.7 relative EPT.

We have avoided a quantitative assessment of the ratio of zeros generated in the case of the Gaussian model, since visually examining the posterior distributions alone confirms there is no sharp peak at $EPT = 0$. This is, of course, expected given the structure of the model (see 2.2.1): it is free to produce continuous values of EPT without any constraints, and so it does not provide inflation of zeros.

In the case of the Hurdle model, however, we were interested in how well it can fit the data by adjusting the parameter π , which is the probability that an EPT observation is exactly zero. As our metric of agreement between the generated and the observed data in this aspect we took the absolute difference between the proportion of zeros generated by the Hurdle model ($N_z^{(rep)}$) and the proportion of zeros present in the corresponding real data ($N_z^{(real)}$). Finally, we averaged the metric over the 1000 replications mentioned above, giving us the mean absolute error:

$$MAE(Z_{ratio}) = \frac{1}{1000} \sum_{rep=1}^{1000} |Z_{ratio}^{(rep)} - Z_{ratio}^{(real)}| \quad (3.2)$$

where $Z_{ratio} = \frac{N_z}{N_j}$. In Figure 3.2 we plot $MAE(Z_{ratio})$, language by language, and, for the sake of visualisation, we sort languages from the best-MAE language (language_0, $MAE(Z_{ratio}) < 0.1\%$), to the worst-MAE language, language_56, for which the model scored an absolute error in percentage of zeros of about 2.7%.

We now compare the generated and observed data on the Kullback-Leibler divergence between their empirical distributions, calculated as

$$KL(p, q) = - \sum_x p(x) \log \left(\frac{p(x)}{q(x)} \right),$$

where we plug in the observed sample for $p(x)$ and the generated one for $q(x)$. It

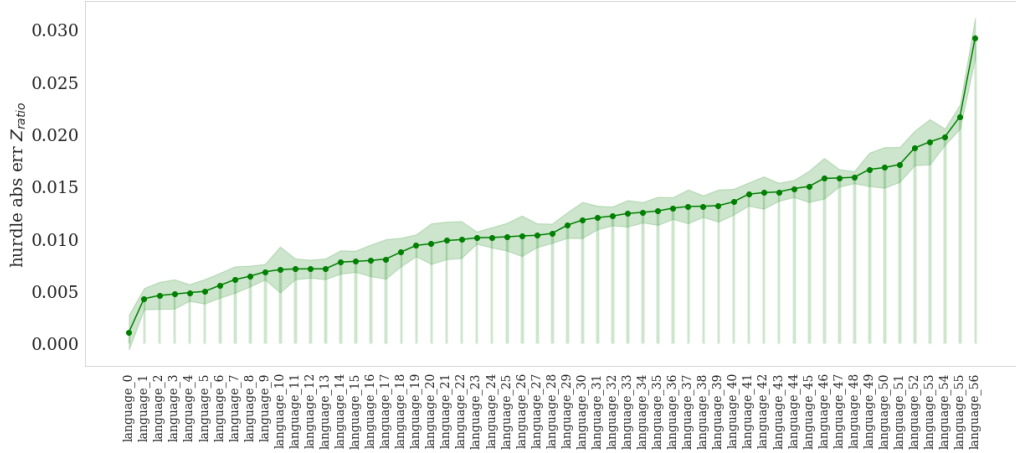


Figure 3.2: MAE of the Z_{ratio} , representing the deviation from the ratio of zeros in the real EPT distribution. The shaded area is the standard deviation of the absolute errors, and the languages on the x-axis are sorted by MAE, ranging from a difference of less than 0.1% of zeros on the left to about 2.7% on the right. Languages have been anonymised for data protection reasons.

represents the average extra information (in nats) needed to encode q assuming that the data is actually distributed as p .

The Hurdle model exhibits lower KL than the Gaussian model, as can be observed in Figure 3.3, which shows the average (over the 1000 replications) ratio between the two divergences:

$$\frac{KL_H}{KL_G} = \frac{KL(R, H)}{KL(R, G)}$$

with R , H , and G being the empirical probability distributions (over the EPT values) of the real data, and the samples generated by the Hurdle and Gaussian models, respectively. For most languages, the ratio between KL_H and KL_G is lower than 1 (for which we have a tie - the horizontal line in the plot), indicating that the Hurdle model generates from a distribution that matches the observed data distribution more closely than the Gaussian model¹. To present a clearer

¹In Figure 3.3 some languages are missing for computational reasons. Sometimes it was hard to find the right binning to build a proper normalised probability distribution over the replicated EPT values provided by the models.

visual indication of the relative deviations of the proportions of the two distances from equality (i.e., 1), we present the plot in log y-scale.

To summarise, we use the posterior predictive check to find any discrepancies in the generated data with respect to the real data and conclude that the Hurdle model reproduces the observed distribution well, matching it both in the inflation of zeros and in the shape of the rest of the distribution. On the other hand, the Gaussian model suffers from some issues, like a non-zero probability for negative EPT, as was apparent at the time of its conception. Nevertheless, the Gaussian model fits the data reasonably well, since we can observe that the discrepancies between its results and the ones produced by the Hurdle model are not that strong, even overperforming it for some languages - i.e. the ones with the KL ratio above 1 (see Figure 3.3).

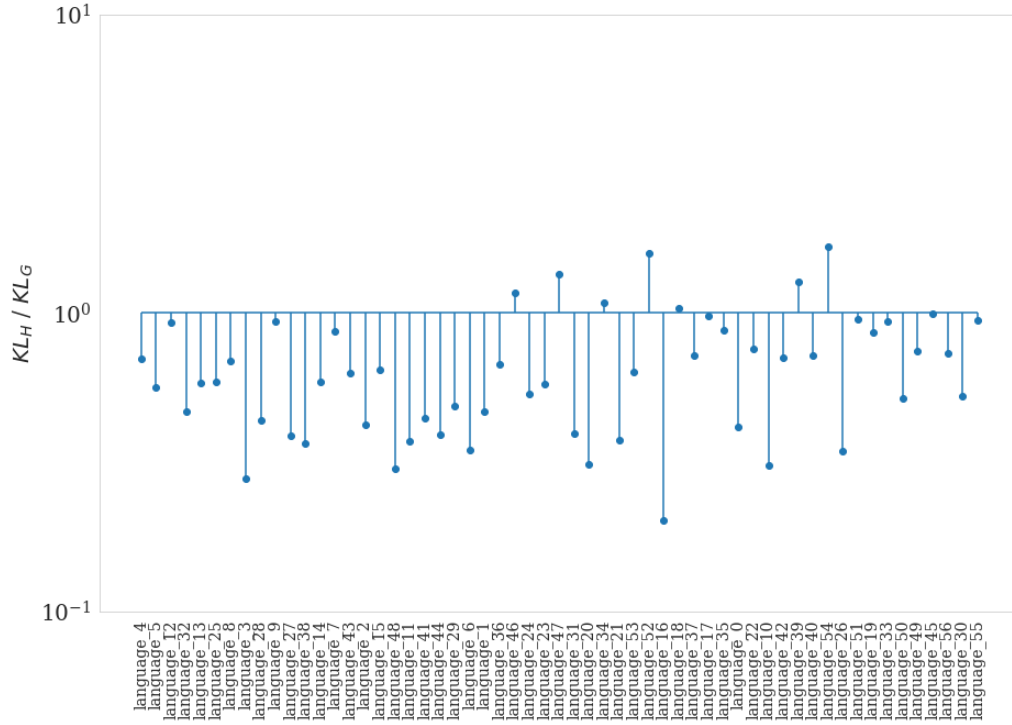


Figure 3.3: The ratio $\frac{KL_H}{KL_G}$ of the Kullback-Leibler divergences between the observed distribution of EPT and the ones reproduced by the models: for most of the languages, the divergence of the Gaussian model (KL_G) is greater than that of the Hurdle model (KL_H), suggesting that for those languages, the Hurdle model fits the distribution better than the Gaussian one. The y-axis is log-scale and the horizontal line represents a "tie" between the two models.

3.1.2 Skills retrieval

The team of translators used by Translated is composed of expert translators, all native speakers of the target language, and specialised in marketing texts, the topic selected for the extraction of our dataset. Language teams leads are selected based on their experience by Translated's project management staff. We will denote such individuals as of skill level 1 (L1). Each is designated a deputy, of skill level 2 (L2). In our study, this selection process provides independent information about translators' skills. It can, therefore, be used to validate our models.

Since most of the translators involved in the experiment act as both translator

<i>group</i>	<i>rev only</i>	<i>tra only</i>	<i>both</i>
skilled (L1)	38%	0	62%
skilled bkp (L2)	13%	0	87%
unknown skill	24%	12%	64%

Table 3.1: *Dataset details for the percentage of linguists per role (translator/reviewer) and skill level*

and reviewer (cf. Table 3.1), we can check how the above translating skill assignments reflect on the linguist characteristics when acting in the role of reviewer. In this validation, we use only the Hurdle model which takes into account the contribution of both reviewers and translators, while the Gaussian model does not.

We split the entire population of translators and reviewers into:

- Level 1: they are the most reliable translators as evaluated by Translated, and the most trusted reviewers. In the following figures, they will be shown in green on the plots.
- Level 2: they are supposed to be more skilled than the average translator on a team, but sometimes they are less certified than the L1. As they act as deputies, we also refer to them as "skilled bkp" (backups). They will be represented in orange.
- Unknown skill: the rest of the population of translators and reviewers. They will be represented in grey.
 - translator only: the translators that are in the "unknown skill" category and are not used as reviewers. Represented in light blue.
 - reviewer only: the reviewers that are in the "unknown skill" category and are not used as translators. Represented in pink.

In the dataset, these linguists can be present in both roles - translator or

reviewer - or in just one of them. The breakdown of roles and skill levels is presented in Table 3.1. On the plots, the "Unknown skill" category alone is broken down further, with pink representing the reviewer-only category and light blue the translator-only category; both the skilled translators - L1 - and the skilled bkp - L2 - will be represented with green and orange respectively, regardless of the role they usually assume.

As a reminder, the Hurdle model assumes that the final EPT score is composed of three multiplicative contributions:

- d_j - the difficulty of the job j , distributed as $d_j \sim \mathcal{HL}(\pi_j \mu_j, \sigma_j)$
- ϵ_t - the translator error propensity, distributed as $\epsilon_t \sim \mathcal{HL}(\pi_t, \mu_t, \sigma_t)$
- β_r - the reviewer bias, distributed as $\beta_r \sim \mathcal{HL}(\pi_r, \mu_r, \sigma_r)$

where the distribution $\mathcal{HL}(\pi, \mu, \sigma)$ has been defined in Eq. 2.8. Given that the final EPT is a product of three random variables d_l, ϵ_t, β_r , it is clear that both translators and reviewers can apply a contraction (or dilation) to the underlying job difficulty. Such an effect is determined by the parameters of the distribution ϵ_t for the translators and β_r for the reviewers. We will thus use the two triplets of parameters - $\{\pi, \mu, \sigma\}$ for ϵ_t and β_r - to characterise the properties of the linguists in the two roles.

Results

We present the results of the analysis with two kinds of plots, both showing the parameters for the 4 groups described above. In Figures 3.4 and 3.6, we show the 95% CI² of the parameters for the translator and reviewer groups respectively, while Figures 3.5 and 3.7 show scatter plots comparing pairs of parameters, which

²All the CI have been calculated with the bootstrap technique.

enables us to characterise both translators and reviewers, respectively. For comparison, we also present the CI of the raw EPT data that has been averaged first over the translators and then over the sample of such averages. The results are first examined separately for the translator and reviewer roles, and, finally, the characteristics of those acting in both roles are presented.

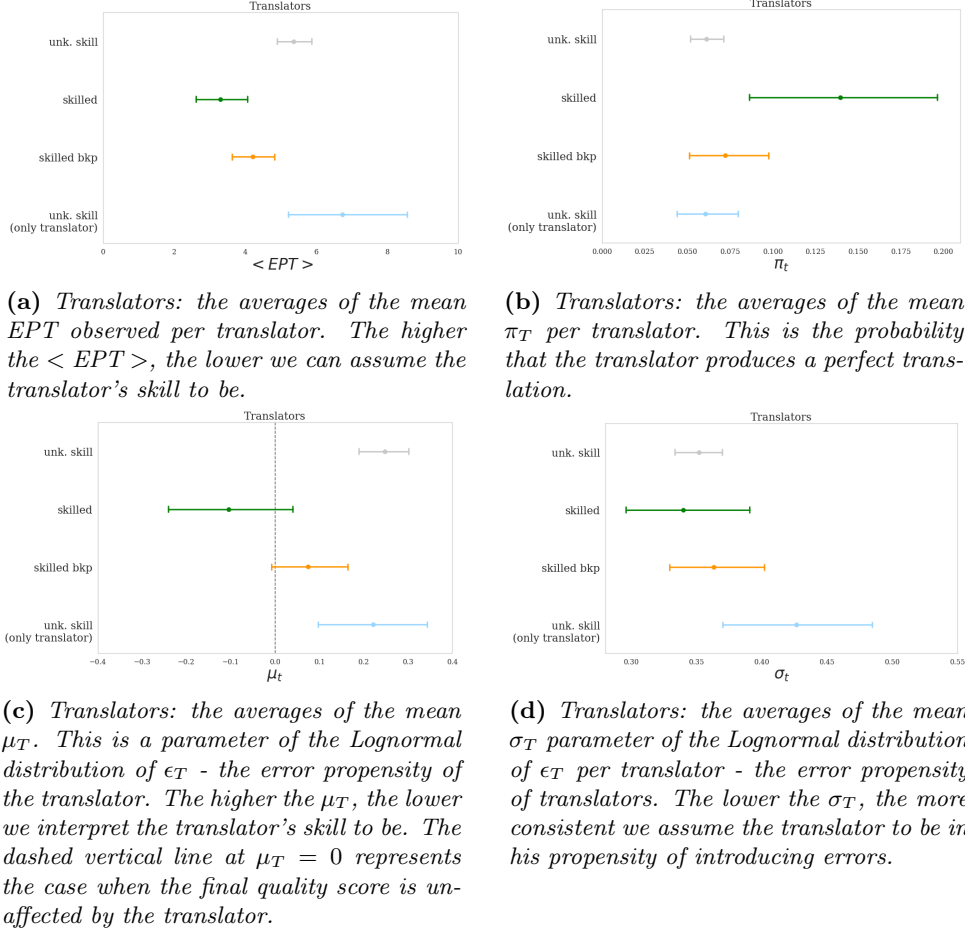


Figure 3.4: Translator assessment: 95% confidence intervals for every parameter of the Hurdle model associated with the translators. The population of translators has been split into the language leads (green), their backups (orange), and the rest: grey for the ones who act both as translators and reviewers at Translated, and light blue for the ones who don't appear in the dataset as reviewers.

