Corpus Linguistics, Translation and Error Analysis

Maria Stambolieva

Centre for Computational and Applied Linguistics & Laboratory for Language Technology

New Bulgarian University

Abstract

The paper presents a study of the French *Imparfait* and its functional equivalents in Bulgarian and English in view of applications in (machine) translation, foreign language teaching and error analysis. The aims of the study are: 1/ based on the analysis of a corpus of text, to validate/revise earlier research on the values of the French *Imparfait*, 2/ to define the contextual factors pointing to the realisation of one or another value of the forms, 3/ based on the analysis of aligned translations, to identify the translation equivalents of these values, 4/ to formulate translation rules, 5/ based on the analysis of the translation rules, to refine the annotation modules of the environment used – the NBU E-Platform for language teaching and research.

1 Context

The paper presents work in progress, partly based on an earlier investigation by the same author (Stambolieva 2004), aiming 1/ to define the Tense-and-Aspect values of French sentences/clauses containing a verb marked for the *Imparfait*; 2/ to describe the linguistic markers linked to each value; 3/ to link these markers to translation equivalents in the two target languages: Bulgarian and English. The software environment is the NBU E-Platform for teaching and research.

The corpora used are the electronic versions of Antoine de Saint-Exupéry's *Le Petit Prince*[1] and its translations in Bulgarian[2] and English[3].

The following procedure was used:

1/ The source text was annotated (POS-tagged and tagged for *Imparfait*-marked forms) in the grammatical analysis module of the E-Platform.

2/ The source text was aligned with the texts in the target languages.

3/ With the respective E-Platform module, two virtual corpora were derived – files with lists of sentences containing a specific annotation value. In this case the corpora contain lists of sentences with *Imparfait*-marked verbal forms and their translation equivalents in the two target languages.

For the analysis of the French sentences in the virtual corpus, the theoretical model proposed by J.-P. Desclés (Desclés 1985, 1990) was adopted – a system organising four main elements: 1/ a system of grammatical forms, 2/ a system of values, 3/ a system of correspondences between 1/ and 2/, 4/ a system of strategies for context analysis. Important studies of the French *Imparfait* and its equivalents in Bulgarian have been published by Zlatka Guentcheva-Desclés (Guentcheva 1990,

[1] https://www.ebooksgratuits.com/html/st_exupery_le_petit_prince.html

[2] http://old.ppslaveikov.com/Roditeli/knigi%20Lqto/anton.sent.ekzuperi-makiat.princ.pdf

[3] http://verse.aasemoon.com/images/f/f5/The_Little_Prince.pdf

Temnikova, I, Orăsan, C., Corpas Pastor, G., and Mitkov, R. (eds.) (2019) *Proceedings of the 2nd Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT 2019)*, Varna, Bulgaria, September 5 - 6, pages 98–104.

https://doi.org/10.26615/issn.2683-0078.2019_012

Guentcheva 1997). A study of the French *Imparfai*t by M. Maire-Reppert (Maire-Reppert 1991) was found to be very useful for some of the values of the forms set out, the rich corpus of examples an the excellent attempt at formalization of the contextual markers of the values of the *Imparfait*. Danchev and Alexieva 1974 and Stambolieva 1987, 1998 and 2008 are contrastive corpus-based studies of contextual markers of tense and aspect in English and their Bulgarian functional equivalents. [4] A pioneering work on the compositionality of aspect in English is Verkuyl 1993.

The rules linking values to forms and context contain the following information:

1/ text element under investigation (indicator) – in our investigation, French verbal lexemes to which the morpheme of the *Imparfait* is attached;

2/ scope of the context where the contextual markers (indices) are found;

3/ contextual markers (indices) – elements of the immediate context which resolve the ambiguity of the indicator;

4/ values attributed to the combined indicator and indices;

5/ pairing of the values to functional equivalents in the target languages.

Thus, on a monolingual plane, we derive value indices of the indicators – the French verbal forms marked for the *Imparfait*. On a bilingual plane, indicators and indices are linked to translation equivalents in the target languages of the investigation.

The software environment of the project is that of the NBU e-Platform for language teaching and research (PLT&R)

2 The NBU E-Platform for Language Teaching and Research

The NBU E-Platform, a recent project of the NBU Laboratory for Language Technology[5], was initially developed as a tool for language teaching/learning: a generator of online training exercises from annotated corpora, with exports to Moodle or other educational platforms. It has since been extended with modules and functionalities allowing research in translation and error analysis and supporting lexicographic projects.

