

KOΠTE

KORPUSPROJEKT ZUR TRANSLATIONSEVALUATION

Presentation of the KOΠTE Corpus

version 2

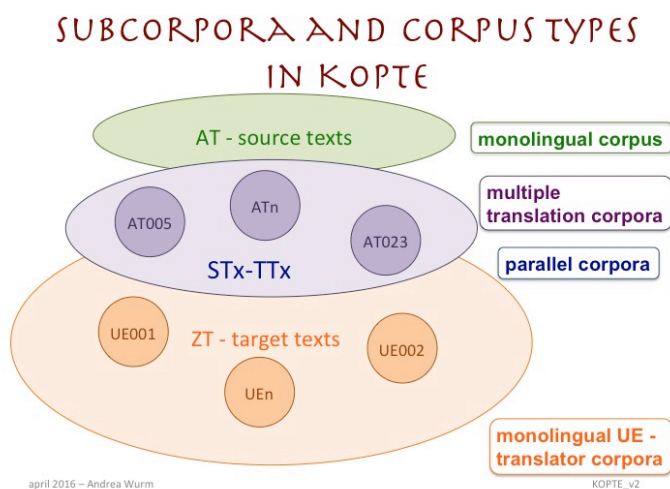
April 2016

Dr. Andrea Wurm
French Section
Institute for Applied Linguistics, Translation and Interpreting
Universität des Saarlandes
Saarbrücken
Germany

a.wurm@mx.uni-saarland.de
KOΠTE@mx.uni-saarland.de

1 Aim of the project

KOPTÉ will enable research on translation evaluation in a university training course for translators and to focus on student's translation problems as well as their problem solving. To achieve this goal, the work involves compiling a large corpus of student translations (Translation Learner Corpus) at Universität des Saarlandes. The languages covered are French and German. Thanks to financial support from the Institute for Applied Linguistics, Translation and Interpreting, the tedious task of transforming the texts into an electronic corpus and encoding them was facilitated by several student helpers. Additionally, students have the opportunity to prepare their final thesis (Diploma, BA, MA) in KOPTÉ and contribute to its annotation by doing so.



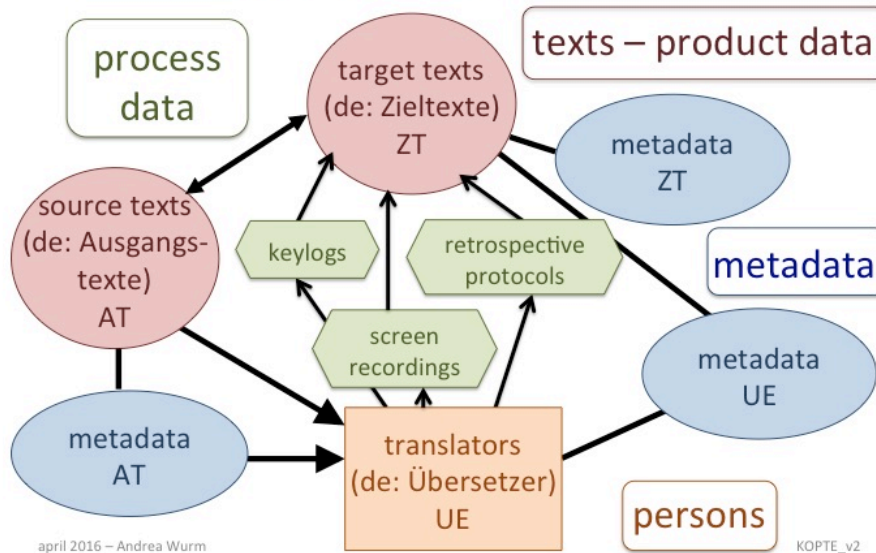
2 Corpus

The KOPTÉ corpus is a multiple Translation Learner Corpus, including many translations of the same source text performed by trainee translators. Our team began to compile the corpus in summer 2009 with the collection of the corrected translations of fourth-year students (in the so-called Klausurenkurs, a translation class (FR-DE) to prepare for final exams ("Diplom-Übersetzer" at the time)). This was continued in the following semesters until winter semester 2011/12. After the diploma course was cancelled in 2012, there were no more students and the class was given up. Students participated for two or more semesters in the class, so the corpus does not only reflect the student's examination level, but also their development in the time before facing final exams. At the moment, texts from BA and MA translation classes, which constitute a longitudinal corpus over the whole academic cycle of students, are being integrated in KOPTÉ.

