# Translation Divergence in Machine Translation

Feng Zhiwei
Applied Linguistics Institute, Ministry of Education of China (ALI/MOE)
Chaonei Nanxiaojie 51
100010 Beijing, China
Tel: +86-10-6526-7983
Email: zwfengde@public.bta.net.cn

## Abstract

The selection of translation equivalence in MT (Machine Translation) depends on the differentiation of translation divergence between the Source Language (LS) and Target Language (TL). In this paper, the different types of translation divergence in MT are discussed. They are the translation divergence in lexical selection, in tense, in thematic relation, in head-switch, in structure, in category, and in conflation. The syntactical, semantic and contextual ambiguity that related with the translation divergence also discussed. The author suggests use the feature vector to represent the co-occurrence cluster, and the co-occurrence cluster based approach in the selection of translation equivalence is described in detail.

**Key Words**

Machine Translation (MT), Source Language (SL), Target Language (TL), translation equivalence, translation divergence, feature vector, co-occurrence cluster

In Machine Translation (MT), we must correctly select the adequate translation equivalence between Target Language (TL) and Source Language (SL). However, the selection of translation equivalence depends on the differentiation of translation divergence. In this paper, we shall discuss the translation divergence in machine translation.

Generally speaking, A MT can include three phases: analysis of SL, transfer from SL to TL, generation of TL.

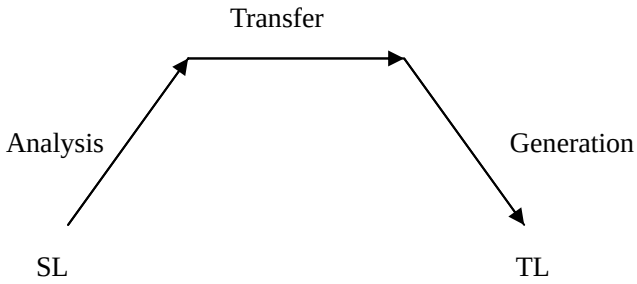



Fig.1. Three phases of MT

The translation divergence exists in all process of MT. In our paper, we shall concentrate to discuss the translation divergence in the transfer. Some problems in analysis and generation that related with transfer are also discussed.

.

**1. Translation Divergence in MT**

1.1 Translation Divergence of Lexical Selection in TL:

In English-Chinese MT, same English word can have different translation equivalence in Chinese.

| E.g. English | Chinese |
|---|---|
| They *play* piano : | “他们*弹*钢琴”(Tamen *tan* gangqin). play – 弹(tan) |
| They *play* violin. : | “他们*拉*小提琴”(Tamen *la* xiaotiqin). play – 拉(la) |
| They *play* football : | “他们*踢*足球”(Tamen *ti* zuqiu). play – 踢(ti) |
| They *play* basketball : | “他们*打*篮球”(Tamen *da* lanqiu) play – 打(da0 |

A Chinese generator in MT must select the appropriate target-language words “tan [弹], la [拉], ti [踢], da [打]” from general notion as “play” in English respectively.

Additional information is required for choosing the relevant term from each target-language-pair.

E.g. In German-English MT,

German: “*kennen*”

English: “*know*” or “*understand*”.

For word “kennen” in German, MT system must select different translation equivalence “know” or ”understand” in accordance with different context in English.

The Selection of Translation Equivalence (STE) is an important problem in MT.

1.2 Translation Divergence in Tense

E.g. Chinese: “我被杭州的风景*吸引住了*”(“Wo bei hangzhou de fengjing *xiyinzhule*.“)

English: “I *was captivated* by the scenery of Hangzhou“

“I *am captivated* by the scenery of Hangzhou”

In Chinese, the tense information (present, past, future) is not overt. The information used to select English language tense depends entirely on the context of the utterance. The second English sentence would be transferred if the speaker is looking at the scenery at the time of speech.

1.3 Translation Divergence in Thematic Relation: Thematic divergence involves an exchange of subject and object positions.

E.g. In Chinese-English MT,

| Chinese | English |
|---|---|
| 王冕死了*父亲* (Wang Mian sile *fuqin*): | Wang Mian’s *father* died |
| OBJ | SUBJ |

“fuqin[父亲]” appears as OBJ in Chinese but the equivalence word “father” appears as SUBJ in English.