Translators In our Hurdle model (used here for the inference of the skill characteristics), the contribution made by the translator to the final quality EPT score

of a translation job is represented by the parameter we call the translator error propensity ϵ_t :

$$\epsilon_t \sim \mathcal{HL}(\pi_t, \mu_t, \sigma_t) \quad (3.3)$$

With this, we can give the following interpretation to the parameters that rule such distribution of ϵ :

- π_t : the translator's contribution to the probability that a translated job will be a perfect translation ($EPT > 0$). The higher the π_t , the more skilled the translator.
- μ_t : the mean of the Gaussian distribution of $\log(\epsilon_t)$. The lower the μ_t , the higher we assume the translator's skill to be, reducing the probability of obtaining a high EPT value in the TQA process.
- σ_t : the standard deviation of the Gaussian distribution described above. The lower the σ_t , the more consistent the translator is in his error propensity.

A good translator should, therefore, have lower values of μ_t and σ_t and higher values of π_t when compared to the average translator.

Bootstrap Confidence Intervals

If we want to find the Confidence Interval of the mean of a population, given a set of observations, we can use the Bootstrap technique ([Efron, 1979](#)): we create multiple resamples (with replacement) of our set of observations, for which we can easily calculate the mean. Repeating this operation multiple times, the set of means will approach a normal distribution (central limit theorem) that we can use to find our confidence interval. The big advantage of the bootstrap technique is that it is a non-parametric method, so we can avoid to assume that our underlying population is normally distributed.

In Figure 3.4, we grouped the translators according to the populations described above and calculated the mean value of the parameters of interest, together with its 95% CI as found by bootstrapping the sub-population of translators.

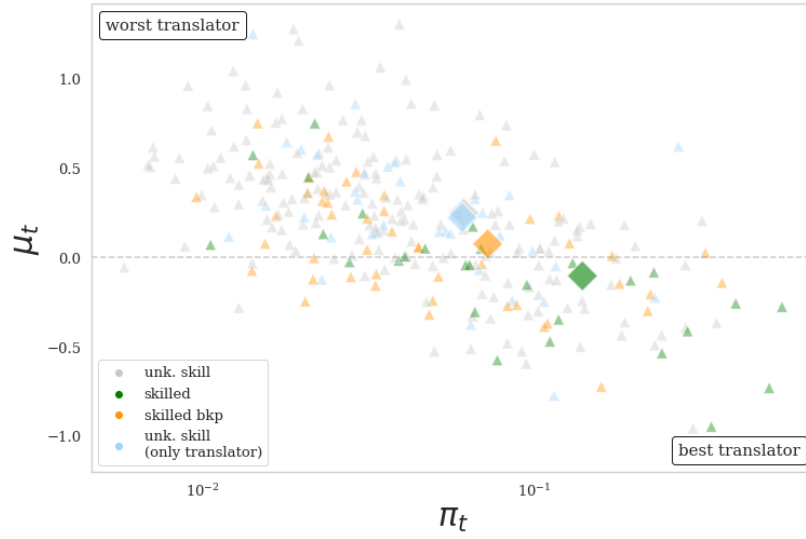
The parameter μ_t (3.4c) confirms the intuition that the most skilled translators - L1 (green) - should have the lowest values: it is more likely that the final translation will have a higher quality (low value of EPT) when translated by them (even excluding the separate process that governs the 0 EPT translations). Using μ_t as the separating criterion, we can distinguish between groups corresponding to L1 (green) and L2 (orange) translators, and considerably more so between L1 and the unknown skill groups (grey and light blue colours). This separation corresponds to the group averages of the raw EPT (c.f. Figure 3.4a), despite those incorporating the joint contributions of job, translator, and reviewer characteristics.

Examining the values of π_t inferred (Figure 3.5a) we observe that the L1 are well separated from the ones with unknown skill. This, however, does not hold for L2, which leads us to suspect that some of the backups to the Level 1 translators might not have the same level of expertise.

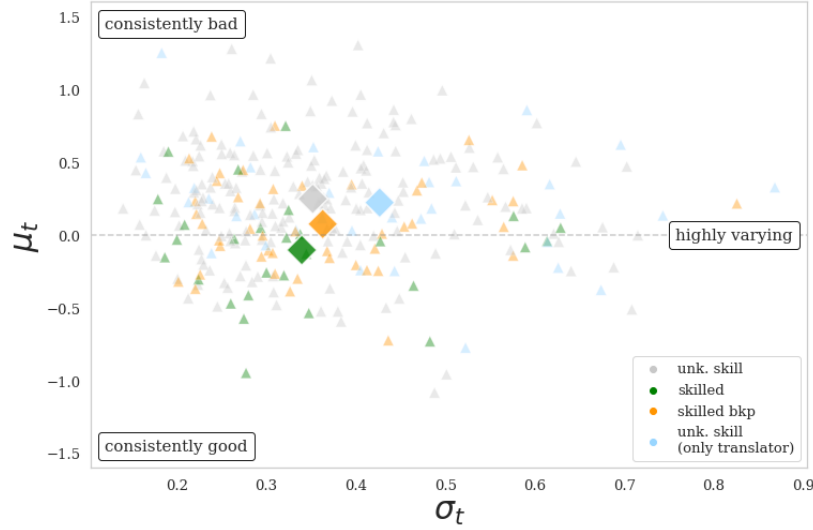
Figure 3.5a shows the location of any individual translator (triangles coloured by their certified skill assignment) in the plane (π_t, μ_t) . We display π_t in log scale for the sake of easier visualisation. The larger squares are the mean values for the respective population (i.e., their centroids) and the dashed horizontal line separates the translators with $\mu_t < 0$ (good) from the ones with positive μ_t (bad). While the orange triangles (L2) are uniformly interspersed within the bounds delimited by the unknown skill translators (grey triangles), the L1 occupy mostly the lower right area of the plane, exhibiting a high probability of producing perfect (horizontal direction) or low-error (vertical direction) translations. If we look at the population centroids (squares), we can indeed observe that the L1 group (green

square) is located towards the lower-right corner of the plane, meaning that, on average, the model assigns higher values of π_t and lower values of μ_t to the L1 group. The unknown skill groups (light blue and grey triangles) instead exhibit the opposite behaviour: most of them have higher μ_t and lower π_t values.

Finally, the confidence intervals on the σ_t estimates do not allow us to state (with a 95% confidence level) that there exists an observable difference between the populations, except in the case of the subgroup within the unknown skill grouping which corresponds to those translators that also assumed the role of reviewer (grey triangles). This looks more consistent (lower σ_t) in its effect than the subgroup comprised of individuals who were translators only (light blue triangles). The relationship between the inferred σ_t and μ_t values is displayed in Figure 3.5b: while there exists a separation between population centroids (squares) in the vertical direction, this does not hold in the horizontal direction, where the separation is minimal, excepting the translator-only subgroup (light blue).



(a) Translators: the higher π_T , the higher the probability of producing perfect translations. The lower μ_T , the lower the probability of making errors even when the translation is not perfect.



(b) Translators: the lower σ_T , the more consistent the effect of the translator. The lower the μ_T the higher the probability of making errors, when the translation is not perfect.

Figure 3.5: Translator assessment: every triangle corresponds to a translator with the colour indicating their certified skill level: green for L1 and orange for L2, while the not-skilled are split between those that act as both reviewers and translators (grey) and those that are translators only (light blue). The squares are the centroids of the respective populations. The parameters inferred by the Hurdle model reasonably reflect external information about translators' skills. The model can reasonably retrieve translators' skills, assessing the L1 as the translators who produce fewer errors and more perfect translations - lower right side of (a) - and are also the most consistent - lower left side of (b).

Reviewers Similarly to the case for translators, in the Hurdle model the effect of the reviewers is also modelled by introducing a multiplicative error into the real translation quality through a parameter that we call *reviewer bias* β_r :

$$\beta_r \sim \mathcal{HL}(\pi_r, \mu_r, \sigma_r) \quad (3.4)$$

In the case of reviewers, the interpretation of the parameters that govern the distribution of β_r is the following:

- π_r : the reviewer's contribution to the probability that a translated text will be a perfect translation ($\text{EPT} = 0$). Here, high values of π_r also have a special interpretation: they correspond to the tendency of the reviewer to not pay too much attention to the review of a particular segment or to have a bias towards low severity error annotations³. In other words, they are prone to producing perfect translations, regardless of the skill of the translator involved or of the job's intrinsic difficulty.
- μ_r : the mean of the Gaussian distribution of $\log(\beta_r)$: when the reviewer finds some errors, a low value of μ_r means that they are generally *lenient*, reducing the probability of the translation scoring a high EPT value in the TQA process.
- σ_r : the standard deviation of the Gaussian distribution of μ_r . The lower the σ_r , the more consistent in their effect will the reviewer be.

As in the case with translators, we also show both the 95% CI of the means of the parameter estimates (Figure 3.6) and the location of any individual reviewer in the planes (π_r, μ_r) and (σ_r, μ_r) (Figure 3.7).

³We remind the reader that the severities *Preferential* and *Repetition* have zero weights in the weighted sum resulting in the final EPT score.

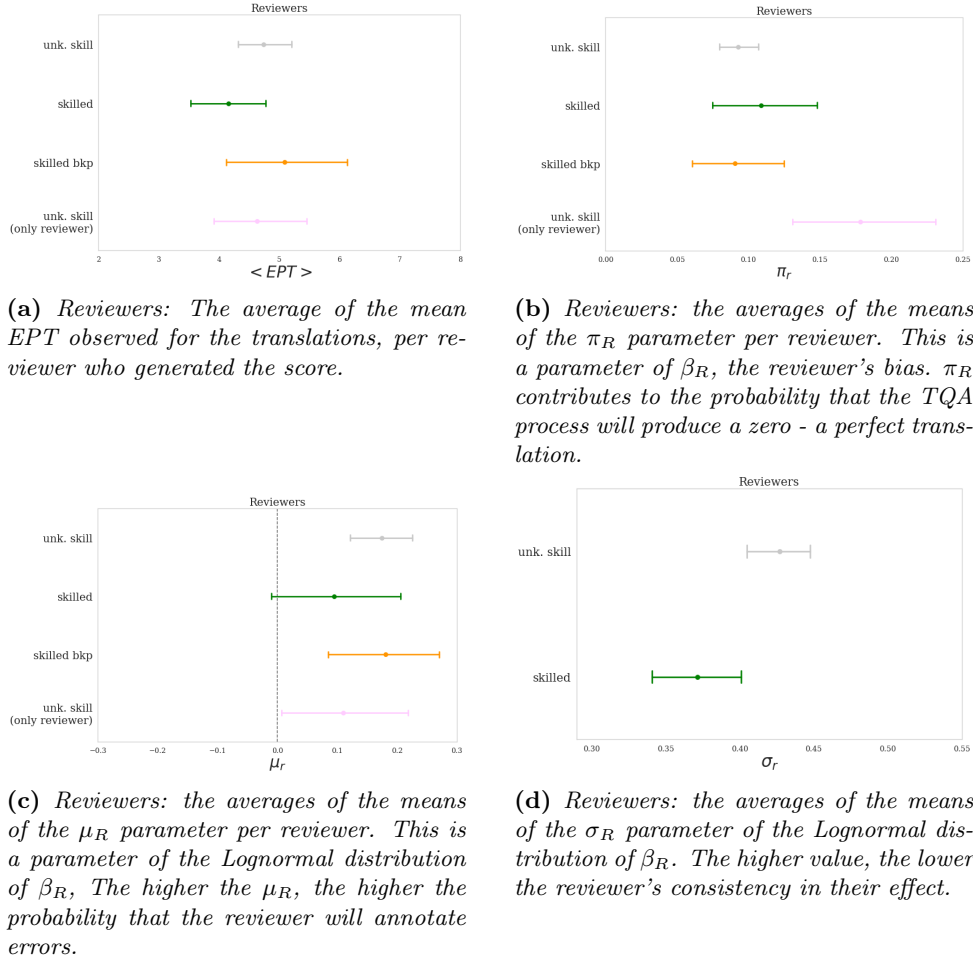
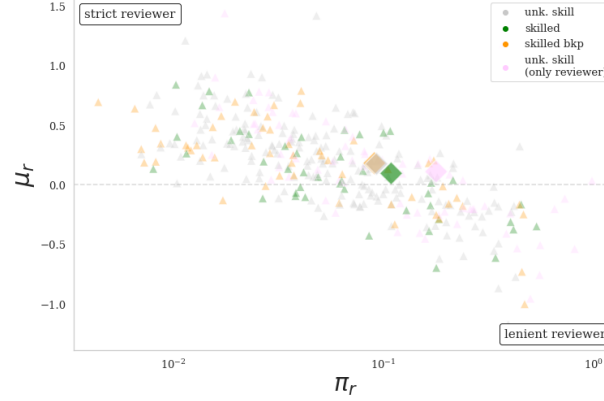
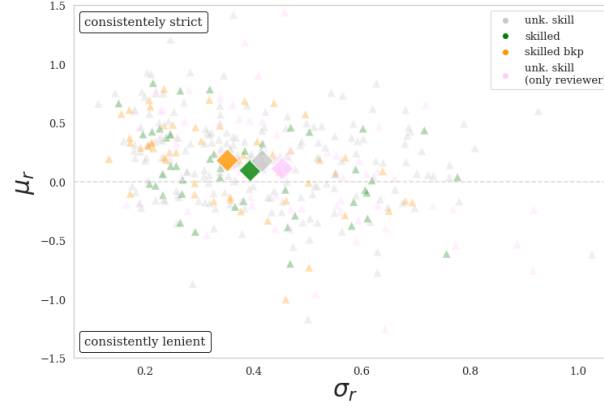


Figure 3.6: Reviewers assessment: 95% confidence intervals for the means of every parameter of the Hurdle model belonging to reviewers. The population of reviewers has been split, separating the language leads (green), their backups (orange), and the others (grey for the ones that serve both as reviewers and translators at Translated, and light blue for the ones who appear in the dataset as reviewers only.)

