Proceedings of the Eighth Biennial International Conference of Greek Studies,
Flinders University June 2009". Flinders University Department of Languages - Modern Greek: Adelaide, 406-418.

Greek Idioms Processing in the Machine Translation System CAT2

Dimitra Anastasiou

This paper describes Machine Translation (MT) and the associated processing of idioms. Particularly, this research examines the rule-based CAT2 MT system and experiments with Greek sentences containing idioms. The paper also provides an in depth discussion of the resources and the procedure which have enhanced the translation of the quality of the idioms for the chosen German-Greek language pair. Greek is a morphologically rich language and the successful processing of Greek idioms within CAT2 has proven that MT can translate idioms correctly, whatever the level of language complexity.

1. Introduction

Machine Translation (MT) is a subfield of computational linguistics, where computer software is used to translate text or speech from one natural language to another. Idioms are, in most cases, multi-word expressions (MWEs) which are lexically fixed, semantically non-compositional, and syntactically, relatively fixed. The translation of idioms is very often a challenge to human translators; translating idioms with MT is a technological breakthrough.

MT started in the mid-1930s, when the first term regarding MT systems was introduced: “translating machines” by French-Armenian Georges Artsrouni and Russian Petr Petrovich Troyanskii (Hutchins, 2004; Hutchins & Lovtskii, 2000). In May 1951, Bar-Hillel[1] was appointed to conduct research at the Massachusetts Institute of Technology (MIT). In June 1952, he convened the first MT conference at MIT. Bar-Hillel collaborated with International Business Machines Corporation (IBM) on a project that resulted in the first public demonstration of an MT system on January 7, 1954.

In the next paragraphs we compare MT’s different architectures, and then introduce CAT2 in the subsection 1.1. CAT2 is a transfer-based multilingual MT system using

[1] Yehoshua Bar Hillel (1915–1975) was an Israeli logician and philosopher who contributed significantly to many linguistic fields, such as computational linguistics, MT, and Information Retrieval (IR).

morphosyntactic rules which are developed and used by academics at the Saarland University, Germany.

Machine Translation (MT) has various architectures: rule-based MT (RBMT), statistical MT (SMT), example-based MT (EBMT), and a hybrid. RBMT appeared around the 1950s, whereas SMT and EBMT, appeared around the 1990s. RBMT's pioneer is the above mentioned Bar Hillel, who in the early 1960s, introduced "Fully Automatic High Quality Machine Translation" (FAHQMT). The pioneer of EBMT is Makoto Nagao (1984) while Brown, Cocke, Pietra, Pietra, Jelinek, Mercer, and Roossin (1988) introduced the first statistical approach to MT. The similarities and differences between EBMT, SMT, and RBMT will now be discussed.

EBMT and SMT have many common points, but also some distinctive differences. The use of a bilingual corpus is a common point, but the SMT approach also makes use of a monolingual corpus. EBMT is distinct from SMT in that it contains symbolic translation knowledge and is not numeric in the form of a distortion and fertility probability model. A distortion and fertility probability model combines all parameters in the most likely manner, as SMT does. Moreover, SMT is based on word frequencies. By contrast, EBMT is based on word sequences and their combinations. More precisely, on one side, in SMT, the process has the following steps: part-of-speech (PoS) tagging, parsing of the tagged sentence to produce a parse tree, re-ordering of the tree randomly, inserting new words in the tree randomly, translating the words at the leaf nodes, and computing the probability of the output. Based on many iterations of the above, the system chooses the most likely translation (Yamada and Knight, 2001). On the other side, EBMT attempts to translate a sentence based on previous translation examples. It retrieves previous examples and identifies parts of these which are relevant to the current sentence. The translations of the corresponding parts are then extracted and re-combined to make a translation of the sentence.

In RBMT there are linguistic rules used, whereas in EBMT there are translation examples. Nagao (1988), supporting EBMT and opposing RBMT, claimed that in RBMT the linguistic features do not deal with all the phenomena occurring in the original text. EBMT, however, deals with real text, as it works on previous real examples, while RBMT focuses more on linguistic rules. The linguistic rules needed for RBMT use are complex, and thus linguistically trained staff is a requirement; the staff component, accordingly, needs time to produce the complex rules.