In English-Chinese MT,

| English | Chinese | |
|---|---|---|
| I read *this book* : | *这本书*我读了(*Zhe ben shu* wo dule). | |
| OBJ | SUBJ | SUBJ |

I eat *this bread* : 这个面包我吃了(*Zhe ge mianbao* wo chile).
OBJ SUBJ SUBJ
"this book' and "this bread" appear as OBJ in English, but the equivalence words "zhe ben shu [这本书]" and "zhe ge mianbao [这个面包]" appear as SUBJ in Chinese.

In English-Spanish MT,
English Spanish
*I* like *Mary* *Mary* *me* gusta ("Mary [to] me pleases")
SUBJ OBJ SUBJ OBJ
"Mary" appears as OBJ in English but as SUBJ in Spanish; analogously, the SUBJ "I" in English appears as OBJ "me" in Spanish.

1.4 Head-Switching Divergence: this divergence occurs commonly across language pairs. In such cases, a main verb (MV) in the SL is subordinated in the TL.
E.g. In English-Chinese MT,
English: We *had* much *snow*.
MV Noun
Chinese: 我们这里下雪很多(women zheli *xiaxue* henduo)
Verb Verb
Main verb "had" in English disappeared in Chinese and noun "snow" in English is realized as verb "xiaxue [下雪]"in Chinese.
E.g. English: My house *is* only five minutes' *walk* from the station.
MV Noun
Chinese: 我的房子离车站只需要走五分钟。(Wode fangzi li chezhan zhi xuyao *zou* 5 fenzhong)
Verb Verb
Main verb (MV) "is" in English disappeared in Chinese, and noun "walk" in English is realized as verb "zou [走]" in Chinese.
E.g. English: My watch may *be* one or two minutes *fast*.
MV Adverb
Chinese: 我的手表也许快了两分钟。(Wode shoubiao yexu *kuaile* liang fengzhong)
Verb Verb
Main verb "be" in English disappeared in Chinese, and adverb "fast" in English is realized as verb "kuaile [快了]" in Chinese.

E.g. English: My watch may *be* one minute slow.
MV Adverb
Chinese: 我的手表也许慢了一分钟。(wode shoubiao yexue *manle* yi fenzhong)
Verb Verb
Main verb "be" in English disappeared in Chinese, and adverb "slow" in English is realized as verb "manle [慢了]" in Chinese.
E.g. In English-German MT,
English: "I *like* to eat"
MV

German: “Ich esse *gern*“ (“I eat likingly“)

Adverb

Main verb “like” in English is realized as an adverbial modifier “gern” in German.

E.g. In English-French MT,

English: “The baby *just* fell”

Adverb

French: “Le bebe *vient de* tomber” (“The baby come (verb-past) of fell”)

MV

The English “just” is translated as the French main verb “venir” which takes the falling event as its complement “de tomber”.

1.5 Structural Divergence: In structural divergence, a verbal argument in SL has a different syntactic realization in the TL

E.g. In English-Chinese MT,

English: She looked down *on the office girls*.

PP

Chinese: 她轻视办公室的女孩 (Ta qingshi *bangongshi de nuhai*)

NP NP

E.g. English: I was looking forward *to the festival*.

PP

Chinese: 我正在盼望着节日 (Wo zheng panwangzhe *jieri*)

NP NP

E.g. English: I will make up *for the lost time*.

PP

Chinese: 我将弥补丢失的时间 (Wo jiang mibu *diushi de shijian*)

NP NP

The PP arguments in English are transferred to NP arguments in Chinese.

E.g. In English-Spanish MT,

English: “John entered *the house*”

NP

Spanish: “Juan entro *en la casa*” (“John entered *in the house*”)

PP

The verbal object is realized as a noun phrase (“the house”) in English and as a prepositional phrase (“en la casa”) in Spanish.

1.6 Category Divergence: Category divergence involves the selection of a TL word that is a categorical variant of the SL equivalent.

E.g. In English-Chinese-German MT,

English: “I am *hungry*”

Adj

Chinese: 我饿了 (Wo *ele* )

Verb Verb

German: “Ich habe *Hunger*” (“I have *hunger*”)

Noun

The predicate is adjective (“hungry“) in English but verb (“ele [饿了]”) in Chinese, and noun (Hunger) in German. This change forces the transfer program to select a different main verb in TL

E.g. In English-Chinese MT,

English: *I* *have* a terrible *toothache*.