While the external certification of the skills of the L1 group seems to be reflected in the model estimates of their translation skills, this doesn't translate to the case when they serve as reviewers. We cannot find a clear separation between the certified experts (L1 and L2) and the rest, at least so far as our interpretation of the parameters would suggest. This can be seen also in the average mean EPT shown in Figure 3.6a: the skilled linguists (the certification pertains to their translation skills and not necessarily their skill as reviewers) do not exhibit a significant difference in terms of average μ_r (Figure 3.6c) when



(a) *Reviewers: The higher the π_R , the more frequently the reviewer will evaluate a translation as perfect. The higher the μ_R , the higher the chance that if the reviewer decides to annotate errors, they will find many. The pink population of those of unknown skill appears to be the least consistent in their effect, while the L2 population appears to be the strictest.*



(b) *Reviewers: In the upper-left area of the plane lie the reviewers with a higher probability of annotating more errors (due to high μ_R) while having a consistent behaviour (related to low σ_R). The L1 and L2 groups are concentrated on the left side of the plane, while in the vertical direction the L2 group and the group of linguists of unknown skill who act as both reviewers and translators seem to be the highest.*

Figure 3.7: *Reviewers assessment: every triangle represents a single reviewer, and the colour indicates their certified skill level: green for L1 and orange for L2, while those of unknown skill are split between those that serve both as reviewers and translators (grey), and those that are reviewers only (pink). In the role of reviewers, the difference in the population centroids of the inferred parameters is not so significant: we can barely distinguish the L2 as stricter than the L1, for which we do not have a reasonable explanation. Looking at σ_r , the populations of skilled and skilled bkp seem to be more consistent in their effect than the others, especially when comparing the L2 to those of unknown skill that are reviewers only.*

compared to those of unknown skill.

The mean of parameter π_r (Figure 3.6b) seems to be higher for the group of reviewers of unknown skill that don't act as translators in our dataset (coloured in pink). This population is composed mainly of linguists who are labelled by Translated as "preferentially reviewer"; the automatic system in charge of assigning different roles to linguists working on a translation job will subsequently prefer such linguists when selecting a reviewer for the job. An aspect to keep in mind is that this label is assigned considering joint feedback of Translated employees and the L1 linguists, therefore such assignments could be prone to human bias: Translated employees could be influenced by the number of times they've observed the reviewer evaluating a translation as perfect and could, therefore, be influenced by this in assigning the label "preferentially reviewer" to such "friendly reviewers". Such selection bias is naturally reflected in higher values of π_r for this subgroup. This could also be the reason why the reviewers-only (pink) group in Figure 3.6b is separated from the other reviewers of unknown skill that also act in the role of translators (grey triangles). This uncertainty in the groups' separation in terms of π_r is also fairly evident in Figure 3.7a where their means are not clearly differentiated, except for unknown skill only-reviewers (pink) along the horizontal π_r axis.

Nonetheless, the L1 and L2 reviewers (coloured green and orange) lie mostly on the left side of the plane (σ_r, μ_r) (Figure 3.7b), meaning that they are more consistent (i.e. have lower variance) than the uncertified linguists when reviewing. Although we do not expect to observe a significant separation in terms of π_r and μ_r between the populations of reviewers examined (we have no reason to suspect a priori that a good translator is a strict or lenient reviewer), we think their values of σ_r should reflect somehow the expertise of a linguist (even if his expertise is certified just for the role of translator): our guess is that a linguist will stabilise

their style as his expertise increases, and will be less prone to inconsistency in reviewing the translations produced by others.

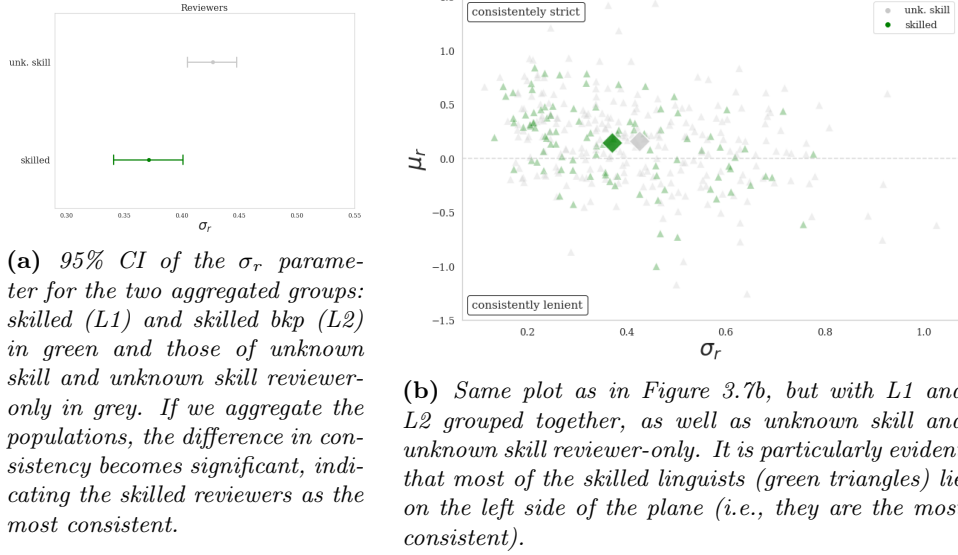


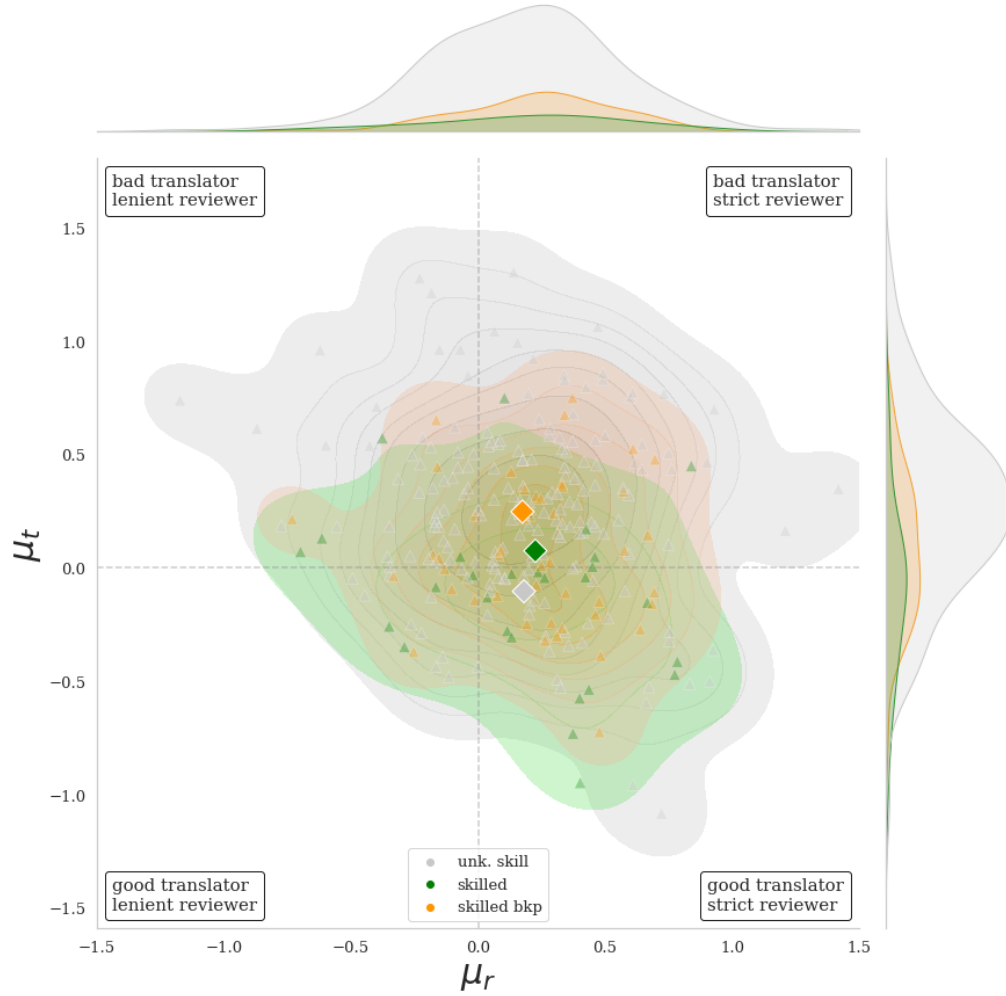
Figure 3.8: Reviewers assessment - skills retrieval: aggregating the skilled groups (L1 and L2) and those of unknown skill (reviewer-only and acting in both roles), the separation between the two groups becomes evident.

We noticed that the reviewer-only category and that of unknown skill show a higher average σ_r value compared to the average of the combined group of skilled and skilled bkp (as can be seen from the group means represented by the squares in Figure 3.7b). We thus compared simultaneously the CI intervals shown in Figure 3.6d, and the scatter plot in Figure 3.7b, this time aggregating into two main groups: skilled (L1 - green - and L2 - orange - together) vs. those of unknown skill (grey and pink triangles). The result shown in Figure 3.8 confirms our initial intuition by inferring that the population of skilled reviewers is more consistent at a statistically significant level (lower σ_r) - with a 95% CI of $CI^{\text{sk}} = [0.34, 0.40]$ - than that of unknown skill, for which the CI is $CI^{\overline{\text{sk}}} = [0.41, 0.45]$.

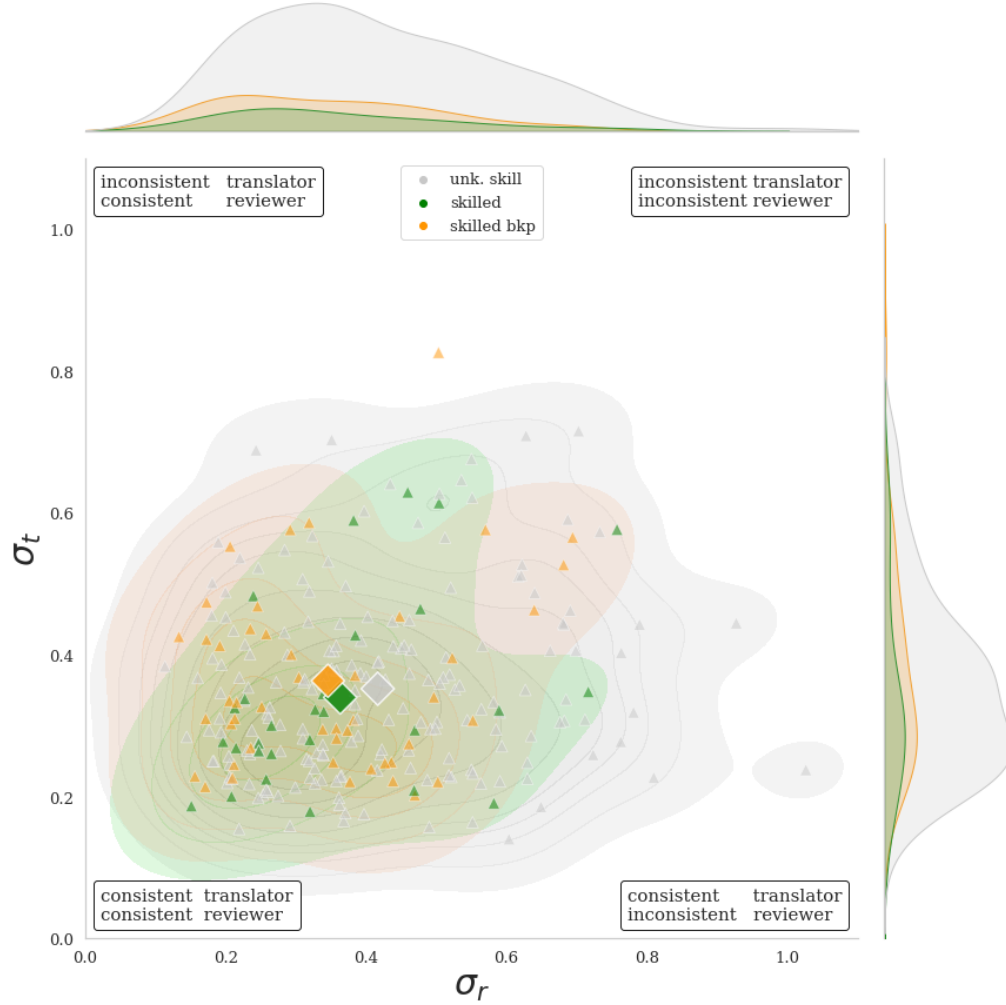
Cross-features Finally, we examine the pairs of parameters (μ_r, μ_t) for the subgroup of linguists who act as both translators and reviewers in our dataset,

aggregating L1 and L2 into one group and those of unknown skill into another (green, orange, and grey triangles in Figure 3.9, respectively). We notice that the L1 group members cluster on the lower-left side of the (σ_r, σ_t) plane (Figure 3.9b), showing a higher consistency both as reviewers and as translators when compared with those of unknown skill. Moreover, looking at the Figure 3.9a, we can clearly see that while the populations of linguists are separated well as translators (in the vertical direction), we cannot say the same for their separation as reviewers (the horizontal direction), meaning that the skill certification is reflected in the behaviour as a translator, but not necessarily so when serving as a reviewer. Therefore, we can claim that assuming that "good translator" implies "good reviewer" is a dangerous assumption to make. Moreover, examining the same plot (Fig. 3.9a), we notice that the lower-left part of the plane (μ_r, μ_t) is quite empty when compared to the other quadrants, which would imply that if a linguist is very good at translating (located in the lower part of the plane) they are also likely to be a strict reviewer (located on the right), and we can reasonably say that we can reasonably say this relationship holds in both directions.

Given that the groups are not strongly separated on either of the plots in Figure 3.9, we can find a considerable portion of linguists, who, despite not being certified by Translated as skilled, actually exhibit the behaviour of a skilled linguist as indicated by the values of their model parameters. It would, therefore, be of interest to check their performance case by case and to, if the findings hold up, use the model as a quantitative metric in choosing whom to certify as a skilled linguist. The model could also help in adjusting the number of translations that Translated assigns to such linguists. When used for this purpose, the model could help improve the overall quality of delivered translations by maximising the number of translations provided by high-quality translators.



(a) Linguists serving both as reviewers (horizontal direction) and translators (vertical direction) - average effect: examining both the coordinates of the group means (squares) and the areas covered by the KDE of their distributions, we can see that while the different groups show a significant separation in terms of μ_t (vertically), they are less clearly separated in their μ_r values (horizontally).



(b) Linguists serving both as reviewers (horizontal direction) and translators (vertical direction) - effect consistency: the L1 and L2 groups cluster in the lower-left area of the plane, as shown by the mean coordinates of the groups, represented by the squares. Those of unknown skill are dispersed more widely in the plane.

Figure 3.9: Linguists assessment - cross-evaluation as reviewers and translators: every triangle is a linguist. The colour indicates their certified skill and the squares are the mean coordinates of the linguists' groupings. The coloured area is the two-dimensional KDE of the population. The horizontal direction gives insights into the linguist's quality-assessment (i.e., review) behaviour, while the vertical one is dedicated to their translating skills.