1.1. Introducing the rule-based MT system CAT2

CAT2[2] is a transfer-based rule-based MT (RBMT) system. Nowadays, Saarland University in Germany makes use of CAT2 in order to make it possible for future

[2] CAT stands for the concepts of *Constructors, Atoms and Translators.*

translators[3] to conduct experiments with several lexicons as well as syntactic and translation rules. CAT2 is adapted to the greatest possible extent, to real-life translation situations. A lot of European and international institutions lend a helping hand to enable the system to translate from and into several, not only western European languages, but also Arabic, Chinese, Japanese, Korean, and Russian. CAT2 is a unification and transfer-based multilingual MT that has been used since 1987 as an alternative to the EUROTRA (EURopean TRAnslation) software program. The EUROTRA was an ambitious MT project that was established and funded by the European Commission, from 1978 until 1992. Its goal was to construct an advanced multilingual transfer system for translation among all Community languages. In 1992 EUROTRA ended, after having achieved the success of increasing the research into computational linguistics. The book *Machine Translation and Greek Language* of Stavrou and Tzevelekou (2000) is divided into two parts: i) the system SYSTRAN of the European Union and ii) the system EUROTRA. Some articles deal with the construction of lexicons, terminology in MT, and linguistic specifications. Generally speaking, topics of Natural Language and Speech Processing are also included in this publication.

This paper's structure is as follows: in section 2 we describe the CAT2 MT flow and its two main grammar rule types, "generators" and "translators". The idiom processing concerning the German-Greek language pair within CAT2 is described in section 3 and more precisely, reference is made to dictionary entries (3.1), translation processes (3.2), and evaluation findings (3.3). A summary is provided in section 4.

2. CAT2 Machine Translation flow

The normal translation path, or the stages an input sentence passes through from the source language (SL) to the target language (TL) in the CAT2 system, will be discussed in this section. In a rule-based MT system, in general, the original text, otherwise called the source text, is first analysed, morphologically and syntactically, in order to obtain a syntactic representation. More precisely, this analysis includes part-of-speech tagging, lexical transfer via dictionaries, alignment, and structural transfer/chunking. This syntactic representation can optionally be refined to a more abstract level. The transfer process then converts this syntactic representation of the SL to a representation of the same level of abstraction in the TL. The SL and TL representations are referred to as "intermediate" representations. From the TL representation, similar stages of analysis, as for the SL, are followed in reverse, in order to generate the target text. This process dates back to Vauquois (1968).

When this normal translation flow fails, then the CAT2 system enters a robust mode, where each word in the input is translated individually, using the same translation path.

[3] The author contributed to the German ↔ Greek language pair during her postgraduate studies at the Saarland University.

In terms of specific CAT2 grammar rules, they are classified into "generators" and "translators", as seen in diagram 1.:

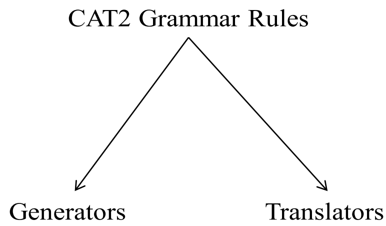

Diagram 1. CAT2 Grammar rules

The "generators" are sets of rules that define the well-formed structures of a representation level. They are a combination of constructors and atoms which make up a level of representation, where the smallest element of a level is called an "atom". In a tree schema, the constructors would represent partial trees, the atoms would represent the leaves, and the translators, any significant relations between the trees (Sharp, 1994).

A generator is introduced in the CAT2 grammar with the following representation level:

@level (LEVEL / TYPE / LANGUAGE)

According to Sharp (1994), a *LEVEL* is an atomic constant that gives a name to the level, a *TYPE* is either a morphological, a syntactic, or a relational structure, and *LANGUAGE* is an atomic constant naming the language being described by the level (1994:15). A constant is an atomic symbol used in programming languages and can be any sequence of alphanumeric characters, including the underscore _, as well as a beginning that starts with a lower case letter or any string of characters delimited by single quotes.

Within CAT2 there are three structures of generators:

a) Morphological structure (MS);
b) Constituent / Syntactic structure (CS); and
c) Interface / Relational structure (IS).

The first two structures, MS and CS, describe the grammar of a language, starting from words (MS) and graduating to phrases and sentences (CS). The morphological generator level analyses the words in a text or generates words during text synthesis. Its purpose is the definition of the structure of a word in terms of its morphemes. Generally speaking, the generator of the MS has to define all the well-formed morphological structures.

The CS deals with phrases and the whole sentence. The tokens delivered by a morphological generator level are used as input by the syntactic generation level. Its aim is to build a constituent structure of a language during the analysis of the input text. In synthesis, it is used to license the provisional structures created by a translator.