Most of the students consented to research done on the basis of their translations; in the Klausurenkurs alone 58 different translators were involved. Researcher and student helpers had to type the first 38 source texts to transform the manuscript translations into a format to be processed. The corpus texts are UTF-8 plain text files, line break only LF. File names are composed of "AT" (Ausgangstext = source text) plus a three-digit text code and "UE" (Übersetzer = translator) and a three-digit translator code, e.g. the version of translator 34 of source text 23 gets the file name "AT023UE034". The source text in this case is "AT023UE000".

Many of the trainee translators filled in a form to deliver translator metadata like languages studied, time and mode of language learning, media use, parent's origin, etc.

DATA TYPES IN KOPTÉ



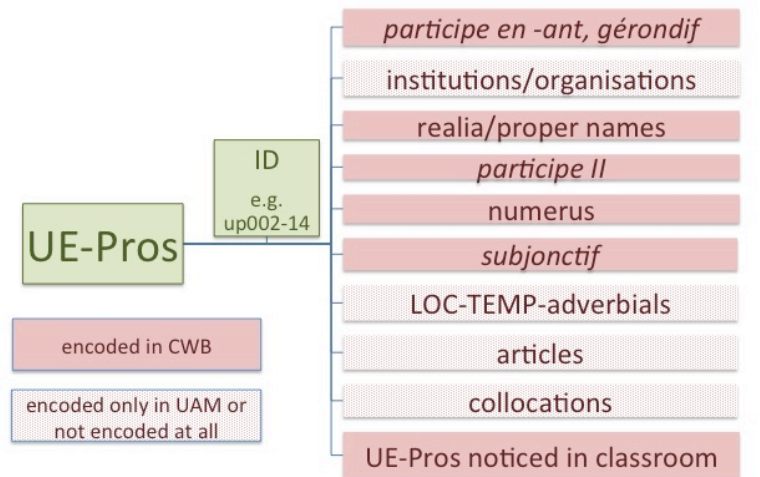
The KOPTÉ core corpus (KOPTÉ-KK) consists of Klausurenkurs texts (AT001-AT077), mainly German translations of newspaper articles from French newspapers, with a slight emphasis on opinion: commentaries or other richly structured texts were chosen deliberately because of the translation problems they present. The concept of the Klausurenkurs was to translate as much of the text (up to 2,400 characters) as possible in 45 minutes to achieve a solid translation. A monolingual dictionary was allowed. After moving into another room equipped with computers for every student (winter semester 10/11, starting with AT039), the students had free access to the Internet. For this reason, the texts no longer needed to have teacher notes with information in case of non-dictionary look-ups.

STUDENT GROUPS IN KOPTÉ

KK	<ul style="list-style-type: none"> • DÜ Diplom-Übersetzer • 2009 - 2012
BA	<ul style="list-style-type: none"> • BA VSLT • 2007 - ...
MA	<ul style="list-style-type: none"> • MA Translationswissenschaft • 2010 - ...
BA+MA	<ul style="list-style-type: none"> • BA VSLT + MA Transl.-wiss. • subgroup of BA and MA

At the moment, the core corpus is being complemented by translations of BA and MA translation students, constituting – apart from the analysis of translation problems and students’ translation solutions – a possibility for long-term studies of translation competence acquisition. KOPE-BA and KOPE-MA consist of different text types. The length of the texts is varying greatly from some lines (product packages) to complete 50-page reports on air quality in the Paris region in 2013. These translations are mostly not evaluated in terms of an evaluation scheme and a grade (except exam translations), but carry corrections from course sessions in form of insertions and deletions that are the result of discussion by students and the teacher. These corrections will be integrated into the translation texts as XML annotations. Furthermore, lemmatization, POS-tagging and the annotation of translation problems will enrich this subcorpus, similar to KOPE-KK (cf. 3 – Corpus Encoding).