Pages 89-93 hidden for data protection reasons

Chapter 4

Conclusions and future works

Translation Quality Assessment is the evaluation of the overall quality of a translation, and can be influenced by many factors: the source texts to be translated do not all share the same difficulty, and, even for the same text, such difficulty can vary, depending on which is the chosen target language. The final translation quality will reflect the skill of the translator who produced it, and its assessment will be further affected by the stylistic features of the reviewer, who can be strict, lenient, and also more or less consistent with his/her judgements, as well as the translator can be more or less consistent in his/her performance.

Here we have taken into account three main features contributing to the quality assessment, namely the job's difficulty, translator's skills, and reviewer's characteristics. We treated the final observed data about the quality as one measurement of a real latent quality-score and we considered the contributes of the translation-job intrinsic difficulty, and the one of both the human linguists involved in the process, as errors affecting such measurements. We created two probabilistic models that differ in the choice of the parameters and in the likelihood function that links such parameters to real data. We designated as "Gaussian" the first one since it considers EPT score (our quality-score data) as a

random variable normally distributed, where the mean corresponds to real quality of the translation, shifted by the reviewer's bias, while the variance represents the reviewer's consistency. The "Hurdle" model, instead, takes its name from the probability distribution through which it faces the characteristics of our dataset of being zero-inflated. EPT here is a random variable that is distributed according to a mixture probability distribution, composed of a Bernoulli and a Lognormal, the first one determining whether the EPT is a zero or it is lognormally distributed. All the parameters of the Hurdle likelihood have an independent contribution coming from the texts, translators, and reviewers, so that we could split and interpret such contributions. The absence of a ground-truth concerning the characteristics that we want to infer, i.e. the real quality-score and the linguists' features, makes it challenging to validate the performance of our models.

Given the availability of only one quality measurement per translation job, we discarded the idea of using a subset of the real data to train the model and the remaining part of the data to test the predictability, hence we used indirect methods to assess the validity of our modelling. In particular, we considered the so-called "posterior predictive check": we checked the ability of the models to reproduce the overall distribution of real data, once having optimised the parameters. The Hurdle model shows to fit better the data, since it can reasonably reproduce the inflation of zeros of the real data, and it prohibits negative EPT values (positive definite) to be produced, while the Gaussian model does not set this constraint. Both the models show to reasonably fit the positive values of the data, even if the Hurdle model produces a probability distribution over the EPT values with a shorter distance from the real one.

As an external (to the model) validation of the information provided by the Hurdle model, we challenged the model in the task of recognising the translators who are certified by Translated to be skilled. The Hurdle model provides infor-

mation about translators and reviewers to which we can give an interpretation in terms of skills and so, we could check if these information correlated with the translators' certification of expertise. The certified expertise shows to be actually reflected in the model inferred parameters concerning linguists in the role of translators, while we showed that such expertise does not directly reflect on linguist's behaviour when playing the role of reviewer. We show, though, that the most valid translators usually are also strict reviewers, and so it is not true that the best translators are also the ones with an unbiased behaviour (that would be desirable in order to have a fair assessment). It can be seen, instead that the higher is the expertise of a translator, the more evident is his consistency, both when translating and when reviewing.

Finally, we designed and engaged an experiment to create new data: a second revision for a set of translations, so that we could have two quality-measurements (provided by two different reviewers) for that set of translations. The quality-scores produced revealed high disagreement among the reviewers, resulting in a poor correlation between the two measurements of the same translation quality. Due to this poor correlation, the estimate of the specific quality-score for the single translation job revealed to be still hard, with this kind of modelling and a single observation in the training set. Both our probabilistic models, indeed, could not beat the predictions made by a trivial "fixed predictor", built to provide always the same score for every translation in the test set: just using as prediction for every quality-score the mean EPT score of the first set of reviews (the training set), the average error, when compared to real data, is lower than the one scored by the models' predictions. Despite that, even if the models cannot manage to make a useful point estimate of the individual quality-score, the Hurdle model proved to be able to reasonably predict the mutual relationship between two different reviewers working on the same set of translations, both in terms of bias and in

terms of number of translations judged as perfect (with no mistakes).

Can human evaluation be considered a golden standard?

In our research we evaluated the process of Translation Quality Assessment in a specific and new scenario, in which all the linguists involved, belonging to both the two groups of translators and reviewers (not mutually-exclusive) were translation experts. In the literature concerning translation studies, researchers usually face the problem of estimating the reliability of the evaluators (reviewers) among a crowd of non-experts, in the framework of crowdsourcing, for instance, in which a requester pays a crowd of non experts to accomplish the quality assessment task (Bojar et al., 2013; Callison-Burch et al., 2008; Denkowski and Lavie, 2010). In our work, instead, we questioned the reliability of expert translators too, that is often taken for granted. When evaluating the quality for translations produced by a Machine Translation system, indeed, such quality is determined by comparing the automatic system outcomes with human-produced reference translations, a set of sentences translated by human translators, that is usually considered to be a golden standard, without further investigations. We claim that the subjective nature of the translation process, and so the one of its quality assessment, plays an important role, also among the professional translators, who showed to be actually biased as reviewers (we have seen that especially the best translators, are usually the strictest), or inconsistent in their evaluations. For this reason, we think that if one wants to get a fair and consistent estimate of the quality of a translation, it is suggested to build a network of evaluations, provided by different reviewers, so that it could be possible to remove the undesired contributions of the human subjectivity, inherent in the reviewers, and to weight more the objective part of the evaluation.

Nonetheless, our research proved that bayesian networks can be an efficient

tool also in the one-review setting, when the task to address is just the one of retrieving translators and reviewers skills. In every situation in which one is interested in understand the real competence of a translator, and he wants to track his performance, therefore, it could be efficient to apply the methods described in this research project, with an evident gain in terms of time and money invested when compared with the usage of multiple reviews on the same translation job, that usually are created ad hoc for the purpose of reliability assessment. In the case of Translated Srl, the single review dataset is naturally built up in the course of the working time, and now they are taking advantage of this dataset by integrating our model in their automatic system that ranks the translators and picks the best available one for the requested translation job. It will be interesting, in the next future, to control the quality trend and to assess the impact of our new ranking system on the overall quality the translations provided.

Future directions

As we mentioned in the section 2.1, Translated invests money and resources to monthly monitor the quality trend of the translations in a language-pair, by asking a second revision pass consecutive to the first one. The first reviewer can apply his edits, and so the second one judges a different translation version, and so the two judgements actually evaluate different objects. This second pass review is always conducted by the best certified translator in the language-pair team (that we called LL in the course of the thesis) and we claim that the usage of just one reviewer, even if with certified expertise, is very likely to introduce a new source of error in the estimation, since we recall that one of our finding is that a good translator is not always a impartial and consistent reviewer.

With that said, we claim that it is more likely to get an unbiased judgement about the overall quality of a batch of translations by reviewing it multiple times

in parallel, rather than with a chain of reviews.

To conclude, we remark that here we are neglecting a feature of our EPT data, that is created by the EPT definition: being it the ratio of errors over the number of words in a source text, clearly there is lower bound, since the text evaluated have a typical maximum length (about 3500 words). This is the effect that we called finite-size effect in the description of the dataset in chapter 2.

Taking explicitly into account this finite-size effect could significantly help the model to catch patterns in the data. For this reason, for the future directions we will investigate the improvements in the performance of the models when feeding them with the number of words of the text and setting constraints on that.

Furthermore, it will be interesting to create a prior on text difficulty by calculating the translation entropy of the source text, that we briefly presented in 1.1.4 as the degree of translation’s uncertainty estimated from the source text to be translated.

Part II

Relatedness estimate through Natural Language Processing

Acknowledgements

This second essay concerning Relatedness estimate through Natural Language Processing is published as a preprint paper ([Tacchella et al., 2021](#)) and it is currently under review-process. I want to thank all the co-authors of the paper, namely Andrea Tacchella, Andrea Zaccaria, Dario Mazzilli and Luciano Pietronero for their valuable contribution.

Abstract

Relatedness is measure of similarity between two human activities, in terms of inputs and contexts needed for their development. The estimate of Relatedness has become of great interest as a tool to inform policies and development strategies in governments, international organisation, and firms, under the idea that it is easier to move between related activities rather towards unrelated ones. In our paper *Relatedness in the Era of Machine Learning* ([Tacchella et al., 2021](#)), we focus on countries and we show that the standard, widespread approach of estimating Relatedness through the co-locations of activities (e.g. Product Space) generates measure of Relatedness that performs worse than trivial auto-correlation prediction strategies. We argue that this is due to the poor signal-to-noise ratio present in international trade data. In our paper we build new methods to measure relatedness, by either redefining the two-products correlations assessment, or by moving from two-products correlations to many-products correlations and finally by using a general network embedding technique, that we call Continuous Projection Space (CPS). Here, I will focus on a new definition of similarities between couples of products through textual description, my main contribution to the paper.

Chapter 5

Relatedness in Economic Complexity

The final results of the textual similarity method that I will describe here, have been integrated in our paper *Relatedness in the Era of Machine Learning* ([Tacchella et al., 2021](#)), and so, although I will focus here on such method, in the following I will present some general results of the paper, necessary for a complete understanding of the context in which we applied and validated the method.

The concept of Relatedness ([Hidalgo et al., 2018](#)) is a key element for both economic and social sciences, with applications ranging from smart specialisation strategies ([Balland et al., 2019](#)) to the study of countries development ([Hidalgo et al., 2007](#); [Zaccaria et al., 2014](#)), to recommender systems ([Lü et al., 2012](#)). Two human activities are considered to be related if they share a common set of capabilities that are needed for their development ([Teece et al., 1994](#)). The larger is the intersection of the needed capabilities, the stronger is the Relatedness. For example, industrial production of radars is closely related to the production of radio

broadcasting apparatus, but not to so much to crude oil refining. In recent years, driven by the increasing popularity and adoption of the Economic Complexity framework (Cristelli et al., 2013; Hidalgo, 2021; Hidalgo and Hausmann, 2009; Hidalgo et al., 2007; Sbardella et al., 2018; Tacchella et al., 2012, 2013; Zaccaria et al., 2014), Relatedness has been gaining importance in informing diversification or specialisation strategies across a wide range of policy making institutions such as the World Bank (Lin et al., 2020) and the European Commission (Pugliese and Tacchella, 2020; Pugliese et al., 2019). Under the idea that it is easier to develop activities that are similar to those already developed in a region, decision makers can rely on a quantitative tool to design policies that can be adapted to several strategic approaches (e.g. vertical or horizontal policies (van Dam and Frenken, 2020)). So a country, or a region, that is currently competitive in the production of radars can have development opportunities in radio broadcasting apparatus, as a consistent set of the needed capabilities is likely already present. Here we address the problem of providing a reliable quantification of the feasibility of this transition, that can be a valuable input for strategic decisions.

Clearly, a poor or inconsistent quantification of Relatedness, and therefore a wrong estimation of the feasibility of transitions to new industries, represents a huge risk for a policy maker basing its decisions on it. In fact, despite the potentially great impact that such ideas can have in shaping policies, an important point that needs to be addressed is the fact that there is no direct way to estimate Relatedness of real-world activities from first principles, i.e. through the quantification of common inputs. While some theoretical work has been done on the combinatorics of very specific or synthetic networks where the inputs layer is observed (such as letters-words or ingredient-recipes networks (Tacchella et al., 2016; Zaccaria et al., 2014)), in any real scenario involving human activities (e.g. industries, technologies, scientific research, etc.) we do not have access to any

"book of recipes", not even to the 'list of possible ingredients', that would allow for a principled computation of Relatedness in terms of shared input capabilities.

For this reason research has been focusing on how to recover an effective measure of Relatedness from location-activities data ([Hidalgo et al., 2007](#); [Neffke et al., 2011](#); [Teece et al., 1994](#); [Zaccaria et al., 2014](#)). The core idea is that if two activities require similar inputs they tend to co-occur within the same locations more than randomly ([Pugliese et al., 2019](#); [Saracco et al., 2017](#); [Teece et al., 1994](#)). Therefore, suitably normalised counts of co-occurrences can be used as a proxy for Relatedness. The problem that we address here is that this proxied estimation is inherently difficult for two reasons:

- The number of activities is very often much larger than the number of locations in which to count co-occurrences. This means that the correlation structure that emerges is mostly random. E.g. in the countries-products case, one would estimate a 5000*5000 (Harmonized System 6-digits) co-occurrence matrix of products out of approximately 170 observations (countries). It is possible to reduce the number of activities by aggregation ([Stojkoski et al., 2016](#); [Zaccaria et al., 2018](#)), but this typically leads to a very coarse grained Relatedness structure that is often trivial and of no practical interest. Also the opposite approach, namely to increase the number of locations by reducing granularity (i.e. going at the subnational level), is of little use due to the fact that harmonised regional-level data is often unavailable, and the number of observations needed to produce a good estimate of a 5000*5000 correlation matrix is easily in the tens or hundreds of thousands ([Bun et al., 2017](#)).
- Very often, locations-activities bipartite networks have a very strong nested structure ([Mariani et al., 2019](#)). As opposed to a block-diagonal structure,

that would immediately lead to a definition of sector-communities, in these networks the Relatedness signal is of second order with respect to the drive towards diversification that generates the nested structure.

The basic idea of counting co-occurrences to infer Relatedness has been refined and generalised in a wide variety of approaches ([Bottazzi and Pirino, 2010](#); [Cimini et al., 2019](#); [Li et al., 2014](#); [Nesta and Saviotti, 2005](#); [Pugliese et al., 2019](#); [Saracco et al., 2017](#); [Teece et al., 1994](#); [Tripodi et al., 2020](#); [Zaccaria et al., 2014](#); [Zhou et al., 2007, 2010](#)). All such approaches give rise to a network of Relatedness relations between couples of activities, i.e., in the language of statistical physics, a two-bodies correlation structure.