The E-Platform integrates: 1/ an environment for creating, organising and maintaining electronic text archives and extracting text corpora; 2/ modules for linguistic analysis: a lemmatiser, a POS analyser; a term analyser; a morphological analyser, a syntactic analyser; an analyser of multiple word units (MWU – including complex terms, analytical forms, phraseological units); a parallel text aligner; a concordancer; 3/ a linguistic database allowing corpus manipulation without loss of information; 4/ modules for the generation and editing of online training exercises. The environment for the maintenance of the electronic text archive organises a variety of metadata which can, individually or in combinations, form the basis for the extraction of text corpora. Following linguistic analysis, secondary ("virtual") corpora can be extracted – lists of sentences containing a particular unit – a lemma (e.g. *it, dislike*), a word form (e.g. *begins*), a MWU (e.g. *has been writing, put off*), a tag (e.g. <intransitive verb>, <comparative degree>, <present perfect progressive tense>, <imparfait>), or a combination of tags. The architecture allows the parallel use of several systems of preprocessing and the comparison of their results for the purpose of making an intelligent choice – which can turn it into an environment for experimentation and research.[6]

[4] Functional equivalence finding is the process, where the translator understands the concept in the source language and finds a way to express the same concept in the target language in the way, in which the **equivalent** conveys the same meaning and intent as the original. (Wikipedia)

[5] NBU CFSR-funded project: https://projects.nbu.bg/projects_inner.asp?pid=642

[6] Cf Stambolieva, Ivanova. Raykova 2018

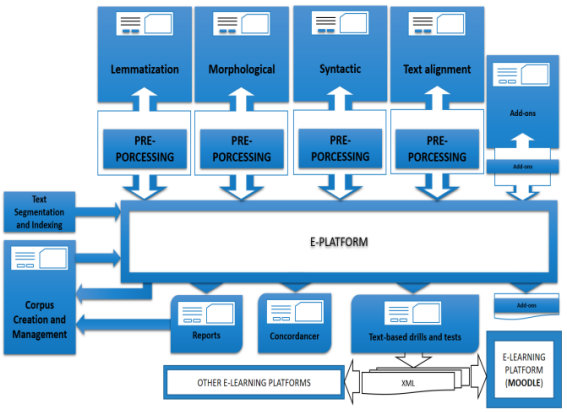

Table 1. Architecture of the E-Platform[7]

The following modules of the platform were extended for the purpose of the project:

- The Text & Corpus organizer
- The annotation modules: Lemmatiser, POS-tagger, Morphological and Syntactic tagger
- The Aligner
- The Virtual Corpus generator.

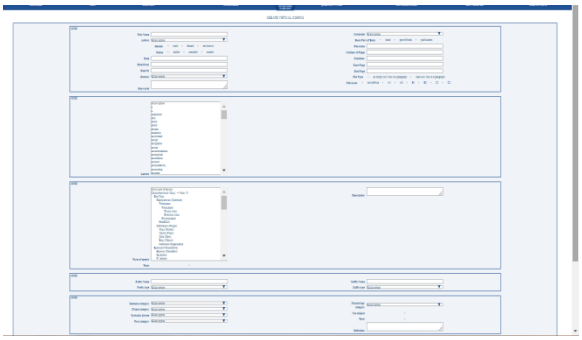

Table 2. The Virtual Corpus generator

A new module combining annotation and alignment was developed as an extension of the Virtual Corpus generator – a generator of virtual corpora coupled with aligned translation equivalents.

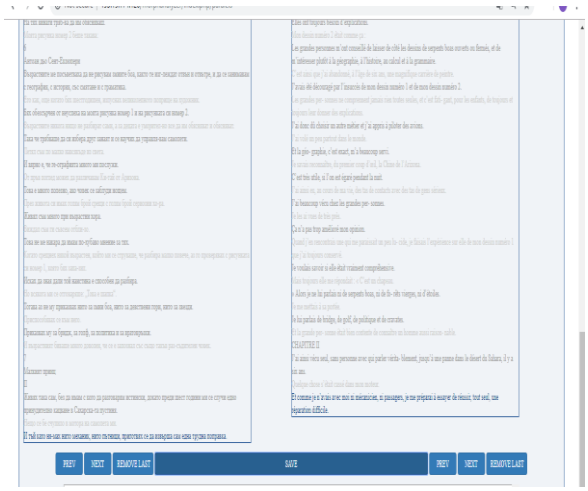

Table 3. Aligning with the E-Platform

3 Values of the *Imparfait* and Its Translation Equivalents in Bulgarian and English

The dominant translation equivalents of the *Imparfait* in our corpus are *The Simple Past Tense* (for English) and *The Past Imperfect* of Imperfective Aspect verbs (for Bulgarian).