TRANSLATION PROBLEMS UE-PROS

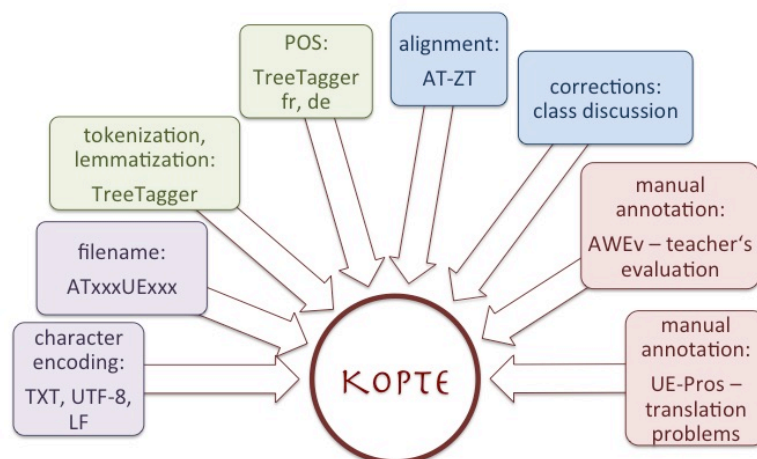


april 2016 – Andrea Wurm

KOPE_v2

3 Corpus Encoding

CORPUS ENCODING

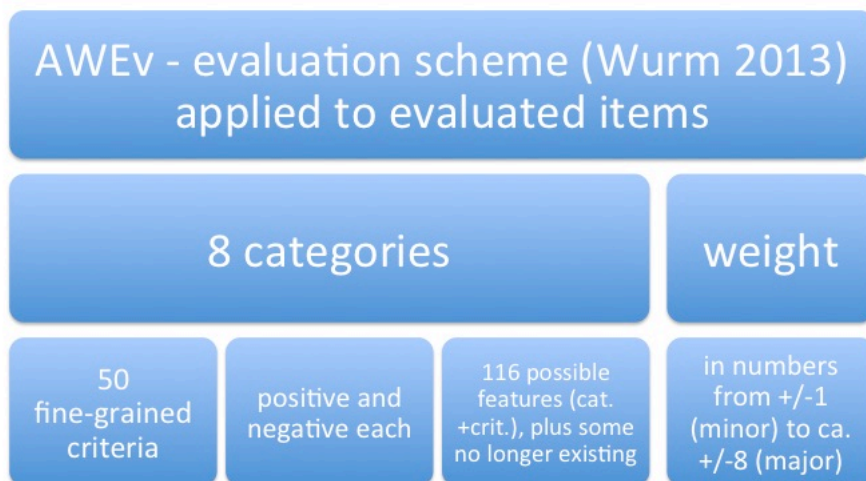


april 2016 – Andrea Wurm

KOPE_v2

Researchers perform automatic tokenization, lemmatization and POS-tagging with TreeTagger (Schmid 1994) and the STTS tagset for German, Achim Stein's tagset for French (Stein 2003). Researchers annotate manually both the teacher's evaluations and passages that have caused translation problems to several students. These translation problems (UE-Pros) are being annotated according to their nature in the source text and to the solutions found by the students. Teacher's evaluation is annotation layer AWEv (Andrea Wurm Evaluation) and consists of positive and negative evaluation, i.e. posev and negev, with numbers indicating the weight of evaluated items (annotation scheme see appendix). Manual annotation is done with UAM Corpus Tool (O'Donnell 2008), which allows free design of annotation schemes and comes with an easy-to-use GUI.