In order to give an objective assessment of the quality of these proxies for Relatedness, we test the ability of these networks to perform an out-of-sample link prediction task in the countries-products bipartite network. We perform the tests in a cross-validation setting that is briefly described in the chapter 6 (for a detailed description we refer to the methods section of [Tacchella et al. \(2021\)](#)). The finding is that the link prediction ability of the co-occurrence methods is generally poor, in some cases only marginally better than the one produced by a random network, and always inferior to trivial prediction strategies. From this evidence one could draw two kinds of conclusions: either (i) Relatedness is an unimportant concept in predicting the development paths of countries, or (ii) the co-occurrence proxy is able to provide a poor quantification of Relatedness. Our findings are strongly in favour of (ii). While we find very poor link-prediction performances from co-occurrence based topologies, at the same time we are able to build much better Relatedness proxies through more advanced machine learning based embedding techniques. However, we also find that describing countries development paths as the sum of binary Relatedness relationships is an oversim-

plification, and that much better results can be obtained considering higher-order interactions (i.e. patterns of absence/presence of many activities) through more complex but less interpretable tree-based model. These models can be learned with very effective strategies of data augmentation (bagging), where many models are learned on randomised subsamples of data and then averaged, and boosting, where the models are learnt in sequence, with each new model trained to minimise the residuals of the previous ones. These strategies allow to better cope with the relative scarcity of data and to learn more complex and effective models. Finally, we mention early attempts to adopt machine learning approaches in economic complexity ([Che, 2020](#); [Palmucci et al., 2020](#); [Tacchella et al., 2020](#)), which however either lack the systematic comparison in prediction tasks we show here, use different data, discuss only very specific test cases, or propose methodologies which are not suitable for the type and amount of data relevant here.

Chapter 6

Forecasting paths for Countries’ development

We consider the temporal countries-product network defined by the bipartite adjacency matrix

$$M_{cp}(t) = \begin{cases} 1 & \text{if } RCA_{cp}(t) \geq 1 \\ 0 & \text{otherwise} \end{cases} \quad (6.1)$$

where $RCA_{cp}(t)$ is the Revealed Comparative Advantage (RCA) ([Alshamsi et al., 2018](#)) of country c in product p in year t , based on COMTRADE data (see the box below). Roughly speaking, $M_{cp} = 1$ means that country c is exporting p in a competitive way.

COMTRADE data

The UN-COMTRADE database^a provides the monetary volumes of the trade flows between countries. Strikingly, the information is provided on a product basis, so we know which country is exporting a given amount of a given product. Since importers' and exporters' declarations do not coincide, suitable reconstruction algorithms are needed in order to achieve a coherent and sanitised dataset. By using a global Bayesian optimisation approach, we produced a denoised dataset (Mazzilli et al., 2021) that permits, by the way, to increase the GDP prediction performance in a considerable way (Tacchella et al., 2018a).

^a<https://comtrade.un.org>

We want to test if co-occurrence based methods are a good proxy for the Relatedness of countries' traded goods. In order to provide a quantitative and objective evaluation of how good these proxies are, we make use of the standard assumption that countries are more likely to develop new products that are related to the ones they already produce (Alshamsi et al., 2018; Hidalgo et al., 2007; Zaccaria et al., 2014). Therefore, our validation criteria are all related to the ability of any inferred Relatedness topology to predict the export basket of a country in year $t + \delta$ given its export basket in year t . Here we discuss the case $\delta = 5$, however different values of δ give similar results. More precisely, we implement a leave-k-countries-out cross-validation strategy (see the box below), so that we learn both Relatedness topologies and predictive models from a set of countries and then we test them on a different, non-overlapping set of countries. We consider various classes of predictive models¹:

¹In the paper (Tacchella et al., 2021) we also present a novel technique that we call Continuous Projection Space (CPS) that shows an improvements over the standard co-occurrences

- Baselines: no co-occurrence models that completely disregard or randomise the co-occurrence signal. The RCA method is based on the auto-correlation of the data.
- Co-occurrences-based: inference of a Relatedness graph based on the bipartite countries- products structure. Co-occurrence based techniques such as the *Product Space* (Hidalgo et al., 2007) and the *Taxonomy Network* (Zaccaria et al., 2014) belong to this class.
- Description-based: Relatedness topology here is based on the textual similarity (Zhelezniak et al., 2019) between product descriptions in the Harmonized System
- Tree-Based: tree-based machine learning algorithms that make a link prediction based on complex patterns of presence-absence of many links (many-body correlations). In particular, here we use XGBoost (XGB) (Chen and Guestrin, 2016; Friedman, 2001).

leave-k-countries-out cross-validation

To validate the link predictions in our temporal bipartite network, we train our models on $C - k$ countries data, and then we evaluate the predictions on the k countries data left out of the training set: we use the trained models to predict all the country-product links of the k countries starting from the year $t_0 + \delta$ (with $t_0 = 2007$ and $\delta = 5$). We repeat the process by excluding other subsets of countries until we have a prediction for all countries.

methods, still preserving their interpretability. Here we do not describe the CPS method, that comes mostly from the work of Tacchella A., and we describe more accurately the description-based method, that is due to the author of this thesis.

We test the predictive models on two link prediction tasks, namely to predict all the country-products links at time $t + \delta$ and to predict the links at time $t + \delta$ that had $RCA(t) < 0.25$. The first task is much less interesting from an economic point of view because of the very strong auto-correlation of the countries-products structure: a model that trivially predicts that $M_{cp}(t + \delta) = M_{cp}(t)$ is typically able to achieve very high scores with every classification metric. On the other hand, the second task measures the ability of the predictive models to forecast new links that are not due to small fluctuations of RCA, i.e. that are more likely to represent genuine economic development. Here we therefore focus on the results of the experiments on task 2². We perform the experiments using 12 years of data, from 2007 to 2018. In particular, for task 2 we compute a prediction score for each 0 element of the M_{cp} matrix from 2007 to 2013 and we validate the predictions by checking whether those links are present 5 years later, i.e. in the 2012-2018 data.

We evaluate the quality of the predictions with several standard classification metrics. In Figure 6.1 we show the results for some of these metrics, chosen to be important for practical applications and to capture different aspects of the prediction task. The metrics presented in Figure 6.1 are:

- *BestF1*: the F1-Score computed at the optimal decision threshold. This score is computed across all the predicted links for task 2.
- *prec@1000*: the precision of the top 1000 predicted links of the M_{cp} matrix in 2018
- *mAP@10*: the mean Average Precision of the top 10 predicted links for each country in 2018

²For the results of the task 1 we refer to the paper [Tacchella et al. \(2021\)](#)

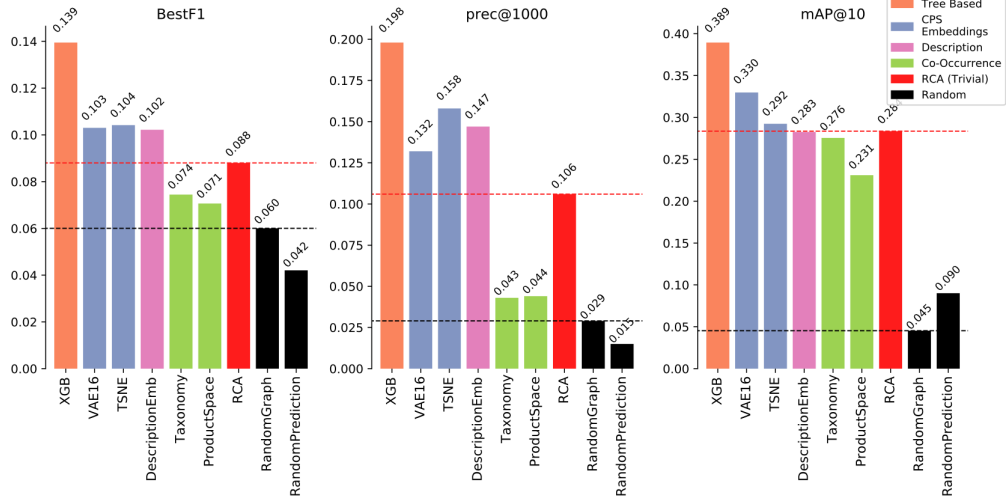


Figure 6.1: Comparison of the prediction performance of different algorithm typologies using different evaluation metrics. Tree-based machine learning techniques (here, XGBoost) overperform all other approaches. Description-based performs better than the RCA auto-correlation baseline, which in turn overperform co-occurrences. Taken from [Tacchella et al. \(2021\)](#). The two blue bars are relative to CPS method that is not presented in this thesis. We refer to the paper for a comprehensive comparison of all the models used.

BestF1 and Precision, therefore, are computed as a function of the predicted links of the whole graph, and we refer to them as global metrics, while AP@10 is computed country by country and then averaged (mAP@10).

6.1 Prediction results

In the global metrics the co-location based Relatedness proxies perform only marginally better than the Random Graph topology and much worse than the trivial RCA baseline, especially in the prec@1000 metric, where RCA alone performs more than twice as good as Product Space and Taxonomy. In the country-level metric, i.e. the mean Average Precision of top 10 more likely links for each country, the Random Graph performs much worse than the co-location topologies, but the Product Space is not able to gain any advantage on the trivial RCA prediction, while the Taxonomy manages to score better. The relatively good

performance of the Random Graph in the global tasks can be explained with the density-based link prediction approach that we use: more diversified countries are more likely to develop new links in the future, and at the same time receive higher scores for all missing links with respect to less diversified countries; this effect is lost in the country-level metrics, that effectively control for diversification by design. It is possible to systematically control for diversification in all density-based prediction strategies (i.e. all the methods presented here with the exception of XGB and RCA) by, e.g. z-scoring all the predictions for each country/year. However, as demonstrated by the relatively good results of Random Graph, diversification per-se is a predictive factor and we obtain that about 45% of the times z-scoring actually worsens the prediction metrics. The full results of these tests are provided in the SI.

Description embeddings and CPS embeddings are able to extract a Relatedness topology that is significantly better in all the global prediction tasks, while in $mAP@10$ only VAE16 provides a significant improvement over Taxonomy. Both CPS implementations are however always better than the RCA baseline. In all tasks, XGB provides a large improvement over all the other pairwise-distances approaches. This is to be expected as the method exploits complex patterns of presence and absence of products in the export baskets of a country and is therefore able to build much more complex decision functions. This comes, of course, at the price of a much less interpretable model. However, the availability of such methods cannot be neglected when the accuracy of the prediction is important, especially for business or policy-relevant decisions.

While a detailed explanation of the

6.1.1 Evaluation metrics

In order to compare the different prediction methodologies we make use of a series of performance indicators, usually adopted for classification tasks. It is important to point out that our results (see for instance fig. 6.1) are highly consistent across different indicators, even if we choose them for covering different aspects of the prediction exercise. Let us now discuss in detail how the prediction performances can be quantitatively evaluated. The specific instance to predict can be *positive* or *negative*, if the corresponding element of the export matrix M_{cp} is equal to 1 or 0, respectively. A *true positive* is a correctly predicted positive instance. Let us focus on the top $k = 1000$ scores of a given algorithm, that is, this algorithm assigns to these couples (c, p) the higher likelihoods to have $M_{cp} = 1$. The indicator $prec@1000$ is defined as the fraction of these 1000 couples for which $M_{cp} = 1$. This is a measure of the global precision of the algorithm. However, this measure takes into account all matrix elements together, while we might be interested in evaluating the prediction performance on a country basis: on average, how much are we precise when recommending a product to a country? To do so we first evaluate the precision country by country, and then we average. This is called the mean precision. Moreover, it is also important to weight our success or failure using the scores rank: we want the highest scores to predict better than the lowest scores. So we have to compute a weighted average. In practice, we use the mean Average Precision $mAP@10$. Let us focus on a single country c first. The Average Precision $AP@k(c)$ is defined as

$$AP@k(c) = \frac{\sum_{i=1}^k prec@k(c) \times rel(i, c)}{min(k, P(c))} \quad (6.2)$$

where $prec@k(c)$ is the precision at k for country c in the list; $rel(k, c)$ is equal to 1 if the item at rank k is 1 (a positive) and 0 otherwise; $P(c)$ is the total number

of positives for country c .

Them, the $mAP@k$ is simply the country average:

$$mAP@k = \frac{\sum_{c=1}^C AP@k(c)}{C} \quad (6.3)$$

Precision-related measures deal with the minimisation of false positives FP. In order to take into account also the problem of false negatives FN, *recall* is usually considered. In general, precision is defined as the ratio between the number of true positives TP and the number of predicted positives, while recall is the ratio between true positives TP and true instances, the real positives P. Since we want a global and balanced measure, we average the two with an harmonic mean, called F1-score. The harmonic mean aggravates the impact of possible small values of one of the two indicators. Since binary classifiers usually provide a continuous set of scores, one has to specify a threshold t above which the score is associated to a positive prediction; as a consequence, precision, recall, and the F1 score will depend on t . We point out that in the computation of $prec@1000$ the threshold choice is derived from the choice $k = 1000$. To have a non trivial and non arbitrary threshold we decided to take the one that maximises the in-sample F1 score, as suggested by . Summarising the above considerations, the best F1 score shown in Fig.6.1 is defined as

$$\text{BestF1score} = \max_t \frac{2}{\text{prec}(t)^{-1} + \text{rec}(t)^{-1}} \quad (6.4)$$

where

$$\text{prec}(t) = \frac{TP(t)}{TP(t) + FP(t)} \quad \text{rec}(t) = \frac{TP(t)}{TP(t) + FN(t)} = \frac{TP(t)}{P} \quad (6.5)$$

These prediction performance measures can be computed for any test set. In

Fig.6.1 we compute the $prec@1000$ and the $mAP@10$ for the last year of our dataset (2018): this bears the interpretation of our result as a recommender system, in which products are suggested as feasible to countries, and countries actually start exporting them. Instead, the best F1-score is computed on all the available years in cross validation, to show that our results are stable and comparable across different periods.

Chapter 7

Prediction methods and Relatedness assessment

All the predictions of the links in the temporal bipartite network countries-products task presented in chapter 6 make use of a specific definition of Relatedness between products, used to make the predictions under the assumption that new links appear where related ones already exist.

7.1 Density-based models

Except for tree-based models, all the others make predictions by creating a score $\mathcal{S}_{c,p}^{t+\delta}$ proportional to the estimated probability that the country c will have, at time $t + \delta$, $RCA > 1$ in the product p . Such score is calculated as

$$\mathcal{S}_{c,p}^{t+\delta} = \frac{\sum_{p'} B_{pp'}^{(\tilde{c})} M_{cp'}^t}{\sum_{p'} B_{pp'}^{(\tilde{c})}} \quad (7.1)$$

where $\tilde{c}$ indicates that the models parameters have been tuned on training data where country c was excluded. We refer as *density based* to the predictions created through the score $\mathcal{S}_{c,p}^{t+\delta}$ because here we are considering the density of products in

which a country has $RCA > 1$ around a target product p , weighting with the relatedness score $B_{pp'}$. The tree-based models, instead, do not consider two-products similarities and they are able to catch more complex correlations between products, paying the price of being less interpretable than the simple co-occurrences methods.