SUBJ MV Noun

Chinese: 我的 牙 痛 得 可怕 (*Wode ya tong* de kepa)

Modifier Noun Verb

The main verb “have” in English disappeared in Chinese, and the noun “toothache” in English is transferred to “Noun + Verb” (ya[牙] + tong[痛]). The subject “I” is transferred to a modifier of noun “ya[牙]” – “wode [我的]”.

E.g. English: “John *is* very *fond* of music”

MV Adj

Chinese :约翰很喜欢音乐 （Yuehan hen *xihuan* yinyue）

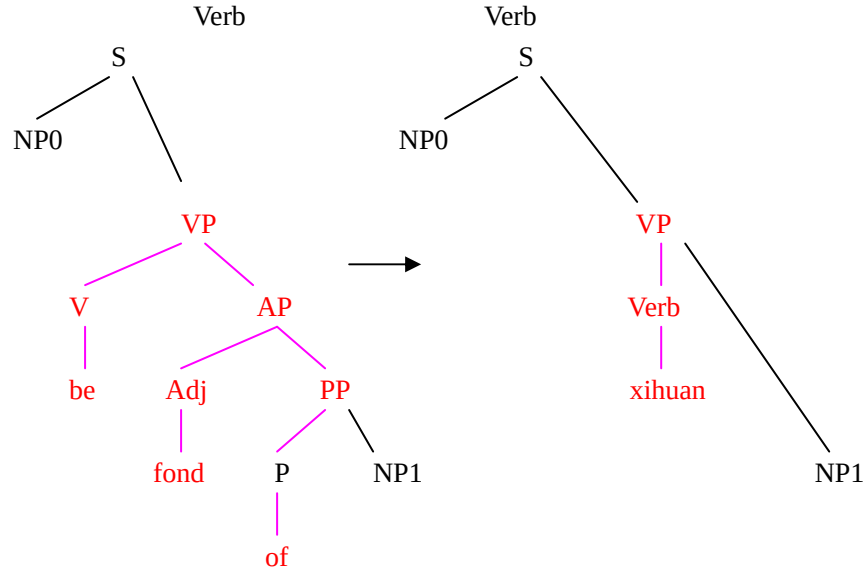


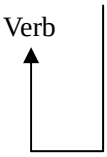

Fig. 2. Mapping SL tree to TL tree

The adjective “fond” in English is transferred to verb “xihuan [喜欢]” in Chinese.

E.g. In English-French MT,

English: “John is *very fond of* music”


French: “John *aime beaucoup* la musique” (“John loves very much the music”)



The English adverb “very” is associated with the predicate “fond of” (instead of with the main verb) whereas in French, the corresponding adverbial “beaucoup” is associated with the main verb “aimer”

1.7 Conflation Divergence: Conflation is the incorporation of necessary participants (or arguments) of a given action. A conflation divergence arises when there is a difference in incorporation properties between the two languages.

E.g. In English-Chinese MT,

English: “I *stabbed* John”

Chinese: 我用刀刺伤了约翰 (Wo *yong dao cishangle* Yuehan.) means “I use knife to wound John”

The main verb “stab” of English is translated as ”yong dao cishang [用刀刺伤]” (means “using knife to wound”) in Chinese. This information is incorporated into the main verb “stab” in English.

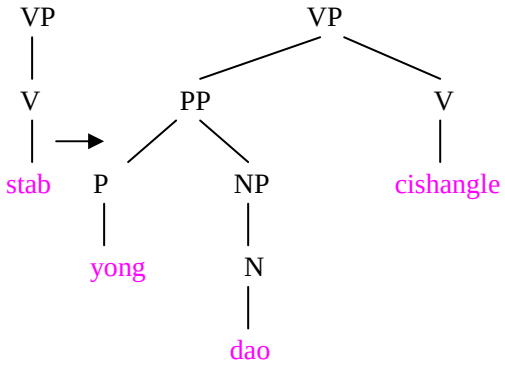



Fig. 3. conflation divergence

1.8 Syntactic Ambiguity

The syntactic ambiguity is a difficult parsing problem in SL analysis.