EVALUATION ANNOTATION



april 2016 – Andrea Wurm

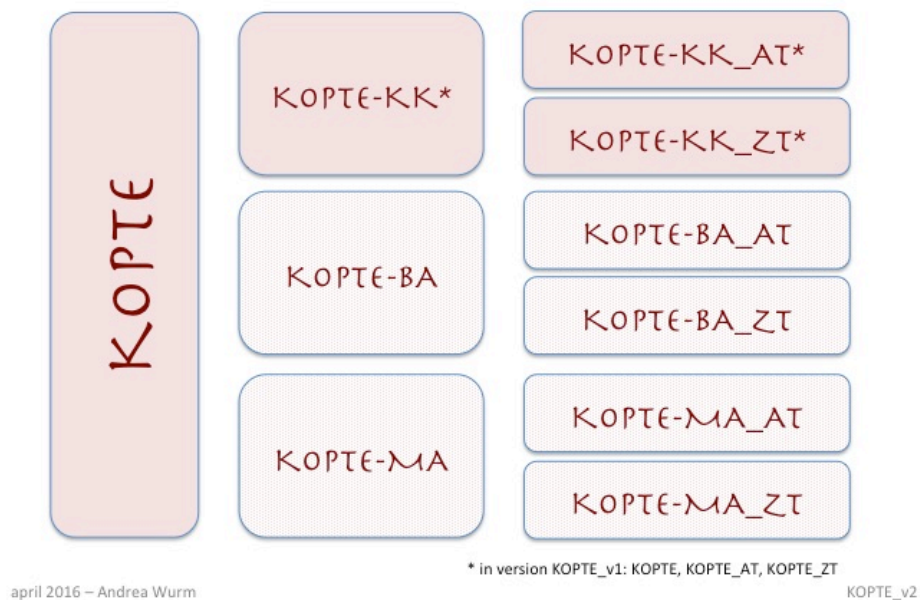
KOPTe_v2

Translator metadata are currently collected in an Excel spread sheet, but will be converted to XML and introduced into the corpus texts for query. The alignment is done with InterText (Vondřička 2010, based on Hunalign) and will be incorporated in KOPTe_v3. The KOPTe corpus is processed with the IMS Corpus Workbench (Evert/Hardie 2011) and queried with its built-in Corpus Query Processor (CQP). KOPTe_v2_KK is now available in CQPweb (Hardie 2012; restricted access) with AT001 through AT077, UE001 through UE058 (i.e. KOPTe-KK) and the following annotations:

- lemmatization (TreeTagger DE and FR)
- POS (TreeTagger DE and FR)
- AWEv (where available, otherwise only negev/posev without differentiation)
- UE-Pros IDs (up001-01, up001-02, ..., up038-05, etc.; not for participles/gerundium and subjunctif)
- UE-Pros classification and translation solutions for

- participles and gerundium (AT001-AT012; annotation layer “Partizipialkonstruktionen”, no IDs)
- past participles (AT001-AT012; annotation layer “Partizip_Perfekt”, UE-Pros ID in this layer)
- grammatical number (AT001-AT077; annotation layer “Numerus”, UE-Pros ID in this layer)
- realia/proper names (AT001-AT004; annotation layer “ER”, ID in layer UE-Pros)
- subjunctive (AT001-AT077; annotation layer “Subjonctif”, no IDs)
- translation problems noticed during correction of target texts (AT001-AT066; annotation layer “UE-Pros”, ID in this layer) with one-layer categorization (e.g. punctuation, realia/proper names, content transfer etc.)

CORPORA IN CWB

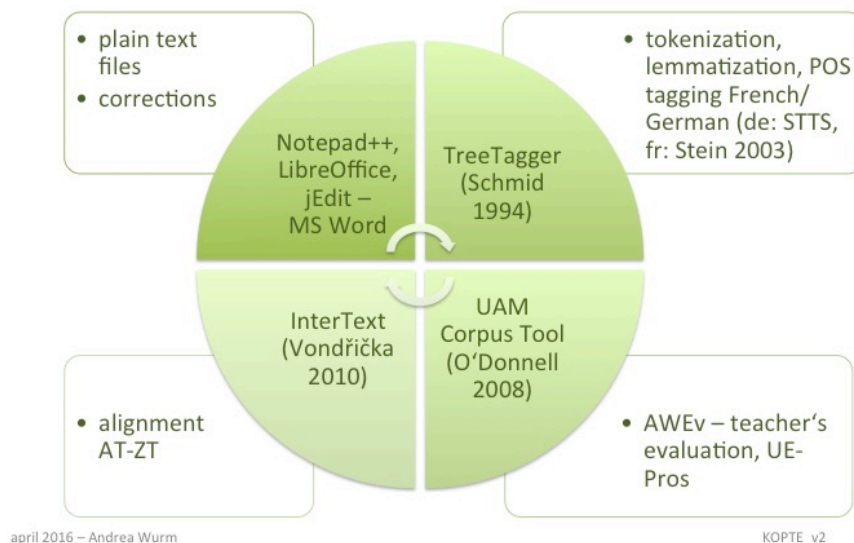


Source texts (KOPTE_v2_KK-AT) are stored up to now independently from target texts (KOPTE_v2_KK-ZT), which leads to two different CWB corpora. In KOPTE_v3, alignment asks for another corpus design.

Additionally, there are annotations already made, but not yet encoded in CWB:

- proper names of institutions/organizations (AT001-AT012)
- local and temporal adverbials (KOPTE-BA, KOPTE-MA)
- selected cohesion phenomena.

CORPUS ENCODING TOOLS



4 Research projects and published works

Since the start of the project, four diploma theses, one MA and six BA theses have been finished. Several papers are already published.