Rule 1:

Given an instance X marked by the morpheme of the Imparfait,

and if X is a member of the set of verbs of a stative archetype,

then the value of the Imparfait is that of "descriptive state"

The Bulgarian translation contains a form of an Imperfective Aspect verb marked for The Past Imperfect

The English translation contains a stative verb marked for The Simple Past Tense.

However, other tense-aspect equivalents also appear: *The Past Continuous Tense* (for English) and *The Present Tense*, *The Past Indefinite Tense* and *The Past Perfect Tense* of Perfective Aspect verbs, *The Future in the Past Tense* (for Bulgarian)

[7] The E-Platform was initially developed by the Central Institute for Informatics and Computer Engineering of the. For its architecture, regular support and update we are indebted to our colleagues from the Informatics department of New Bulgarian University Dr. Mariyana Raykova.and Dr. Valentina Ivanova.

– hence the necessity to identify the different values of the *Imparfait* and their contextual markers. Maire-Reppert (Maire-Reppert, op. cit.) proposes a very fine-grained set of values of the French *Imparfait* for situation types and for registers, including seven subtypes of states,[8] three subtypes of processes, events, iterative situations, conditions and formulae of politeness. Based on her findings and an analysis of our corpus, we have arrived at a set of rules (94 in all), the general form of which is presented with the following simple rule for **Descriptive States**:

Ex. 1 *Il* ***représentait*** *un serpent boa qui digérait un éléphant*. *– Тя* ***изобразяваше*** *змия боа, която смила слон*. *– It* ***was*** *a picture of a boa constrictor digesting an elephant*.

Rule 1 indicates the necessity to extend the annotation module with subcategories/subtypes of verbal lexemes. For the description of the values of the *Imparfait*, six lists of verbal subtypes were drawn up: 1. Stative locative verbs, 2. *Verba dicendi*, 3. Stative link verbs, 4. Stative full verbs, 5. Change of State verbs 6. Dynamic full verbs with a closed right-hand bound (so-called "conclusives" – e.g. *perdre, mourir, comprendre)*.

Along with the lists of verbs, 15 more lists were drawn up: of adverbial expressions of frequency or place; of nouns belonging to the semantic subgroups 'characteristic feature', or 'item of clothing'; 'taking'expressions, as e.g. *se servir de*, *utiliser*, *employer*, etc.

A similar rule, with non-stative verbs, has been formulated for **Processes in Development**. The Bulgarian translations contain a Past Imperfect form of an Imperfective Aspect verb. The English equivalents can appear in both the Past Continuous Tense and the Past Simple Tense (which is the unmarked member of the opposition).

Ex. 2 *Comme le petit prince* ***s'endormai****t, je le pris dans mes bras, et me remis en route. – Малкият принц* ***заспиваше****, аз го взех на ръце и отново тръгнах*. *– As the little prince dropped off to sleep, I took him in my arms and set out walking once more.*

The following main **triggers of asymmetry** in the translation equivalents were identified:

1/ **The Sequence of Tenses** is part of the grammatical systems of French and English, but not of Bulgarian:

Ex. 3 *J'avais ainsi appris une seconde chose très importante: c'est que sa planète d'origine* ***était*** *à peine plus grande qu'une maison* ! *– Така узнах второ, много важно нещо: че неговата родна планета* ***е*** *малко по-голяма от къща*! *– I had thus learned a second fact of great importance: this was that the planet the little prince came from* ***was*** *scarcely any larger than a house*!

Rule 2 relies on syntactic annotation – it involves marking sentences as Simple, Compound and Complex Sentences, and clauses (at least) as Main and Subordinate.

2/ **New State** is typically marked by a verb of dynamic archetype (although French source sentences can also appear with the verb *être* in the *Imparfait*). The English translations contain a *Simple Past* tense form of the verb (including *to be*), while a verb of dynamic archetype (of Perfective Aspect) must appear in the Bulgarian translations, marked for *The Past Perfect Tense.* The contextual markers defining the situation as non-descriptive are adverbial expressions appearing in Change-of-State lists, as well as adverbial expressions which do not appear in lists of expressions marking processes in development – such as *pendant, pendant que, tandis que, alors que*. etc.

Rule 3.