E.g. I saw the man on the hill *with the telescope*.

PP “with a telescope” can be analyzed as the modifier of verb “saw”, it can also be analyzed as the modifier of noun “hill”. It is the “PP-attachment” in the English parsing.

If we cannot decide the PP-attachment relation in SL analysis, it will become a difficult problem in transfer form SL to TL because the MT program does not know how to transfer the PP.

1.9 Complex Semantic ambiguity

In SL analysis, the homography and metonymy are the complex semantic ambiguity.

- Homography:

E.g. in “The box was in the *pen*”, the “pen” is a writing instrument or some sort of enclosed space (i.e. a play pen or pig pen), it is a homograph.

Bar-Hillel said: FAHQMT is unattainable with the technology of the time that could not guarantee word sense choice.

His now-famous example:

“Little John was looking for his toy box. Finally, he found it. The *box was in the pen*. John was very happy.”

The word “pen” in the emphasized sentence has at least two meanings – a writing pen and a playpen. Bar-Hillel conclusion was that “no existing or imaginable program will enable an electronic computer to determine that the word *pen* in the given sentence within the given context has the second of the above meaning.”

However, Bar-Hillel suggests: any means for resolving difficulties in language analysis and generation are appropriate, even if their theoretical nature is not yet understood by science.

If the Word Sense Disambiguation cannot be resolved in SL analysis, it will be very difficult to find the translation equivalence in TL.

- Metonymy: E.g. “While driving, John swerved and hit a tree.” MT system must determine that it is *John* who is driving but *John’s car* that hit the tree (metonymy resolution). The metonymy in SL will increase the difficulty in transfer. But if the TL correlate is always the same as SL, we can keep this metonymy in the transfer and generation.

## 1.10 Contextual ambiguity

The contextual ambiguity is also a problem in the SL analysis.

E.g. “The computer outputs the data; *it* is fast.”

“The computer outputs the data; *it* is stored in ASCII.”

In the context of a computer manual, determining the appropriate antecedent for the word “it” can be solved by distinguishing between storable objects and non-storable objects (storable [yes or no]) and between objects with a speed attribute and those without (speed fast/slow).

The anaphor ambiguity in SL analysis is important for Natural Language Understanding (NLU). However, for MT, we may keep the anaphor ambiguity in transfer and generation, if it doesn’t give a remarkable influence for the understanding of TL. .

## 1.11 Complex contextual ambiguity

The complex contextual ambiguity is a kind of PP-attachment in SL analysis.

E.g. “John hit the dog *with a stick*”.

This ambiguity could be resolved by remembering from the earlier text that John was carrying a stick to protect himself or there were several dogs, one of which had a stick. It is a kind of PP-attachment of English. Since the translation divergence of Chinese from English, it is a difficult problem in transfer from English to Chinese.

## 1.12 What sort of Computation is MT?

E.g “The soldiers fired at the women and I saw several of *them* fall.”

In this English sentence, “them” can be construed to refer either to “women” or to “soldiers” When this sentence was translated to Chinese, French, Spanish or Russian, MT system must determine the gender of “them”. The correct gender can be only determined if we know whether the soldiers or women fell.

In this case, the depth of analysis of MT must be very deep. MT is a sort of computation that needs deep analysis. However, it is important to realize that there are genuinely different theories of how to do MT, and only practical experiment will show which ones are right

.

## 2. Co-occurrence cluster based approach for Selection of Translation Equivalence (STE)

### 2.1 Hanks approach

Based on the corpus analysis, P Hanks found that the words including in the context of homograph can be classified to different co-occurrence clusters in accordance with the different meanings of the target word. E.g. the co-occurrence words including in the context of homograph "*bank*" can be classified to two co-occurrence clusters:

Cluster A: money, notes, loan, account, investment, clerk, official, robbery, vault, working in a, First national, of England.

Cluster B: river, swim, boat, east, west, south, on top of.

The co-occurrence words in cluster A correlated with the meaning "financial institution" of "bank". E.g.

"A *bank* can hold the *investments* in an *account* in the client's name." – bank1

The co-occurrence words in cluster B correlated with the meaning "sloping mound". E.g.