Nathalie Mattick (BA) delivered a fine analysis of French past participles and the translation strategies used by students. Katharina Redeker (BA) in turn shed some light on the subjunctive, a mode that is non-existent in German. Helen Klein (BA) did a contrastive study of article use, while Johanna Czibulinski (BA) queried collocates from the source texts and compared their translations. Anke Fechter (BA) ventured herself on the unlaboured field of translator profiles, seeking methods to use data from KOPE to get a grip on translator's personalities and competence acquisition. Vanessa Konzok (MA) was the first to work with BA/MA translations and tracked down progress in the development of translational competence using as an indicator the positioning of local and temporal adverbials in the sentence, a feature in which French and German differ quite a lot. Jana Stöckeler (DÜ) did an experiment with Translog (Jakobsen/Schou 1999), logging the translation process of AT026 for nine students also having handed in several translations to the corpus. Stefan Eich (DÜ) focused on proper names of institutions and organizations with regard to their problem causing potential, while Aleksandra Jagurinoska (DÜ) analysed grammatical number as a translation problem. Katja Palgen (DÜ) studied, in a contrastive way, French present participles and gerundium, analysing their rendering by different translators (see figure "Translation Problems – Example"). Her excellent diploma thesis was published as a monograph in 2011. She graduated as the best student of her year.

TRANSLATION PROBLEMS ENCODING

annotation layers with ID and categorisation	UE-Pros	problems arising from classroom corrections
		one-layer categorisation
	Partizip_Perfekt	solution encoding
	Numerus	solution encoding
annotation layers without ID, with categorisation	Partizipialkonstruktionen [participe en –ant, g�erondif]	solution encoding
	ER [realia/proper names]	ID in layer UE-Pros hierarchical categorisation in AT001-004
	Subjonctif	solution encoding

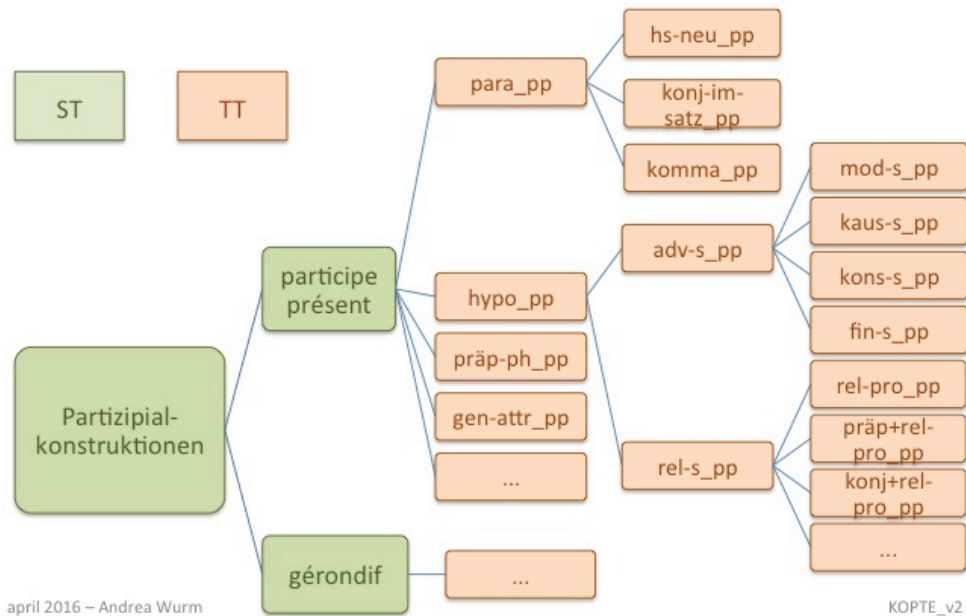
april 2016 – Andrea Wurm

KOPT E_v2

In several articles up to date, I have been using data from KOPT E. Wurm (2012) is a rather qualitative study of two cohesion phenomena, one from the corpus alone, and the other with translation process data from the Translog experiment (keystroke logging files, recorded and transcribed retrospective protocols). A quantitative study of proper names and culture specific items as a translation problem has been presented at the GAL conference 2011 in Bayreuth and is now published in transkom.eu (Wurm 2013a). Another paper deals with a German unique item, "Pronominaladverbien", and their cohesive use without prompting in the source text (Wurm 2014). With colleagues, we have been working on translation evaluation in human and MT settings, using KOPT E texts and my evaluations to compare with MT evaluation metrics (Vela, Schumann, Wurm 2014a, 2014b). Finally, this description of KOPT E_v2 is available online (Wurm 2015), based on the description already existing for Version 1 (Wurm 2013b).