An analysis of well-formed sentences is required for the translation process. Therefore, it is useful to examine further each representation level, both with a focus on the degree of its well-formedness and via a transfer of representations in one language, from one level to another. That is the functionality of an interface between SL and TL. The interface structure (IS) depicts the semantic structure, or the predicate-arguments structure. According to the predicate-arguments grammar, the sentence consists of the verb, its arguments, and adjuncts. The relational level defines an interface structure of a phrase or a sentence. Two languages are then related via their interface structures rather than their syntactic structures (Sharp, 1994).

The structures of generators or representation levels and their order can be seen below in the Figure 1. The following figure shows the transfer-based approach (Sharp, 1994:3).

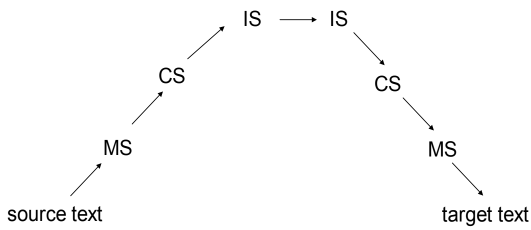

Figure 1. Representation levels in CAT2

It is noteworthy that Kaplan and Bresnan (1982:175–231) proposed a lexical-functional grammar (LFG) with two descriptions for every sentence of the language which it generates, namely c-structure and f-structure. According to LFG grammar, the constituent structure (or "c-structure") of a sentence is a conventional phrase structure (PS), which is represented as an ordered tree graph. It indicates the surface grouping and ordering of words and phrases in a sentence. The functional structure (or "f-structure") provides a more detailed representation of grammatical relations between words and phrases, as traditionally expressed by subject, direct object, and so on. The representation in f-structures deals with agreement by using features like number and gender.

Grammar rules or "translators" are rules which map the structures at one representation level to structures at an adjacent level. A translator is introduced with the following representation level:

@level: @level (LEVEL1 ⇔ LEVEL2)

LEVEL1 and *LEVEL2* are the names of the syntactic and/or relational level generators. The ⇔ symbol that separates the two level names, symbolises that the translator level is bidirectional. The levels named in the *@level* directive must correspond to the syntactic and relational generator levels (Sharp, 1994:17).

Four rule types are included in CAT2 grammar, two for generators (b-rules, f-rules) and three for translators (t-rules, tf-rules, mw-rules) (Streiter, Schmidt Wigger, Haug and Sharp, 1994:10):

a) Building rules (b-rules)

These rules describe the possible structures at the level in question.

b) Feature rules (f-rules)

The f-rules describe the feature content of the generated structures. They apply feature constraints to objects created by b-rules. F-rules cannot modify phrase structures in any way, nor can they delete or change existing features, but they can add features.

c) Transfer rules (t-rules)

Transfer rules transform structures from *LEVEL1* to *LEVEL2* or from *LEVEL2* to *LEVEL1*. T-rules may be bidirectional or unidirectional.

d) Transfer feature rules (tf-rules)

Tf-rules are analogous to f-rules in a generator, in the sense that they manipulate features, not structures. They transfer features from source structures to target structures. Tf-rules are also bidirectional.

e) Multi-word rules (mw-rules)

The multi-word rules transfer word lists to other word lists and can be bidirectional. They identify lexical units which are compounds, words which consist of more than one lexeme. These lexemes connect with each other to generate a word node which can be now analysed by the syntactic parser, for example, the French MWE *pomme de terre* (*Kartoffel*):

```
MWE = [{lex=pomme},{lex=de},{lex=terre}]  <=> [{lex=pomme_de_terre}]
```

Mw-rules can also separate into single parts of the contractions of article and preposition, for example, in Greek (*sto-στο*):

```
Cont = [{lex=p,string=sto}]  <=>
                    [{lex=p,string=se},{lex=det,string=to}]
```

where "lex" is specified as a preposition "p" and a determiner/article "det". Thus, the contraction *sto* (*στο*) consists of the preposition *se* (*σε*) and the article *to* (*το*).

3. Idiom Processing Study in CAT2

This section looks at idiom processing within the CAT2 system. More precisely, we present the dictionary entries (subsection 3.1) and describe their translation process through MT (3.2). An evaluation of the idiom processing is found in subsection 3.3.

3.1. Dictionary entries

11 idiom dictionary entries (listed below) have been added to the Greek-German transfer dictionary and correspondingly 11 sentences, containing an idiom, to the Greek corpus; each sentence contains an idiom stored in the dictionary. 7 out of 11 sentences contain continuous idioms and 4 discontinuous idioms. A distinction has been made between continuous idioms which have adjacent constituents, and discontinuous idioms which have non-adjacent constituents (gaps between constituents).