"As the agriculture development on the *east bank*, the *river* will shrink even more." – bank2

### 2.2 Teubert approach

Wolfgang Teubert proposed the similar approach to deal with WSD (Word Sense Disambiguation). "Mole" is a homograph with three different meanings.

Mole1: a stone wall of great strength built out into the see from the land as a defence against the force of the waves, or to act as a road.

Mole2: a small, dark brown, slightly raised mark on a person's skin, usually there since birth.

Mole3: a type of small insect-eating animal with very small eyes and soft dark fur, which digs holes and passages underground and makes its home in them.

The co-occurrence clusters of "mole" are:

Cluster A: harbour, water, boat

Cluster B: skin, spot, dark.

Cluster C: garden, dig, underground.

The co-occurrence words in cluster A correlated with the meaning of mole1; the co-occurrence words in cluster B correlated with the meaning of mole2; the co-occurrence words in cluster C correlated with the meaning of mole3.

If we have to translate the sentence "The doctor had a look at the *dark mole* on Anna's *skin*",

In the context of "mole" we find, among other words, the co-occurrence words "dark" and "skin". These context words will be checked against the context profiles for "mole" listed in the co-occurrence cluster. The closest match will be the best choice – mole2.

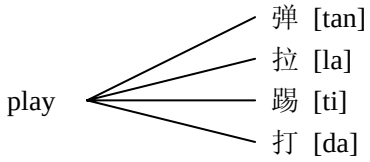
The co-occurrence cluster approach can be used also in the transfer to deal with the translation divergence, and to select the adequate translation equivalence in TL.

### 2.3 Co-occurrence cluster in MT

In English-Chinese MT, we can classify the context of the target word in SL as different clusters. The different ambiguous translated words in TL of this target word will be listed together with

their context profiles in different clusters. The context profiles will help us to determine the one adequate sense when we have to translate this target word to TL. For example,

They *play* piano – “他们弹钢琴”(Tamen *tan* gangqin).
They *play* violin. – “他们拉小提琴”(Tamen *la* xiaotiqin).
They *play* football. – “他们踢足球”(Tamen *ti* zuqiu).
They *play* basketball – “他们打篮球”(Tamen *da* lanqiu)




We can classify the words in context of “play” to four clusters:
Cluster A: piano, organ, pipe organ (musical instrument played by pressing narrow bars).
Cluster B: violin, cello, harp, guitar (stringed musical instrument)
Cluster C: football.
Cluster D: basketball, volleyball, baseball.
In the transfer, the selection translation equivalence of “play” depends on the type of clusters.

| Type of cluster | | Translation equivalence |
|---|---|---|
| Cluster A | – | 弹 [tan] |
| Cluster B | – | 拉 [la] |
| Cluster C | – | 踢 [ti] |
| Cluster D | – | 打 [da] |

Of course, this kind of research work must be based on the aligned parallel corpus.

## 2.4 Feature vector

In the corpus processing of WSD, we can use the feature vector to represent the co-occurrence cluster. A simple feature vector consisting of numeric values can easily encode the most frequently used linguistic information.

E.g.

If we have the co-occurrence cluster for WSD of “bank”:
Cluster A: money, notes, loan, account, investment, clerk, official, robbery, vault, working in a, First national, of England.
Cluster B: river, swim, boat, east, west, south, on top of.
For sentence “A *bank* can hold the *investments* in an *account* in the client’s name.”, the feature vector of cluster A is:
[0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0]
The feature vector of cluster B is:
[0, 0, 0, 0, 0, 0, 0]
So we can select the sense that represented by cluster A as the adequate sense of bank in this sentence. -- “financial institution”
For sentence “As the agriculture development on the *east bank*, the *river* will shrink even more.”, the feature vector of cluster A is:
[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
The feature vector of cluster B is:

[1, 0, 0, 1, 0, 0, 0]
So we can select the sense that represented by cluster B as the adequate sense of bank in this sentence: “sloping mound”.

Similarly, in sentence “The doctor had a look at the *dark* mole on Anna’s *skin*”, the feature vector of cluster A, B and C for WSD of “mole” are respective as following:
Cluster A: [0, 0, 0]
Cluster B: [1, 0, 1]
Cluster C: [0, 0, 0]
Therefore, we can select the sense that represented by cluster B as the adequate sense of “mole’ in this sentence: “a small, dark brown, slightly raised mark on a person’s skin, usually there since birth.”