Current research covers questions like the influence of translator's biography on translation quality, statistical analysis of evaluated items (Can we observe clustering of evaluation criteria?), establishment of translator and text profiles based on the metadata and the (evaluated) translations, possible correlations between text features and translation quality.

TRANSLATION PROBLEMS EXAMPLE FOR SOLUTION ENCODING



5 Conclusion

KOPTÉ is a corpus and a research project covering a huge variety of thematic approaches and allowing translation scholars to ask completely new questions. As far as long established strands of research are concerned, they can be reconsidered using empirical data not available beforehand. Students are keen on preparing their theses in KOPTÉ, because – as they tell me – they can immediately see the relevance of their research contributions. Future translators therefore do not only deliver their translations for research purposes. They get to work with them on their own while doing their first steps in the scholarly world.

To sum up, KOPTÉ can foster a multitude of work in the area of translation evaluation, translation competence (acquisition) and contrastive studies. It can be a node in the growing network of Translation Learner Corpora and bring forward pieces in the puzzle of results gained from them. This will help us to understand better what translation is and how it can be taught.

6 Project papers

Czibulinski, Johanna (2015): *Französische Substantiv-Adjektiv-Kollokationen und ihre Übersetzung ins Deutsche. Eine quantitative und qualitative Fallstudie anhand des KOPTÉ-Korpus und die Methodik.* unpublished Bachelor thesis, Institut für Angewandte Sprachwissenschaft sowie Übersetzen und Dolmetschen, Universität des Saarlandes, Saarbrücken

Eich, Stefan (2011): *Die Eigennamen von Institutionen und Organisationen in der Übersetzung Französisch-Deutsch. Eine Untersuchung anhand des KOPTÉ-Korpus.* unpublished diploma thesis, Institut für Angewandte Sprachwissenschaft sowie Übersetzen und Dolmetschen, Universität des Saarlandes, Saarbrücken

Fechter, Anke (2015): *Erstellung von Übersetzerprofilen.* unpublished Bachelor thesis, Institut für

Angewandte Sprachwissenschaft sowie Übersetzen und Dolmetschen, Universität des Saarlandes, Saarbrücken

- Jagurinoska, Aleksandra (2012): *Numeruskategorien als Übersetzungsproblem im KOPTÉ-Korpus (Französisch-Deutsch)*. unpublished diploma thesis, Institut für Angewandte Sprachwissenschaft sowie Übersetzen und Dolmetschen, Universität des Saarlandes, Saarbrücken
- Klein, Helen (2014): *Fehlerquellen bei Lernenden des Studiengangs Übersetzung hinsichtlich der Unterschiede im Sprachsystem Französisch zum Sprachsystem Deutsch: Artikelgebrauch im Sprachvergleich*. unpublished Bachelor thesis, Institut für Angewandte Sprachwissenschaft sowie Übersetzen und Dolmetschen, Universität des Saarlandes, Saarbrücken
- Konzok, Vanessa (2013): *Vom Bachelor zum Master – ein Einblick in die Entwicklungsstadien von Übersetzern*. unpublished Master thesis, Institut für Angewandte Sprachwissenschaft sowie Übersetzen und Dolmetschen, Universität des Saarlandes, Saarbrücken
- Mattick, Nathalie (2012): *Die Übersetzung von Partizipialkonstruktionen mit Participe passé aus dem Französischen ins Deutsche im Rahmen des Projekts KOPTÉ*. unpublished Bachelor thesis, Institut für Angewandte Sprachwissenschaft sowie Übersetzen und Dolmetschen, Universität des Saarlandes, Saarbrücken
- Palgen, Katja (2011): *Kontrastive Grammatik und Übersetzungsdidaktik. Übersetzungsprobleme und Lösungsstrategien beim Übersetzen von Gérondif und Participe présent aus dem Französischen ins Deutsche im Rahmen des Projekts KOPTÉ*. Saarbrücken: VDM Verlag
- Redeker, Katharina (2013): *Subjonctif als Übersetzungsproblem in den französisch-deutschen Übersetzungen des KOPTÉ-Korpus – Wie übertragen Übersetzungslerner den Subjonctif ins Deutsche?* unpublished Bachelor thesis, Institut für Angewandte Sprachwissenschaft sowie Übersetzen und Dolmetschen, Universität des Saarlandes, Saarbrücken
- Stöckeler, Jana (2010): *KOPTÉ: Korpusbasierte Translationsevaluierung – Übersetzungsprozess und Übersetzungsprodukt*. unpublished diploma thesis, Institut für Angewandte Sprachwissenschaft sowie Übersetzen und Dolmetschen, Universität des Saarlandes, Saarbrücken
- Vela, Mihaela; Schumann, Anne-Kathrin; Wurm, Andrea (2014a): "Beyond Linguistic Equivalence. An Empirical Study of Translation Evaluation in a Translation Learner Corpus." *Proceedings of HaCat Workshop at EACL 2014*. <https://sites.google.com/site/hacat2014/program/papers> (13.06.14)
- Vela, Mihaela; Schumann, Anne-Kathrin; Wurm, Andrea (2014b): "Human Translation Evaluation and its Coverage by Automatic Scores." *Proceedings of MTE Workshop at LREC 2014*. <http://mte2014.github.io/> (13.06.14)
- Wurm, Andrea (2012): "Kohäsive Makrostrukturen in Übersetzungen von Studenten auf Diplomniveau. Eine übersetzungsdidaktische Reflexion." Atayan, Vahram; Wienen, Ursula (eds): *Sprache – Rhetorik – Translation. Festschrift für Alberto Gil zu seinem 60. Geburtstag*. Frankfurt a.M.: Peter Lang, 431-440
- Wurm, Andrea (2013a): "Eigennamen und Realia in einem Korpus studentischer Übersetzungen (KOPTÉ)." *trans-kom* 6 [2]. <http://trans-kom.eu>. 381–419
- Wurm, Andrea (2013b): "Presentation of the KOPTÉ Corpus." describes KOPTÉ_v1, formerly online on <http://fr46.uni-saarland.de/index.php?id=3807>, now substituted for Wurm (2016)
- Wurm, Andrea (2014): "Kohäsion, Korpora und der Erwerb von Translationskompetenz. Text- und korpuslinguistische Analysen anhand des KOPTÉ-Korpus." Kerstin Kunz, Elke Teich, Silvia Hansen-Schirra, Stella Neumann, Peggy Daut (eds.): *Caught in the Middle – Language Use and Translation*. A Festschrift for Erich Steiner on the Occasion of his 60th Birthday. Saarbrücken: universaar Universitätsverlag des Saarlandes. 429–441
- Wurm, Andrea (2016): "Presentation of the KOPTÉ Corpus – Version 2." online <http://fr46.uni-saarland.de/index.php?id=3807>