In this section we briefly explain how the models that appear in the evaluation showed in figure 6.1 work, paying more attention in the technique through which we assessed the Relatedness between couples of products by evaluating the similarities of their textual descriptions¹.

Baselines

In order to control the performance of the models under examination, we set up two trivial baseline models (whose results are represented in black and in red in figure 6.1). Exception made for the last black bar on the right in the plots of figure 6.1, that represents a totally random prediction, both *Random Graph* and *RCA* baselines have a prediction power that lies in different reasons. Both the baseline methods uses the score presented in equation 7.1 in order to predict the link for the couple country-product (c, p) in the year $t + \delta$.

The **RCA method** simply uses, for the year $t + \delta$, the same $RCA_{c,p}^t$ as the score $S_{c,p}^{t+\delta}$, discarding any product-density or co-occurrence with other products. This trivial method still preserves a certain forecasting power because of the strong auto-correlation of the export database, also overperforming the methods obtained from the network of co-occurrences (described in the next paragraph).

Random Graph model, instead, calculates the score $S_{c,p}^{t+\delta}$ using exactly the

¹A detailed review about how these models work can be found in the Methods section of our paper (Tacchella et al., 2021)

equation 7.1, but randomising the Relatedness matrix $B_{pp'}$. Mostly thanks to the nested structure of the M_{cp} matrix (Bustos et al., 2012), also this method still makes predictions better than the completely random method: more diversified countries are more likely to become even more diversified than non-diversified countries.

Co-occurrence-based

This methods basically consider two products as related, if it is usual that a country have both of them in its export basket. Hence, here the Relatedness matrix is a projection of the bipartite country-product network into a monopartite network of products. More precisely, the Relatedness between products p and p' is obtained with

$$B_{pp'} = \sum_c M_{cp} M_{cp'}. \quad (7.2)$$

The different normalisation factor distinguishes between *Product Space* method (that controls for the fact that more ubiquitous products have more co-occurrences) and *Taxonomy* method (that also gives less weight to the co-occurrence if it happens in countries with high diversification, more likely to share many products).

7.2 Tree-based models

As we mentioned before, these methods do not take into account just two-products correlations, but instead they manage to find more complex correlations between many products. The way they deal with the scarcity of data, that could create the risk of learning spurious correlations, is the technique of *bagging* and *boosting*: they create many models using a randomly chosen subset of all the available features (bagging) and these *weak* models are trained in sequence, with each new model trained by giving focus to the training examples that generated the largest

losses in the previously learned models (boosting).

The tree models are trained to specifically predict $M_{cp}^{t+\delta}$ starting from $\overrightarrow{RCA_c^t}$, i.e. from all the RCA values for the country c at time t .

The price that these models pay for better predict the future export basket of a country, is the interpretability of the results, since we cannot have anymore a clear vision about the pattern in the past data that created a specific prediction.

7.3 Textual descriptions' similarity

One of the proposed methodologies to quantify the Relatedness between products makes use of their textual description, provided by the Harmonized System, an internationally standardised system that contains a nomenclature of all the products of the world trade. We applied a Natural Language Processing technique to mathematically calculate the distance between two textual descriptions.

In literature we can find several methods with which to mathematically calculate a distance between two sentences, and therefore to be able to establish possible semantic links. Each of these is based on the fundamental concept of creating a vector, defined in a mathematical space, which is the geometric representation of the single sentence, so as to be able to apply the known distance formulas most suited to the specific case.

One of the most used methods to obtain such representation is Bag-of-Words (BoW) [Harris \(1954\)](#); [Salton et al. \(1975\)](#); [Sanderson \(2010\)](#), according to which a N-dimensional vector is created for each sentence (where N is the number of words in the overall set of words available in the dataset, i.e. its vocabulary). Each component, representing a word, contains the number of occurrences of that word in the sentence.

Other methods, developed later thanks to the evolution of Deep Learning

techniques [Kim \(2014\)](#); [Subramanian et al. \(2018\)](#); [Zhelezniak et al. \(2018\)](#) outperformed the BoW models. They create vectorial representations by using neural networks, specifically trained to minimise the distance (usually the Cosine Similarity) between paraphrases, in a supervised setting.

In a recent article, published in April 2019 [Zhelezniak et al. \(2019\)](#), the authors show a new technique that they prove to have good performances on different tasks, overcoming the most recent models of deep learning too. They call this technique Fuzzy-Bag-of-Words (FBoW).

We decided to adopt this last technique for our task, consisting in the assessment of the similarities between text descriptions of the products in an unsupervised setting.

7.3.1 Fuzzy Bag of Words

As the name suggests, this new technique is a modification of the BoW method, which we will therefore briefly describe. BoW is based on the concept that the larger is the number of words two sentences share, the more similar the sentences will be. Mathematically the vector representations of the individual sentences are obtained as follows

$$\mathbf{b} = \sum_{i=1}^N c_i \mathbf{e}^{(i)} \quad (7.3)$$

Where c_i is the number of occurrences of the i -th word and $\mathbf{e}^{(i)}$ is the i -th canonical vector of the $\mathbb{R}^N$ basis. If you want to use the Jaccard distance, then, you will have

$$J_{A,B} = \frac{A \cap B}{A \cup B} \quad (7.4)$$

Where union and intersection will be given respectively by the total number

of words in the two sentences represented by the sets A and B and by the number of common words in the two. Since the only information that is preserved with this technique is whether the specific word of the vocabulary is present (at most how many times it is repeated) in the sentence or not, a limitation is that it isn't possible to capture those semantic similarities defined by words that are similar, but not the same, as the following example makes clear.

Let's assume we have two sentences $A = \text{"I have a cat"}$ and $B = \text{"He has one dog"}$. It is easy to verify that the number of common words among the sentences A and B is zero, so it follows that the BoW method, will produce zero similarity between A and B , calculated with the formula (7.4). This is obviously a problem, since we realize that the words "cat" and "dog" are quite similar, living in the same context as "pets".

The modification of FBoW aims to correct this limitation: each component of the vector representation will no longer be 0 or c_i integer, but it will be a real number that will establish the similarity of the word w_i with all the words in the vocabulary. This similarity is defined by operations performed on pretrained word-embeddings. In doing so, even though the sentences don't share any word, there will be words that bring positive values to the similarity score between the two sentences, as long as they belong to similar contexts.

To provide a better understanding of the method, we describe below the mathematical steps that lead to the similarity assessment of two sentences.

Suppose we want to establish a similarity score between two descriptions s_1 and s_2 , respectively composed of N_1 and N_2 words. As already anticipated, the starting point for FBoW is a set of pretrained word-vectors (for the purpose of this article, the Elmo-Embeddings [Peters et al. \(2018\)](#) model was used): a matrix of size $N \times d$, where N is the number of words of the whole vocabulary and d is the size of the pretrained vectors. Each i -th word of the description s will then

be represented by a vector

$$\mathbf{w}_i = \mathbf{e}^{(i)} \mathbf{W} \quad (7.5)$$

of size $1 \times d$. At this point it is possible to establish a distance between the i -th word and each word of the vocabulary, identifying the context to which the word eventually belongs

$$\mu^{(i)} = \mathbf{w}_i \mathbf{W}^T \quad (7.6)$$

Here $\mu^{(i)}$ is a vector of dimension $1 \times N$. Each component of $\mu^{(i)}$ indicates the proximity of the i -th word of the sentence s to each word of the vocabulary.

Once that you have calculated $\mu^{(i)}$, you have a vector for each word of s , but you are interested in a vector that identifies the full sentence's proximity to all the words in the vocabulary, and not just the individual words' proximity. The first thing you could think to do, therefore, is to take the average proximity across all the words of the sentence for each component: $\mu^{s_1} = \frac{1}{N_1} \sum_{i=1}^{N_1} c_i \mu^{(i)}$ (weighting it with the number c_i of occurrences of the i -th word in the sentence s_1), but the operation that in fact you would like to perform is the union between two sets, which for the fuzzy-sets is correctly realized only taking the maximum value of proximity, scored by the closest word of s_1 , to the relevant word of vocabulary s (in fact $\{\text{cat}\} \cup \{\text{cat}\} = \{\text{cat}\}$ exactly like $\max\{1, 1\} = 1$ and $\{\text{cat}\} \cap \emptyset = \{\text{cat}\}$, like $\max\{1, 0\} = 1$ and not 0.5), then

$$\mu^{s_1} = \max_{i=1}^{N_1} c_i \mu^{(i)} \quad (7.7)$$

With which you obtain a vector representation μ^{s_i} (of size $1 \times N$) for each sentence considered. Finally it is possible to calculate the Jaccard distance between

the two sentences, remembering the concept of union (max) and intersection (min) for the fuzzy-sets

$$J_{A,B} = \frac{A \cap B}{A \cup B} \quad \rightarrow \quad FJ\{\mu^A, \mu^B\} = \frac{\sum_{i=1}^{N_1+N_2} \min\{\mu_i^A, \mu_i^B\}}{\sum_{i=1}^{N_1+N_2} \max\{\mu_i^A, \mu_i^B\}} \quad (7.8)$$

Having used as total of the available words the union of the words that make up the two sentences A and B, instead of all the words in the vocabulary (note that c_i would still be null for both vectors μ , if you considered words neither in A nor in B).

Discussion

The concept of Relatedness can be an extremely powerful tool to understand development dynamics and to inform policy decisions. By quantifying the proximity between activities in terms of knowledge, inputs, and infrastructures, it can help designing paths to diversification or specialisation strategies, based on empirical evidence. However, while the idea is extremely appealing, we show that current methods to estimate Relatedness from data perform poorly in predicting the trajectories of countries, even when compared to trivial alternative approaches.

In order to overcome these limitations we have introduced novel embedding techniques and we challenged their Relatedness assessment in the task of predicting the new products that a country will add to its export basket, starting from the ones in which the country is already competitive. We showed that in global metrics the technique of assessing similarities between products by looking at their textual description, overperforms the standard co-occurrences techniques, still preserving the same interpretability of the standard methods. Tree-based models, instead, can catch more complex patterns by looking at many-products correlation, at the price of interpretability, as reflected in their better performance in predicting the future export basket of the countries, assessed with every metrics used here.

Bibliography

- S. Agrawal, G. Foster, M. Freitag, and C. Cherry. Assessing Reference-Free Peer Evaluation for Machine Translation. *arXiv preprint arXiv:2104.05146*, 2021.
- A. Alshamsi, F. L. Pinheiro, and C. A. Hidalgo. Optimal diversification strategies in the networks of related products and of related research areas. *Nature communications*, 9(1):1–7, 2018. ISSN 2041-1723.
- D. R. Amancio, L. Antiqueira, T. A. S. Pardo, L. Da, F. Costa, O. N. Oliveira, and M. G. V. Nunes. Complex networks analysis of manual and machine translations. Technical Report 4, 2008. URL www.worldscientific.com.
- D. R. Amancio, M. G. Nunes, O. N. Oliveira, T. A. Pardo, L. Antiqueira, and L. Da. Using metrics from complex networks to evaluate machine translation. *Physica A: Statistical Mechanics and its Applications*, 390(1):131–142, 1 2011. ISSN 03784371. doi: 10.1016/j.physa.2010.08.052.
- V. Ambati. *Active learning and crowdsourcing for machine translation in low resource scenarios*. PhD thesis, Carnegie Mellon University, 2012.
- L. Antiqueira, M. G. Nunes, O. N. Oliveira, and L. d. F. Costa. Strong correlations between text quality and complex networks features. *Physica A: Statistical Mechanics and its Applications*, 373:811–820, 1 2007. ISSN 03784371. doi: 10.1016/j.physa.2006.06.002.

- P.-A. Balland, R. Boschma, J. Crespo, and D. L. Rigby. Smart specialization policy in the European Union: relatedness, knowledge complexity and regional diversification. *Regional studies*, 53(9):1252–1268, 2019. ISSN 0034-3404.
- M. Ballerini, N. Cabibbo, R. Candelier, A. Cavagna, E. Cisbani, I. Giardina, V. Lecomte, A. Orlandi, G. Parisi, A. Procaccini, M. Viale, and V. Zdravkovic. Interaction ruling animal collective behavior depends on topological rather than metric distance: Evidence from a field study. Technical report, 2008. URL www.pnas.org/cgi/content/full/.
- C. Beckner, N. C. Ellis, R. Blythe, J. Holland, J. Bybee, J. Ke, M. H. Christiansen, D. Larsen-Freeman, W. Croft, and T. Schoenemann. Language Is a Complex Adaptive System: Position Paper The "Five Graces Group". *Language Learning Research Club*, 59:1–26, 2009. ISSN 0023-8333.
- S. Bhardwaj, D. A. Hermelo, P. Langlais, G. Bernier-Colborne, C. Goutte, and M. Simard. Human or Neural Translation? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6553–6564, 2020.
- M. Bloodgood and C. Callison-Burch. Using Mechanical Turk to build machine translation evaluation sets. *arXiv preprint arXiv:1410.5491*, 2014.
- O. Bojar, C. Buck, C. Callison-Burch, C. Federmann, B. Haddow, P. Koehn, C. Monz, M. Post, R. Soricut, and L. Specia. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria, 8 2013. Association for Computational Linguistics. URL <https://aclanthology.org/W13-2201>.
- O. Bojar, C. Buck, C. Federmann, B. Haddow, P. Koehn, J. Leveling, C. Monz, P. Pecina, M. Post, and H. Saint-Amand. Findings of the 2014 workshop on sta-

- tistical machine translation. In *Proceedings of the ninth workshop on statistical machine translation*, pages 12–58, 2014.
- O. Bojar, R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, S. Huang, M. Huck, P. Koehn, Q. Liu, and V. Logacheva. Findings of the 2017 conference on machine translation (wmt17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, 2017.
- G. Bottazzi and D. Pirino. Measuring industry relatedness and corporate coherence. *Available at SSRN 1831479*, 2010.
- L. Bowker. Towards a methodology for a corpus-based approach to translation evaluation. *Meta: journal des traducteurs/Meta: Translators’ Journal*, 46(2): 345–364, 2001. ISSN 0026-0452.
- J. Bun, J.-P. Bouchaud, and M. Potters. Cleaning large correlation matrices: Tools from Random Matrix Theory. *Physics Reports*, 666:1–109, 2017. ISSN 0370-1573. doi: <https://doi.org/10.1016/j.physrep.2016.10.005>. URL <https://www.sciencedirect.com/science/article/pii/S0370157316303337>.
- S. Bustos, C. Gomez, R. Hausmann, and C. A. Hidalgo. The dynamics of nestedness predicts the evolution of industrial ecosystems. *PloS one*, 7(11):e49393, 2012. ISSN 1932-6203.
- C. Callison-Burch, C. S. Fordyce, P. Koehn, C. Monz, and J. Schroeder. Further meta-evaluation of machine translation. In *Proceedings of the third workshop on statistical machine translation*, pages 70–106, 2008.
- M. Carl and M. J. Schaeffer. Why translation is difficult: A corpus-based study of non-literality in post-editing and from-scratch translation. *HERMES-Journal of Language and Communication in Business*, (56):43–57, 2017. ISSN 1903-1785.