The sentences have been manually constructed, though different from each other, both regarding morphology and syntax, in favour of diversity and the avoidance of tailoring or "foreseeing" the good evaluation results, by entering similar sentences. The idiomatic expressions have been classified into the following syntactic categories: noun phrase (NP), preposition phrase (PP), and verbal phrases (NP plus Verb and PP plus Verb).

a) NP

das A und O
to alfa kai to wmega (το άλφα και το ωμέγα)
the end-all and be-all

der Stein des Anstoßes
h petra tou skandalou (η πέτρα του σκανδάλου)
bone of contention

tote Hose
psofia pragmata (ψόφια πράγματα)
nothing doing

b) PP

mit Müh und Not
me ta chilia zoria (με τα χίλια ζόρια)
limpingly/reluctantly

c) NP-V

seinen Kopf durchsetzen
pataw podi (πατάω πόδι)
get one's way

reinen Tisch machen
ksekayarizw logariasmous (ξεκαθαρίζω λογαριασμούς)
get things straight

Eindruck schinden
kanw figoura (κάνω φιγούρα)
impress

die Zeit totschlagen
skotwnw muges (σκοτώνω μύγες)
kill time

d) PP-V

im siebten Himmel sein
eimai ston ebdomo ourano (είμαι στον έβδομο ουρανό)
be in seventh heaven/be on cloud nine

auf die Nase fallen
spaw ta moutra mou (σπάω τα μούτρα μου)
come a cropper

auf die falsche Karte setzen
pontarw se lathos xarti (ποντάρω σε λάθος χαρτί)
bet on the wrong horse

3.2. Idiom translation process

Within CAT2, the necessary resources to translate idioms from one language to another are the grammars of each language, the transfer dictionaries, and the multi-word expression rules. These three stages are explained below. An SL corpus is helpful, as users do not have to type the input examples; rather, they have them stored in the corpus.

a) Addition of the dictionary entries to the transfer dictionary; these entries correspond to *t*-rules on the interface/relational level. Presented below are an NP idiom dictionary entry (i), a PP idiom (ii), and an idiomatic VP (iii):

   (i) *stein = {lex='h petra tou skandalou'}.[] <=> {lex='der Stein des Anstoßes'}.[].*

   (ii) *müh = {lex='ta chilia zoria'}.[] <=> {lex='Müh und Not'}.[].*

   (iii) *tisch = {lex=ksekayarizw,frame= {arg1={semf=pers}, arg2={nlu=logariasmos, agr={num=plu}}}}.[] <=> {lex=machen,frame= {arg1={semf=pers}, arg2={lex='reinen Tisch', agr={num=sing}}}}.[].*

   It is noteworthy that PPs are "coded" as NPs and then, during the translation process, the preposition is automatically added in (ii) based on the word-for-word translation. The entry found in (iii) will now be discussed in more detail. *Logariasmos* literally means *Rechnung/bill*, whereas *ksekayarizw* has only a figurative meaning: "reinen Tisch machen / to get things straight".[4] In (iii) we

[4] However, the idiom "ksekayarizw logariasmous" corresponds more clearly to the German idiom "reinen Tisch machen" than the one-word verb "ksekayarizw" does alone.

"match" the Greek verb with the German one (*ksekayarizw-machen*), and the Greek noun with the German NP (*logariasmos-reinen Tisch*). Although, literally seen, there is no translation equivalence, this kind of matching serves our purposes and leads to a successful translation result. The defined agreement in (iii) should be noted, as the Greek noun *logariasmos,* as an idiom, occurs in the plural form, whereas the German correspondent in this idiom, *Tisch*, is represented as a singular number. This method should not be generalisable to the non-idiom cases, but only to multi-word expressions whose components are connected with each other. To give an example, *clean the table* should not be given as an entry, because *clean* can come with other *nouns* too, and semantically these components do not form an entity. The MT process can handle the *clean the table* in the same way as *clean the room/carpet*. It should be pointed out that the idiom's single parts should not necessarily be stored in the dictionary in order to obtain the correct idiomatic translation, unless we want to test whether CAT2 translates correctly, for example, with the phrase, *plhrwnw logariasmous* (Rechnungen bezahlen/pay bills). In this case the following rules must also be added to the dictionary:

logariasmos1=
{lex=logariasmos,vlex~=ksekayarizw}.[] <=> {lex=’Rechnung’}.[].
logariasmos2= {lex=logariasmos,vlex=ksekayarizw,agr={num=plu}}.[]
<=> {lex=’reinen Tisch’,agr={num=sing}}.[].