Given an instance X marked by the morpheme of the *Imparfait*

and if X has a dynamic archetype

and if X is in a list of verbs of the Conclusive type

[8] (Descriptive (état descriptif), Resultant, (état résultant), Inferential (état à valeur inférentielle), of Acquired experience (état à valeur d'expérience), Passive (état passif), New (nouvel état) or Permanent state (état permanent)).

and it there is, in the same clause, a phrase belonging to list of temporal expressions

then the value of X is that of "new state"

The Bulgarian translation contains a form of a verb marked for *The Past Perfect of Perfective Aspect verbs*

The English verb contains a verb in *The Past Simple Tense.*

Ex. 4 *Le premier ministre arrivait. On entra en conférence.* (Corpus of M.-Reppert)

Rule 2

Given an instance X marked by the morpheme of the *Passé Composé* or *Passé Simple* ,

and if X is in the list of verb 'Verba dicendi',

and given an instance of a verb Y marked by the morpheme of the *Imparfait* within a Subordinate Clause introduced by the Conjunction *que*

then the value of the *Imparfait* is that of "permanent state"

The Bulgarian translation contains a form of a verb marked for *The Present*.

The English verb contains a verb in *The Past Simple*, *The Past Continuous* or *The Perfect Perfect Tense.*

For the New State translation rules, the Bulgarian forms must be tagged for Aspect. The values of this category are part of the POS-tagger of the e-Platform.

Rule 3 is one of the 9 New State rules formulated for New States and their translations.

3/ **Real Conditions**. French verbs appearing in the Subordinate Clause of Real Conditions introduced by the conjunction *si* often appear in the *Imparfait*. The English tense form in the translation equivalent is in most cases in the *Simple Present Tense*; the Bulgarian one is in the *Present Tense*.

Ex. 5 *Elle serait bien vexée, se dit-il, si elle voyait ça. – Ако види това – каза си той, -- ще бъде обидена. – If she sees that, he thought, she will be hurt.*

4/ **Iterative situations**. For this value, the data from the two corpora have been described in 19 rules; the cases of asymmetry are restricted to predictable, structure-induced cross-language transformations. The general rule is presented below:

Rule 5.

Given an instance X marked by the morpheme of the *Imparfait*

and if an element, member of a list of adverbs of frequency (*parfois, quelquefois, plusieurs fois,* etc.), appears in the same clause,

then the value of X is that of "Iterative Situation".

The Bulgarian translation contains a form of a verb marked for *The Past Imperfect Tense*

The English verb contains a verb in *The Past Simple Tense* OR *Past Continuous Tense* OR *a would/used to + Infinitive* structure.

5/ **Expression of Politeness**. This value of the *Imparfait* allows the speakers to grant their interlocutors – as a sign of politeness or reserve – the option to oppose, as it were, the process:

Rule 6.

Given an instance X marked by the morpheme of the *Imparfait*

and if the clause contains a *verbum dicendi*,

and if the main clause contains a personal pronoun in the first or second person singular or plural or a nominal syntagm from a list of polite forms of address,

then the value of X is that of "Expression of Politeness"

The Bulgarian translation contains a form of a verb marked for *The Past Imperfect Tense*

The English verb contains a verb in *The Past Simple Tense.* OR *a modal form*, e.g. *would like + to-infinitive.*

Ex. 6 *Je voulais vous dire que je ne pourrai pas venir demain.# Je venais dire à Madame que le déjeuner était servi*.

Rule 4.

Given an instance X marked by the morpheme of the *Imparfait*

and if X appears in a Subordinate Clause introduced by the Conjunction *si*

and if the main clause contains a verbal form marked for the *Conditionnel*,

then "Real Condition" can be the value of X.

The Bulgarian translations contain a form of a verb marked for *The Present Tense*

The English translations contain a verb in *The Present Simple Tense.*

Contextual markers for this value of the Imparfait are: 1/ the presence of *verba dicendi*, 2/ personal pronouns for the first and second person singular or plural in the same clause, or a nominal syntagm from a list including *Madame, Mademoiselle, Monsieur*. The Bulgarian translations appear in the Present Tense, the English translations – in the Present Simple tense.

6/ **The Non-Evidential mood** [9] in Bulgarian. The contextual factors triggering this type of French & English vs Bulgarian asymmetry are yet to be analyzed before the formulation of the translation rules.

4 Conclusions

The analysis of the corpus indicates that the formulation of translation rules for the French *Imparfait* involves lexical, morphological and syntactic annotation of the micro context of the tense marker (the verbal lexeme) and of the macrocontext of the sentence/clause.

The verbal lexemes forming the microcontext of the *Imparfait* marker fall into several subclasses, which have been added to the tagsets in the annotation modules of the e-Platform. The macrocontext of the verbal forms, i.e. their left and right hand environment, must be syntactically tagged for sentence type and clause status and function, along with the standard parts-of-the sentence and POS-tagging. These values were added to the annotation set of the syntactic module.