- S. Castilho, S. Doherty, F. Gaspari, and J. Moorkens. Approaches to Human and Machine Translation Quality Assessment. pages 9–38. 2018. doi: 10.1007/978-3-319-91241-7{_}2.
- E. Chatzikoumi. How to evaluate machine translation: A review of automated and human metrics. *Natural Language Engineering*, 26(2):137–161, 2020. ISSN 1351-3249.
- N. X. Che. Intelligent Export Diversification: An Export Recommendation System with Machine Learning. 2020.
- T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.
- G. Cimini, T. Squartini, F. Saracco, D. Garlaschelli, A. Gabrielli, and G. Caldarelli. The statistical physics of real-world networks. *Nature Reviews Physics*, 1(1):58–71, 2019. ISSN 2522-5820.
- J. G. Cragg. Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods. *Econometrica*, 39(5):829–844, 1971. ISSN 00129682, 14680262. doi: 10.2307/1909582. URL <http://www.jstor.org/stable/1909582>.
- M. Cristelli, A. Gabrielli, A. Tacchella, G. Caldarelli, and L. Pietronero. Measuring the intangibles: A metrics for the economic complexity of countries and products. *PloS one*, 8(8):e70726, 2013. ISSN 1932-6203.
- J. Daems, L. Macken, and S. Vandepitte. Quality as the sum of its parts: A two-step approach for the identification of translation problems and translation quality assessment for HT and MT+ PE. In *Proceedings of MT Summit XIV*

- Workshop on Post-Editing Technology and Practice*, volume 2, pages 63–71, 2013.
- L. Dalla'Asta, A. Baronchelli, A. Barrat, and V. Loreto. Nonequilibrium dynamics of language games on complex networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 74(3), 2006. ISSN 15502376. doi: 10.1103/PhysRevE.74.036105.
- G. De Sutter, B. Cappelle, O. De Clercq, R. Loock, and K. Plevoets. Towards a corpus-based, statistical approach to translation quality: Measuring and visualizing linguistic deviance in student translations. *Linguistica Antverpiensia, New Series—Themes in Translation Studies*, 16, 2017. ISSN 2295-5739.
- M. Del Vicario, G. Vivaldo, A. Bessi, F. Zollo, A. Scala, G. Caldarelli, and W. Quattrociocchi. Echo Chambers: Emotional Contagion and Group Polarization on Facebook. *Scientific Reports*, 6, 12 2016. ISSN 20452322. doi: 10.1038/srep37825.
- M. Denkowski and A. Lavie. Exploring normalization techniques for human judgments of machine translation adequacy collected using Amazon Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 57–61, 2010.
- M. Denkowski, H. Al-Haj, and A. Lavie. Turker-assisted paraphrasing for English-Arabic machine translation. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 66–70, 2010.
- B. E. Dimitrova. *Expertise and explicitation in the translation process*, volume 64. John Benjamins Publishing, 2005. ISBN 9027294267.

- S. Doherty. Translations| The Impact of Translation Technologies on the Process and Product of Translation. *International journal of communication*, 10:23, 2016. ISSN 1932-8036.
- B. Efron. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1–26, 1 1979. doi: 10.1214/aos/1176344552. URL <https://doi.org/10.1214/aos/1176344552>.
- J. Eyckmans, W. Segers, and P. Anckaert. Translation assessment methodology and the prospects of European collaboration. In *Collaboration in language testing and assessment*, volume 26, pages 171–184. Peter Lang, 2012. ISBN 363163529X.
- S. Flaxman, S. Mishra, A. Gandy, H. J. T. Unwin, T. A. Mellan, H. Coupland, C. Whittaker, H. Zhu, T. Berah, J. W. Eaton, M. Monod, P. N. Perez-Guzman, N. Schmit, L. Cilloni, K. E. Ainslie, M. Baguelin, A. Boonyasiri, O. Boyd, L. Cattarino, L. V. Cooper, Z. Cucunubá, G. Cuomo-Dannenburg, A. Dighe, B. Djaafara, I. Dorigatti, S. L. van Elsland, R. G. FitzJohn, K. A. Gaythorpe, L. Geidelberg, N. C. Grassly, W. D. Green, T. Hallett, A. Hamlet, W. Hinsley, B. Jeffrey, E. Knock, D. J. Laydon, G. Nedjati-Gilani, P. Nouvellet, K. V. Parag, I. Siveroni, H. A. Thompson, R. Verity, E. Volz, C. E. Walters, H. Wang, Y. Wang, O. J. Watson, P. Winskill, X. Xi, P. G. Walker, A. C. Ghani, C. A. Donnelly, S. Riley, M. A. Vollmer, N. M. Ferguson, L. C. Okell, and S. Bhatt. Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature*, 584(7820):257–261, 8 2020. ISSN 14764687. doi: 10.1038/s41586-020-2405-7.
- M. Freitag, G. Foster, D. Grangier, V. Ratnakar, Q. Tan, and W. Macherey.

- Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. 4 2021. URL <http://arxiv.org/abs/2104.14478>.
- J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001. ISSN 0090-5364.
- M. Gellerstam. Translationese in Swedish novels translated from English. *Translation studies in Scandinavia*, 1:88–95, 1986.
- A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 1995. ISBN 0429258410.
- T. E. Gorochowski, A. Matyjaszkiewicz, T. Todd, N. Oak, K. Kowalska, S. Reid, K. T. Tsaneva-Atanasova, N. J. Savery, C. S. Grierson, and M. Di Bernardo. BSim: an agent-based tool for modeling bacterial populations in systems and synthetic biology. 2012. ISSN 1932-6203.
- Y. Graham, T. Baldwin, A. Moffat, and J. Zobel. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, 2013.
- Y. Graham, T. Baldwin, and N. Mathur. Accurate evaluation of segment-level machine translation metrics. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1183–1191, 2015.
- Y. Graham, B. Haddow, and P. Koehn. Translationese in Machine Translation Evaluation. 6 2019. URL <http://arxiv.org/abs/1906.09833>.
- J. Hale. Uncertainty about the rest of the sentence. *Cognitive science*, 30(4): 643–672, 2006. ISSN 0364-0213.

- C. Han. Translation quality assessment: a critical methodological review. *Translator*, 26(3):257–273, 2020. ISSN 17570409. doi: 10.1080/13556509.2020.1834751.
- L. Han. Machine Translation Evaluation Resources and Methods: A Survey. 5 2016. URL <http://arxiv.org/abs/1605.04515>.
- Z. S. Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.
- H. Hassan, A. Aue, C. Chen, V. Chowdhary, J. Clark, C. Federmann, X. Huang, M. Junczys-Dowmunt, W. Lewis, M. Li, S. Liu, T.-Y. Liu, R. Luo, A. Menezes, T. Qin, F. Seide, X. Tan, F. Tian, L. Wu, S. Wu, Y. Xia, D. Zhang, Z. Zhang, and M. Zhou. Achieving Human Parity on Automatic Chinese to English News Translation. 3 2018. URL <http://arxiv.org/abs/1803.05567>.
- C. A. Hidalgo. Economic complexity theory and applications. *Nature Reviews Physics*, pages 1–22, 2021.
- C. A. Hidalgo and R. Hausmann. The building blocks of economic complexity. *Proceedings of the national academy of sciences*, 106(26):10570–10575, 2009. ISSN 0027-8424.
- C. A. Hidalgo, B. Klinger, A.-L. Barabási, and R. Hausmann. The product space conditions the development of nations. *Science*, 317(5837):482–487, 2007. ISSN 0036-8075.
- C. A. Hidalgo, P.-A. Balland, R. Boschma, M. Delgado, M. Feldman, K. Frenken, E. Glaeser, C. He, D. F. Kogler, and A. Morrison. The principle of relatedness. In *International conference on complex systems*, pages 451–457. Springer, 2018.
- J. M. Hilbe. *Modeling count data*. Cambridge University Press, 2014. ISBN 1139992295.

- M. D. Hoffman and A. Gelman. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.
- J. S. Holmes. *Translated!: papers on literary translation and translation studies*. Number 7. Rodopi, 1988. ISBN 9062037399.
- J. House. *Translation quality assessment: A model revisited*. Gunter Narr Verlag, 1997. ISBN 3823350757.
- J. House. Translation quality assessment: Past and present. In *Translation: A multidisciplinary approach*, pages 241–264. Springer, 2014.
- J. Howe. *Crowdsourcing: How the power of the crowd is driving the future of business*. Random House, 2008. ISBN 1905211112.
- E. Jauniaux, A. Bhide, A. Kennedy, P. Woodward, C. Hubinont, S. Collins, G. Duncombe, P. Klaritsch, F. Chantraine, J. Kingdom, L. Grønbeck, K. Rull, B. Nigatu, M. Tikkanen, L. Sentilhes, T. Asatiani, W. C. Leung, T. Alhaidari, D. Brennan, E. Kondoh, J. I. Yang, M. Seoud, R. Jegasothy, S. Espino y Sosa, B. Jacod, F. D’Antonio, N. Shah, D. Bomba-Opon, D. Ayres-de Campos, K. Jeremic, T. L. Kok, P. Soma-Pillay, N. Tul Mandić, P. Lindqvist, T. B. Arnadottir, I. Hoesli, U. Jaisamrarn, A. Al Mulla, S. Robson, and R. Cortez. FIGO consensus guidelines on placenta accreta spectrum disorders: Prenatal diagnosis and screening. *International Journal of Gynecology and Obstetrics*, 140(3):274–280, 3 2018. ISSN 18793479. doi: 10.1002/ijgo.12408.
- J. S. Jordan, N. Srinivasan, and C. van Leeuwen. The role of complex systems theory in cognitive science, 11 2015. ISSN 16124790.
- Y. Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

- G. S. Koby, P. Fields, D. R. Hague, A. Lommel, and A. Melby. Defining translation quality. *Revista tradumàtica: traducció i tecnologies de la informació i la comunicació*, (12):413–420, 2014. ISSN 1578-7559.
- P. Koehn and C. Monz. Manual and automatic evaluation of machine translation between european languages. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 102–121, 2006.
- W. A. Kretzschmar and W. A. Kretzschmar Jr. *Language and complex systems*. Cambridge University Press, 2015. ISBN 1107100453.
- H. P. Krings. *Repairing texts: Empirical investigations of machine translation post-editing processes*, volume 5. Kent State University Press, 2001. ISBN 087338671X.
- A. L. Lagarda, D. Ortiz-Martínez, V. Alabau, and F. Casacuberta. Translating without in-domain corpus: Machine translation post-editing with online learning techniques. *Computer Speech and Language*, 32(1):109–134, 2015. ISSN 10958363. doi: 10.1016/j.csl.2014.10.004.
- T.-Y. Lai. Reliability and validity of a scale-based assessment for translation tests. *Meta: Journal des traducteurs/Meta: Translators’ Journal*, 56(3):713–722, 2011. ISSN 0026-0452.
- S. Lauscher. Translation quality assessment: Where can theory and practice meet? *Translator*, 6(2):149–168, 2000. ISSN 17570409. doi: 10.1080/13556509.2000.10799063.
- V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union, 1966.

- M.-X. Li, V. Palchykov, Z.-Q. Jiang, K. Kaski, J. Kertész, S. Micciche, M. Tumminello, W.-X. Zhou, and R. N. Mantegna. Statistically validated mobile communication networks: the evolution of motifs in European and Chinese data. *New Journal of Physics*, 16(8):083038, 2014. ISSN 1367-2630.
- J. Lin, M. Cader, and L. Pietronero. What African Industrial Development Can Learn from East Asian Successes. 2020.
- V. Loreto, A. Baronchelli, A. Mukherjee, A. Puglisi, and F. Tria. Statistical physics of language dynamics. *Journal of Statistical Mechanics: Theory and Experiment*, 2011(04):P04006, 2011. ISSN 1742-5468.
- L. Lü, M. Medo, C. H. Yeung, Y.-C. Zhang, Z.-K. Zhang, and T. Zhou. Recommender systems. *Physics reports*, 519(1):1–49, 2012. ISSN 0370-1573.
- K. Marais. *Translation theory and development studies: A complexity theory approach*. Routledge, 2014. ISBN 0203768280.
- M. S. Mariani, Z.-M. Ren, J. Bascompte, and C. J. Tessone. Nestedness in complex networks: Observation, emergence, and implications. *Physics Reports*, 813:1–90, 2019. ISSN 0370-1573. doi: <https://doi.org/10.1016/j.physrep.2019.04.001>. URL <https://www.sciencedirect.com/science/article/pii/S037015731930119X>.
- R. Martínez. A deeper look into metrics for translation quality assessment (TQA): A case study. *Miscelánea: A Journal of English and American Studies*, 49, 2014. ISSN 1137-6368.
- J. M. Martínez Martínez and E. Teich. Modeling routine in translation with entropy and surprisal: A comparison of learner and professional translations. *Kreativität und Hermeneutik in der Translation. Tübingen: Narr Francke Attempto Verlag*, 2017.

- N. Martínez Melis and A. Hurtado Albir. Assessment in translation studies: Research needs. *Meta: journal des traducteurs/Meta: Translators' Journal*, 46(2):272–287, 2001. ISSN 0026-0452.
- N. Mathur, T. Baldwin, and T. Cohn. Towards efficient machine translation evaluation by modelling annotators. In *Proceedings of the Australasian Language Technology Association Workshop 2018*, pages 77–82, 2018.
- D. Mazzilli, A. Tacchella, and L. Pietronero. In preparation. 2021.
- M. Mézard, G. Parisi, and M. A. Virasoro. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, volume 9. World Scientific Publishing Company, 1987. ISBN 9813103914.
- M. Miccheli, A. Leban, A. Tacchella, A. Zaccaria, D. Mazzilli, and S. Bratières. A Bayesian approach to translators' reliability assessment. *arXiv e-prints*, page arXiv:2203.07135, 3 2022.
- R. Muñoz Martín. From process studies to cognitive translatology. *EST Newsletter*, 46:20–21, 2015.
- F. Neffke, M. Henning, and R. Boschma. How do regions diversify over time? Industry relatedness and the development of new growth paths in regions. *Economic geography*, 87(3):237–265, 2011. ISSN 0013-0095.
- L. Nesta and P. P. Saviotti. Coherence of the knowledge base and the firm's innovative performance: evidence from the US pharmaceutical industry. *The Journal of Industrial Economics*, 53(1):123–142, 2005. ISSN 0022-1821.
- E. A. Nida. *Toward a science of translating: with special reference to principles and procedures involved in Bible translating*. Brill Archive, 1964.