The "logariasmos2" rule makes clear that *logariasmos* should be translated as *Rechnung* only under the condition that the verb can be anything apart from *ksekayarizw*. By contrast, the "logariasmos1" rule lacks this constraint.

b) Construction of rules at the morphology level in the Greek (i) and correspondingly in the German (ii) grammar:

(i) Greek Grammar
petra = {role=gov,lex=’h petra tou skandalou’,lemma=’h petra tou skandalou’,cat=n, gen=fem,agr={num=sing},semf=abs}.[].
zoria = {lex=’ta chilia zoria’,lemma=’ta chilia zoria’,cat=n,case=acc}.[].

(ii) German Grammar
stein = {role=gov,lex=‘der Stein des Anstoßes‘,cat=n,gen=masc, agr={num=sing}, semf=abs}.[].
müh = {lex=‘Müh und Not‘,lemma=‘Müh und Not‘,cat=n}.[].

As for the idiomatic VPs, these are not stored as continuous strings, as the verb can be permutated with consequent, syntactically discontinuous, phenomena. Hence the verbs and their nominal/prepositional part are separately stored as normal entries.

c) Constructions of multiword-rules in the German grammar:

stein={lex=’der_Stein_des_Anstoßes’,cat=n} <==>
[{lex=art,lemma=der,cat=det},
{lex=’Stein’,lemma=’Stein’,cat=n},
{lex=art,lemma=des,cat=det},
{lex=’Anstoßes’,lemma=’Anstoßes’,cat=n}].

müh={lex=’Müh_und_Not’,cat=n} <==>
[{lex=’Müh’,lemma=’Müh’,cat=n},
{lex=und,lemma=und},
{lex=’Not’,lemma=’Not’,cat=n}].

tisch={lex=’reinen_Tisch_machen’,cat=v, transitivity=trans} <==>
[{lex=reinen,lemma=reinen,cat=a},
{lex=’Tisch’,lemma=’Tisch’,cat=n},
{lex=machen,lemma=machen,cat=v}].

These multiword-rules aim to bring the idiom’s constituents together, in order to be seen as an entity. Also the parts-of-speech are added to help the morphological generation.

It is noteworthy that the multiword-rules are needed only in the grammar of the TL; thus if we translate from Greek into German, then multiword-rules suffice only in the German grammar.

3.3. Evaluation findings

Five examples tested within CAT2 and their translation outputs are presented below:

1. *H petra tou skandalou einai h gunaika* (Η πέτρα του σκανδάλου είναι η γυναίκα)
   Der Stein des Anstoßes ist die Frau (*The bone of contention is the woman*)[5]

2. *Autos diabazei ena biblio me ta chilia zoria* (Αυτός διαβάζει ένα βιβλίο με τα χίλια ζόρια)
   Er liest ein Buch mit Müh und Not (*He is reading a book reluctantly*)

3a. *Egw ksekayarizw logariasmous* (Εγώ ξεκαθαρίζω λογαριασμούς)
Ich mache reinen Tisch

3b. H gunaika ksekayarizei logariasmous (Η γυναίκα ξεκαθαρίζει λογαριασμούς)
Die Frau macht reinen Tisch (The woman gets things straight)

[5] The German sentences are the output; the English sentences are only provided here for the non-German speakers.

3c. *Egw ksekayarizw shmera logariasmous* (Εγώ ξεκαθαρίζω σήμερα λογαριασμούς)
Ich mache heute reinen Tisch (*I get things straight today*)

All idioms, both continuous and discontinuous ones, are correctly translated. Hence the evaluation of CAT2 based on our resources brought very successful results, reaching 100% recall, as well as precision rates. Also, when the verb was conjugated in (3b), there were no encountered problems. This finding contrasted markedly with outcomes obtained using commercial MT systems (Anastasiou, 2008). The translation of sentences with discontinuous idioms (3c) is successfully performed, as long as the aforementioned appropriate multiword-rules are available.

It should be pointed out that the complexity of matching and accordingly, translating discontinuous phrases, is much higher than the complexities encountered when matching continuous phrases. Matching a discontinuous phrase of length *m* on a sentence of length *n* may lead to a huge number of retrieved entries in the order of:

$$O\binom{n}{m}$$

By contrast, for continuous phrases, there is a maximum of ($n - m + 1$) matches. For example, a discontinuous phrase of 5 words on a 15-word sentence, can be matched in more than 3,000 possible ways, whereas a continuous phrase may lead to 11 possible matches (Carl, 2007:67).