Our findings also indicate that simple identification of WHEN-type adverbial modification [10] is not sufficient to define the temporal values of the French *Imparfait*. They confirm the need to include frequency expressions – as proposed in the guidelines and methods formulated by I. Mani et al (Mani et al, 2001) and J.-P. Desclés (Desclés 1997).

An extended set of annotation values was found to be necessary for the description of those values of the *Imparfait*-marked sentences/ clauses where the morpheme does not mark temporality.

5 Applications

The analysis of the contextual and translation rules of the French *Imparfait* is part of a larger task – the development of a multilingual annotated corpus of

[9] The (Non)Evidential Mood is an epistemic grammatical mood. It indicates that the utterance is based on what the speaker has/has not seen with their own eyes, or heard with their own ears.

[10] As e.g. in Vazov 1999

Tense and Aspect with rules for value identification and translation. As our examples and rules indicate, the corpus of aligned translations can be used not only to derive monolingual contextual rules (with or without rules for translation equivalence in a target language), but also to assign possible values in the source language based on translation equivalents.

The rules formulated by analyzing the aligned corpora of text will be tested in a system of automatic tense-and-aspect translation. The types of cross-language asymmetry can be integrated both in machine translation applications and in the test generating modules of the E-Platform. Student translations in the target language will be automatically tested against the target language equivalents of the corpus for appropriateness of tense-and-aspect values.

Our final objective in developing the corpora and providing input rules is to create an automatic or machine-assisted training system allowing:

1/ the choice between alternative values given an input of contextual markers;

2/ the proposal of contextual markers given an input of values;

3/ the choice between alternative target language Tense/Aspect values based on source text context analysis;

4/ the choice between source text values based on markers in the target text;

5/ error analysis and assessment of machine or student generated target texts.

References

Danchev & Alexieva 1974. Izborat mejdu minalo svarsheno i minalo nesvarsheno vreme pri prevoda na past simple tense ot angliyski na balgarski ezik. In : Yearbook of Sofia University, Faculty of classical and modern languages, vol. LXVII, 1, pp 249-329

Desclés 1985. J.-P, Desclés. Représentation des connaissances : archétypes cognitifs, schèmes conceptuels et schémas grammaticaux. Actes sémiotiques VII ; No 69-70

Desclés 1990 : J.-P, Desclés. The concepts of state, process, event and topology. General Linguistics, vol. 29, No 3. The Pennsylvania State University Press. University Park and London, 159-200

Desclés et al. 1997. J.-P, Desclés, E. Cartier, A. Jackiewicz, J.-L. Minel. Textual processing and contextual exploration method. In : CONTEXT'97, pp 189-197, Brasil, Rio de Janeiro

Guentcheva 1990. Zlatka Guentcheva. Temps et aspect: exemple du bulgare contemporain. CNRS, Paris

Guentcheva 1997. Imparfait, aoriste et passé simple : confrontation de leurs emplois dans des textes bulgares et français. In : J.-P. Desclés et al. 1997. Textual processing and contextual exploration method. In CONTEXT'97, pp 189-197, Brasil, Rio de Janeiro

Maire-Reppert 1991. Danièle Maire-Reppert. Les temps de m'indicatif du français en vue d'un traitement informatique: Imparfait. CNRS, Paris

Mani et al. 2001. I. Mani, L. Ferro, B. Sanheim, G. Wilson. Guidelines for annotating temporal information. In: Notebook Proceedings of Human Language Technology Conference 2001, pp 299-302, San Diego, California

Stambolieva 1997. TO BE and SAM in the systems of English and Bulgarian. PhD Dissertation, Sofia University Sv. Kliment Ohridski

Stambolieva 1998. "Context in Translation". Proceedings of the Third European Seminar "Translation Equivalence". Montecatini Terme, Italy, October 16-18 1997. The TELRI Association. Institut fur deutsche Sprache, Mannheim & The Tuscan Word Centre, pp. 197-204

Stambolieva 2008. Maria Stambolieva. Building Up Aspect. Peter Lang Academic Publishers

Stambolieva, Ivanova, Raykova 2018. M. Stambolieva, V. Ivanova, M. Raykova. A Platform for Language Teaching and Research (PLT & R). CLARIN annual conference, Pisa 2018

Vazov 1999. N. Vazov. Context-scanning strategy in temporal reasoning. In: Modeling and Using Context, CONTEXT Conference 1999, Springer-Verlag

Verkuyl 1993. Henk Verkuyl. A Theory of Aspectuality. The interaction between temporal and atemporal structure. Cambridge Studies in Liguistics. Cambridge University Press