7 References

- Evert, Stefan; Hardie, Andrew (2011): "Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium." *Proceedings of the Corpus Linguistics 2011 conference*, University of Birmingham, UK. http://cwb.sourceforge.net/doc_links.php
- Hardie, Andrew (2012): "CQPweb – combining power, flexibility and usability in a corpus analysis tool." *International Journal of Corpus Linguistics* 17 (3);

http://cwb.sourceforge.net/doc_links.php. 380-409

Jakobsen, Arnt Lykke; Schou, Lasse (1999): "Translog documentation." Hansen, Gyde (ed): *Probing the Process In Translation: Methods and Results*. Copenhagen Studies in Language, 24. Copenhagen: Samfundslitteratur. 149-184

O'Donnell, Michael (2008): "The UAM CorpusTool: Software for corpus annotation and exploration." Bretones Callejas, Carmen M. et al. (ed): *Applied Linguistics Now: Understanding Language and Mind / La Lingüística Aplicada Hoy: Comprendiendo el Lenguaje y la Mente*. Almería: Universidad de Almería. 1433-1447. <http://www.wagsoft.com/Papers/index.html>

Schmid, Helmut (1994): "Probabilistic Part-of-Speech Tagging Using Decision Trees." *Proceedings of International Conference on New Methods in Language Processing, Manchester, UK*. <http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/>

Stein, Achim (2003): French tagset for TreeTagger. <http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/french-tagset.html>

Vondřička, Pavel (2010): InterText – parallel text alignment editor. <http://wanthalf.saga.cz/intertext>

Appendix

Andrea Wurm's evaluation scheme = annotation scheme (AWEv)

The weight of evaluated items is indicated with numbers ranging from 1 (minor) to ~8 (major). The scheme is in German, but an English translation of categories and criteria is indicated in italics, following the German descriptor.