- R. Niehus, P. M. De Salazar, A. R. Taylor, and M. Lipsitch. Using observational data to quantify bias of traveller-derived COVID-19 prevalence estimates in Wuhan, China. *The Lancet Infectious Diseases*, 20(7):803–808, 7 2020. ISSN 147344457. doi: 10.1016/S1473-3099(20)30229-2.
- F. J. Och. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st annual meeting of the Association for Computational Linguistics*, pages 160–167, 2003.
- J. Olive, C. Christianson, and J. McCary. *Handbook of natural language processing and machine translation: DARPA global autonomous language exploitation*. Springer Science & Business Media, 2011. ISBN 1441977139.
- A. Palmucci, H. Liao, A. Napoletano, and A. Zaccaria. Where is your field going? A machine learning approach to study the relative motion of the domains of physics. *PloS one*, 15(6):e0233997, 2020. ISSN 1932-6203.
- K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. Technical report, 2002.
- G. Parisi. Complex Systems: a Physicist’s Viewpoint. Technical report, 2002.
- M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.
- E. Pugliese and A. Tacchella. Economic complexity for competitiveness and innovation: A novel bottom-up strategy linking global and regional capacities. Technical report, 2020.
- E. Pugliese, G. Cimini, A. Patelli, A. Zaccaria, L. Pietronero, and A. Gabrielli. Unfolding the innovation system for the development of countries: coevolution

- of Science, Technology and Production. *Scientific reports*, 9(1):1–12, 2019. ISSN 2045-2322.
- S. T. Radev, F. Graw, S. Chen, N. T. Mutters, V. M. Eichel, T. Bärnighausen, and U. Köthe. OutbreakFlow: Model-based Bayesian inference of disease outbreak dynamics with invertible neural networks and its application to the COVID-19 pandemics in Germany. *PLoS Computational Biology*, 17(10), 10 2021. ISSN 15537358. doi: 10.1371/journal.pcbi.1009472.
- H. H. A. Rêgo, L. A. Braunstein, G. D’Agostino, H. E. Stanley, and S. Miyazima. When a text is translated does the complexity of its vocabulary change? Translations and target readerships. *PLoS ONE*, 9(10), 10 2014. ISSN 19326203. doi: 10.1371/journal.pone.0110213.
- K. Reiss and H. J. Vermeer. Groundwork for a general theory of translation. *Tubingen: Niemeyer*, 101, 1984.
- C. Rhee, S. S. Kadri, R. L. Danner, A. F. Suffredini, A. F. Massaro, B. T. Kitch, G. Lee, and M. Klompas. Diagnosing sepsis is subjective and highly variable: A survey of intensivists using case vignettes. *Critical Care*, 20(1), 4 2016. ISSN 1466609X. doi: 10.1186/s13054-016-1266-9.
- I. Rivera-Trigueros. Machine translation systems and quality assessment: a systematic review, 2021. ISSN 15728412.
- G. Salton, A. Wong, and C.-S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.
- A. Samoilenko, F. Karimi, D. Edler, J. Kunegis, and M. Strohmaier. Linguistic neighbourhoods: explaining cultural borders on Wikipedia through multilingual co-editing activity. *EPJ Data Science*, 5(1), 12 2016. ISSN 21931127. doi: 10.1140/epjds/s13688-016-0070-8.

- M. Sanderson. Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008. ISBN-13 978-0-521-86571-5, xxi+ 482 pages. *Natural Language Engineering*, 16(1):100–103, 2010.
- F. Saracco, M. J. Straka, R. Di Clemente, A. Gabrielli, G. Caldarelli, and T. Squartini. Inferring monopartite projections of bipartite networks: an entropy-based approach. *New Journal of Physics*, 19(5):053022, 2017. ISSN 1367-2630.
- A. Sbardella, E. Pugliese, A. Zaccaria, and P. Scaramozzino. The role of complex analysis in modelling economic growth. *Entropy*, 20(11):883, 2018.
- C. E. Shannon. A Mathematical Theory of Communication. Technical Report 3, 1948.
- G. M. Shreve. Translation as a complex adaptive system: A framework for theory building in cognitive translatology. In *The Routledge Handbook of Translation and Cognition*, pages 69–87. Routledge, 2020. ISBN 1315178125.
- A. Sîrbu, V. Loreto, V. D. P. Servedio, and F. Tria. Opinion Dynamics: Models, Extensions and External Effects. In V. Loreto, M. Haklay, A. Hotho, V. D. P. Servedio, G. Stumme, J. Theunis, and F. Tria, editors, *Participatory Sensing, Opinions and Collective Awareness*, pages 363–401. Springer International Publishing, Cham, 2017. ISBN 978-3-319-25658-0. doi: 10.1007/978-3-319-25658-0_{_}17. URL https://doi.org/10.1007/978-3-319-25658-0_17.
- X. Song, T. Cohn, and L. Specia. BLEU deconstructed: Designing a better MT evaluation metric. *International Journal of Computational Linguistics and Applications*, 4(2):29–44, 2013.

- L. Steels. Modeling the cultural evolution of language. *Physics of life reviews*, 8 (4):339–356, 2011. ISSN 1571-0645.
- V. Stojkoski, Z. Utkovski, and L. Kocarev. The impact of services on economic complexity: Service sophistication as route for economic growth. *PloS one*, 11 (8):e0161633, 2016. ISSN 1932-6203.
- S. Subramanian, A. Trischler, Y. Bengio, and C. J. Pal. Learning general purpose distributed sentence representations via large scale multi-task learning. *arXiv preprint arXiv:1804.00079*, 2018.
- A. Tacchella, M. Cristelli, G. Caldarelli, A. Gabrielli, and L. Pietronero. A new metrics for countries’ fitness and products’ complexity. *Scientific Reports*, 2, 10 2012. ISSN 20452322. doi: 10.1038/srep00723.
- A. Tacchella, M. Cristelli, G. Caldarelli, A. Gabrielli, and L. Pietronero. Economic complexity: conceptual grounding of a new metrics for global competitiveness. *Journal of Economic Dynamics and Control*, 37(8):1683–1691, 2013. ISSN 0165-1889.
- A. Tacchella, R. Di Clemente, A. Gabrielli, and L. Pietronero. The build-up of diversity in complex ecosystems. *arXiv preprint arXiv:1609.03617*, 2016.
- A. Tacchella, D. Mazzilli, and L. Pietronero. A Dynamical Systems approach to GDP Forecasting: Supplementary Information Regularization of Export Data and construction of the Countries-Products matrix. Technical report, 2018a.
- A. Tacchella, S. Romano, M. Ferraldeschi, M. Salvetti, A. Zaccaria, A. Crisanti, and F. Grassi. Collaboration between a human group and artificial intelligence can improve prediction of multiple sclerosis course: A proof-of-principle study [version 1; referees: 1 approved, 2 approved with reservations]. *F1000Research*, 6, 2018b. ISSN 1759796X. doi: 10.12688/f1000research.13114.1.

- A. Tacchella, A. Napoletano, and L. Pietronero. The language of innovation. *PloS one*, 15(4):e0230107, 2020. ISSN 1932-6203.
- A. Tacchella, A. Zaccaria, M. Miccheli, and L. Pietronero. Relatedness in the era of machine learning. *arXiv preprint arXiv:2103.06017*, 2021.
- D. J. Teece, R. Rumelt, G. Dosi, and S. Winter. Understanding corporate coherence: Theory and evidence. *Journal of economic behavior & organization*, 23(1):1–30, 1994. ISSN 0167-2681.
- A. Toral, S. Castilho, K. Hu, and A. Way. Attaining the Unattainable? Re-assessing Claims of Human Parity in Neural Machine Translation. *Proceedings of the Third Conference on Machine Translation (WMT)*, 1(research papers): 113–123, 2018. doi: 10.18653/v1/W18-64012. URL <https://doi.org/10.18653/v1/W18-64012>.
- G. Tripodi, F. Chiaromonte, and F. Lillo. Knowledge and social relatedness shape research portfolio diversification. *Scientific reports*, 10(1):1–12, 2020. ISSN 2045-2322.
- M. Turchi, M. Negri, M. F. Fondazione, and B. Kessler. Coping with the Subjectivity of Human Judgements in MT Quality Estimation. Technical report, 2014. URL <http://www.quest.dcs.shef.ac.uk/>.
- A. van Dam and K. Frenken. Vertical vs. Horizontal Policy in a Capabilities Model of Economic Development. *arXiv preprint arXiv:2006.04624*, 2020.
- D. van Ravenzwaaij, P. Cassey, and S. D. Brown. A simple introduction to Markov Chain Monte-Carlo sampling. *Psychonomic Bulletin & Review*, 25(1):143–154, 2018. ISSN 1531-5320. doi: 10.3758/s13423-016-1015-8. URL <https://doi.org/10.3758/s13423-016-1015-8>.

- A. Vespignani, H. Tian, C. Dye, J. O. Lloyd-Smith, R. M. Eggo, M. Shrestha, S. V. Scarpino, B. Gutierrez, M. U. Kraemer, J. Wu, K. Leung, and G. M. Leung. Modelling COVID-19, 6 2020. ISSN 25225820.
- C. Waddington. A positive approach to the assessment of translation errors. In *Actas del I Congreso Internacional de la Asociación Ibérica de Estudios de Traducción e Interpretación*, volume 2, pages 409–426, 2003.
- L. Wang, R. Song, Z. Qu, H. Zhao, and C. Zhai. Study of China’s publicity translations based on complex network theory. *IEEE Access*, 6:35753–35763, 6 2018. ISSN 21693536. doi: 10.1109/ACCESS.2018.2849680.
- T. Watanabe, J. Suzuki, H. Tsukada, and H. Isozaki. Online large-margin training for statistical machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 764–773, 2007.
- A. Way. Quality expectations of machine translation. Technical report, 2018. URL <https://www.taus.net/think-tank/news/press-release/size-machine-translation-market-is-250->.
- J. S. White, T. A. O’Connell, and F. E. O’Mara. The ARPA MT evaluation methodologies: evolution, lessons, and future approaches. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas*, 1994.
- J. Williams. *Theories of translation*. Springer, 2013. ISBN 1137319380.
- M. Williams. The assessment of professional translation quality: Creating credibility out of chaos. *TTR: traduction, terminologie, rédaction*, 2(2):13–33, 1989. ISSN 0835-8443.

- M. Yildiz. A critical perspective on the translation quality assessments of five translators organizations: ATA, CTTIC, ITI, NAATI, and SATI. *RumeliDE Dil ve Edebiyat Araştırmaları Dergisi*, (18):568–589, 2020. ISSN 2148-7782.
- A. Zaccaria, M. Cristelli, A. Tacchella, and L. Pietronero. How the taxonomy of products drives the economic development of countries. *PloS one*, 9(12): e113770, 2014. ISSN 1932-6203.
- A. Zaccaria, M. Cristelli, R. Kupers, A. Tacchella, and L. Pietronero. A case study for a new metrics for economic complexity: The Netherlands. *Journal of Economic Interaction and Coordination*, 11(1):151–169, 4 2016. ISSN 18607128. doi: 10.1007/s11403-015-0145-9.
- A. Zaccaria, S. Mishra, M. Z. Cader, and L. Pietronero. Integrating services in the economic fitness approach. *World Bank Policy Research Working Paper*, (8485), 2018.
- O. Zaidan and C. Callison-Burch. Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 1220–1229, 2011.
- J. Zehnalová. Tradition and Trends in Translation Quality assessment. Technical report, 2013. URL <https://www.researchgate.net/publication/294260655>.
- W. Zhao, G. Glavaš, M. Peyrard, Y. Gao, R. West, and S. Eger. On the limitations of cross-lingual encoders as exposed by reference-free machine translation evaluation. *arXiv preprint arXiv:2005.01196*, 2020.
- Y. Zhao, J. Zhang, C. Zong, Z. He, and H. Wu. Addressing the Under-Translation

- Problem from the Entropy Perspective. Technical report, 2019. URL www.aaai.org.
- V. Zhelezniak, D. Busbridge, A. Shen, S. L. Smith, and N. Y. Hammerla. Decoding Decoders: Finding Optimal Representation Spaces for Unsupervised Similarity Tasks. *arXiv preprint arXiv:1805.03435*, 2018.
- V. Zhelezniak, A. Savkov, A. Shen, F. Moramarco, J. Flann, and N. Y. Hammerla. Don’t Settle for Average, Go for the Max: Fuzzy Sets and Max-Pooled Word Vectors. *arXiv preprint arXiv:1904.13264*, 2019.
- T. Zhou, J. Ren, M. Medo, and Y.-C. Zhang. Bipartite network projection and personal recommendation. *Physical review E*, 76(4):046115, 2007.
- T. Zhou, Z. Kuscsik, J.-G. Liu, M. Medo, J. R. Wakeling, and Y.-C. Zhang. Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences*, 107(10):4511–4515, 2010. ISSN 0027-8424.

Pages 145-152 hidden for data protection reasons

Acknowledgements

This research project was funded by Translated SRL, a leading provider of professional translation services. I am grateful to Marco Trombetti, CEO of Translated, for continuous support and determining feedback, especially on validation; to several Translated colleagues: to the IT team for help with data acquisition and curation, especially Damiano Barbati and Emilio Spatola for helping me with Translated database and Domenico Lupinetti and Francesco Elefante for the valuable assistance with MateCat; to the Quality Management team with validation and mapping to business needs (special thanks to Helen Corfield); to the Project Management team for discussion of results and feedback, and especially to Antonella Giudici for her infinite availability and to Francesca Piscitelli and Daniele Cocozza for organisation and curation of the English-Malay experiment (Subsection 3.1.3). I am indebted to Marco Trombetti and Luciano Pietronero, Head of Centro Ricerche Enrico Fermi, Rome, for instigating the institutional collaboration which made this research project possible. I want to thank Sébastien Bratières for his supervision and conceptualisation of the entire work conducted in collaboration with Translated SRL. I want to sincerely thank my colleagues and friends Andrea Tacchella, Dario Mazzilli, and Andrea Zaccaria for their valuable ideas, that have been essential for both TQA research project and Relatedness one. Finally, I want to thank my advisor Luciano Pietronero for giving me the opportunity to work on this research project and for his supervision.

Marco Miccheli, Rome, April 2022