4. Summary

In this paper the unification and transfer-based multilingual MT system CAT2 that has been used since 1987 at Saarland University for research purposes, has been described. More information about CAT2 system can be found in Haller (1993), Sharp (1994), and several other publications.[6] The author has improved the Greek into German language pair in relation to idiomatic processing. After the research revealed differences between EBMT, SMT, and RBMT, specific rule sets of the RBMT system, CAT2, with respect to the generators and translators, were developed. Idiom dictionary entries were listed and put through the stages of the actual idiom processing translation path, adding analogous examples. The paper closed with an evaluation of the German translation output of Greek input sentences, containing both continuous and discontinuous idioms.

6 All publications related to CAT2 are at: http://www.iai-sb.de/forschung/content/view/51/63/ 2 August 2010.

Bibliography

Anastasiou, 2008

Dimitra Anastasiou, "Identification of idioms by MT's hybrid research system vs. three commercial systems". In: *Proceedings of the European Machine Translation Conference* (*EAMT*), 22nd–24th September, Hamburg, Germany.

Brown, Cocke, Pietra, Pietra, Jelinek, Mercer, Roossin, 1988

Peter Brown, John Cocke, Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, Robert L. Mercer, Paul S. Roossin, "A Statistical Approach to Language Translation", in: *Proceedings of the 12th Conference on Computational Linguistics* (*COLING*) 1988, Budapest, Hungary, 71–76.

Carl, 2007

Michael Carl, "METIS-II: The German to English MT System". In: *Proceedings of MT Summit* XI, ed. Bente Maegaard: 65–72. Copenhagen, Denmark.

Haller, 1993

Johann Haller, "*CAT2 – Vom Forschungssystem zum präindustriellen Prototyp*". In: *Sprachtechnologie: Methoden, Werkzeuge, Perspektiven 1993,* ed. H. P. Pütz, J. Haller. 282–303: Hildesheim.

Hutchins, 2004

William John Hutchins, "Two precursors of machine translation: Artsrouni and Trojanskij". In: *International Journal of Translation*, John Benjamins, Vol. 16 (1), 11–31.

Hutchins and Lovtskii, 2000

William John Hutchins and Evgenii Lovtskii, "Petr Petrovich Troyanskii (1894–1950): a forgotten pioneer of mechanical translation". In: *Machine Translation* 15 (3): 187–221, The Netherlands, Kluwer Academic Publishers.

Kaplan and Bresnan, 1982

Ronald M. Kaplan and Joan Bresnan, "Lexical-Functional Grammar: A formal system for grammatical representation". In: *The Mental Representation of Grammatical Relations*, Joan Bresnan. 173–281: Cambridge, Mass.: MIT Press.

Nagao, 1984

Makoto Nagao, "A Framework of a Mechanical Translation between Japanese and English by Analogy Principle". In: A. Elithorn, R. Banerji (eds), *Artificial and Human Intelligence*, Amsterdam, North-Holland, 173–180.

Nagao, 1988

Makoto Nagao, "Language engineering: the real bottle-neck of natural language processing". In: *Proceedings of the 12th COLING 1988*. 448: Budapest, Hungary.

Sharp, 1994

Randall Sharp, *CAT2 Reference Manual*, Version 3.6, Institut der Gesellschaft zur Förderung der Angewandten Informationsforschung e.V. an der Universität des Saarlandes, Saarbrükken, Germany.

Stavrou and Tzevelekou, 2000

Μελίτα Σταύρου και Μαρία Τζεβελέκου, *Η Μηχανική Μετάφραση και η Ελληνική Γλώσσα*. Athens: Καστανιώτης.

Streiter, Schmidt Wigger, Haug and Sharp, 1994

Oliver Streiter, Antje Schmidt Wigger, Ute Haug and Randall Sharp, "Getting started with CAT2 Machine Translation System", http://www.iai-sb.de/docs/getting_started.pdf

Yamada and Knight, 2002
Kenji Yamada and Kevin Knight, "A Decoder for Syntax-Based Statistical MT" (K.). In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (ACL), 2002, Philadelphia, Pennsylvania, 303–310.

Vauquois, 1968
Bernard Vauquois, "A survey of formal grammars and algorithms for recognition and transformation in mechanical translation". In: *International Federation for Information Processing*, Vol. 2, Edinburgh, UK, 1114–1122.