annotation feature (markings A. Wurm)	negev	posev
Formale Gestaltung <i>form</i>	n-form	p-form
Fa Absatzgestaltung <i>paragraphs</i>	nf-a	pf-a
Ff Formatierung <i>formatting</i>	nf-f	pf-f
Fg Makrostrukturelle Gliederungssignale <i>macrostructural hints</i>	nf-g	pf-g
Fi Interpunktion <i>punctuation</i>	nf-i	pf-i
Fl Layout <i>lay out</i>	nf-l	pf-l
Fn Nachbearbeitung <i>editing</i>	nf-n	pf-n
Fo Orthographie <i>orthograph</i>	nf-o	pf-o
Ft Typographie <i>typography</i>	nf-t	pf-t
Fu Umfang <i>text amount</i>	nf-u	pf-u
Struktur des Textes/Kohärenz <i>structure</i>	n-struktur	p-struktur
Sa Aufbau/Thematische Progression <i>logical structure/"thematic progression"</i>	ns-a	ps-a
Sb Abbildungen <i>illustrations</i>	ns-b	ps-b
Sk Herstellbarkeit von Kohärenz <i>coherence construction</i>	ns-k	ps-k
Sm Makrostruktur <i>macrostructure</i>	ns-m	ps-m
Ss Sequenzierung <i>sequencing</i>	ns-s	ps-s
St Themenentfaltung <i>development of theme</i>	ns-t	ps-t
Struktursignale/Kohäsion <i>cohesion</i>	n-kohaes	p-kohaes
Kb Bezugnahme <i>reference</i>	nk-b	pk-b
Kk Konnexion <i>connection</i>	nk-k	pk-k
Ks Struktur <i>structuring</i>	nk-s	pk-s
Stil/Register <i>stylistics/register</i>	n-register	p-register
Rf Rhetorische Figuren/Stilmittel <i>rhetorical figures</i>	nr-f	pr-f
Rn Norm <i>norm</i>	nr-n	pr-n
Rr Register/Textsorte <i>register/genre</i>	nr-r	pr-r
Rs Stilebene <i>style</i>	nr-s	pr-s

Grammatik <i>grammar</i>	n-gramm	p-gramm
Ga Artikel/Determinanten <i>determiners</i>	ng-a	pg-a
Gg Genus <i>gender</i>	ng-g	pg-g
Gk Konstruktion <i>construction (inflection etc.)</i>	ng-k	pg-k
Gm Modalität <i>modality</i>	ng-m	pg-m
Gs Satzbau/Wortstellung <i>syntax/word order</i>	ng-s	pg-s
Gt Tempus <i>tense</i>	ng-t	pg-t
Gw Wortbildung <i>morphology</i>	ng-w	pg-w
Lexik/Semantik <i>lexis/semantics</i>	n-lexik	p-lexik
Le Einbettung <i>semantic relations</i>	nl-e	pl-e
Lg Textsemantik <i>text semantics/meaning</i>	nl-g	pl-g
Li Idiomatik <i>idioms</i>	nl-i	pl-i
Lm Mengenangaben <i>quantities</i>	nl-m	pl-m
Lp Darstellungsperspektive <i>perspective</i>	nl-p	pl-p
Lt Terminologie <i>terminology</i>	nl-t	pl-t
Lu Un-Sinn <i>non-sense</i>	nl-u	pl-u
Lw Wortsemantik <i>word semantics</i>	nl-w	pl-w
Translatorische Probleme <i>translational problems</i>	n-trans	p-trans
Td Defekt im AT <i>defectuous source text</i>	nt-d	pt-d
Te Eigennamen/Realia <i>proper names/culture specific items</i>	nt-e	pt-e
Tf Spezielle Funktion eines Textelements <i>special function of a textual element</i>	nt-f	pt-f
Ti Ideologie <i>ideology</i>	nt-i	pt-i
Tl Lokalisierung/Kulturspezifik <i>localization</i>	nt-l	pt-l
Tm Mengenangaben/Zahlen <i>weights, measures etc.</i>	nt-m	pt-m
Tn Normvorschriften im Geltungsbereich <i>standards, laws etc.</i>	nt-n	pt-n
Tp Pragmatik <i>pragmatics</i>	nt-p	pt-p
Tr Recherche <i>documentation</i>	nt-r	pt-r
Tv explizite Vorgaben des Übersetzungsauftrags (mit Spezifizierung wie F, Lt, T, Rt etc.) <i>explicit postulations of the translation brief</i>	nt-v	pt-v
Tz Zitate/Anspielungen <i>citations/allusions</i>	nt-z	pt-z
Intention/Funktion <i>function</i>	n-intent	p-intent
If Funktionstypen <i>types of functions</i>	ni-f	pi-f
Iz Zielanordnung <i>goal dependence</i>	ni-z	pi-z