2.5 TranslationBase in CCL
Now a TranslationBase of Machine Translation is developing in CCL (Centre for Corpus Linguistics) of Birmingham University headed by Prof. Dr. Wolfgang Teubert. This project shall use the cluster based approach to select the translation equivalence in MT. This project will facilitate, improve and speed up human translation, it will make possible machine translation of real, unrestricted natural language, and it will be used for a range of further Chinese-English language technology, including collaborative multimedia document authoring and publishing, the creation of parallel language versions, cross-lingual terminology support, summarization, quality assurance of translations, and also for language learning**.** It is a remarkable development in current MT research.
The co-occurrence cluster for the selection of translation equivalence of English verb “play” in Chinese sentence is relatively easy, we can adopt the feature vector to deal with this problem.

## 3. Some suggestions

We may use the co-occurrence cluster based approach to select the Translation Equivalence in MT. However, when the translation divergence between SL and TL is not so large, we need not to analyze the SL into too deep level or understand the meaning of SL. So we propose following suggestions in MT.

3.1 Can one avoid treatment of meaning?

- If the structure of SL and TL is similar, we can avoid treat the meaning of SL. For example, in the PP-attachment, why we waste time to detect and represent the meaning of the input string when the target language correlate is always the same?
- Ben Ari said: “It must kept in mind that the translation process does not necessarily require full understanding of the text. Many ambiguities may be preserved during translation, and thus should be presented to the user (human translator) for resolution.” (*Translation ambiguity rephrased*, 1988)
- Isabelle and Bourbeau said: “Sometimes, it is possible to ignore certain ambiguities, in the hope that the same ambiguities will carry over in translation. This is particularly true in system like TAUM-aviation that deals with only one pair of closely related languages within a severely restricted sub-domain. The difficult problem of prepositional phrase attachment, for example, is frequently bypassed in this way.

Generally speaking, however, analysis is aimed at producing an unambiguous intermediate representation." (*TAUM-AVIATION: its technical features and some experimental results*, 1985)

3.2 Varying the acceptability threshold

If MT text is not intended for publication, then the quality standards for MT can be effectively lowered without tangible harm.

There are different applications of MT:

- Dissemination;
- Assimilation;
- Interchange : immediate translation, reading the text of Internet.
- Information access.

According to the user of MT, MT can be classified as following types:

- For the watcher: MT-W;
- For the reviser: MT-R;
- For the translator: MT-T
- For the author: MT-A

Different users have different threshold in MT acceptability.

3.3 Partial automation in MT

In some cases, we may use the partial automation technique as follows:

- Post-editing (SYSTRAN system)
- Interactive computer environment – translator's workstation.
- Translation memory (TRADOS system).

3.4 Restricting the Ambiguity of source text

We can also restrict the ambiguity of text in SL.

- Choosing a sufficiently narrow subject domain (like as TAUM-METEO). TAUM-METEO was developed at the University of Montreal and delivered to the Canadian Weather Service for everyday routine translations of weather reports from English into French. The system operates very successfully, practically without human intervention. Its vocabulary consists of about 1500 items, about half of which are place names. There is very little lexical ambiguity in the system because words are expected to be used in only one of their senses – namely, the one that belongs to the sub-world of weather phenomena. Of course, finding well-delineated, self-sufficient, and useful sub-language is a very difficult task.
- Human pre-editor modifies the text to include only the words and constructions that the MT system is able to process automatically.

3.5 MT combines with other kinds of text processing:

The good information retrieval may result from less than high-quality translation results. The information retrieval can be done without a "blanket" coverage of the text, without succeeding in analyzing each and every input word.

# Reference


1. Sergei Nirenburg, Yorick Wilks, Machine Translation, <Advances in computers>, Volume 52, 2000, Academic Press.
2. Bonnie J. Dorr, Pamela W. Jordan, John W. Benoit, A survey of Current Paradigms in Machine Translation, <Advances in Computers>, Volume 49, 1999, Academic Press.
3. Feng Zhiwei, Fundamental of Computational Linguistics, 2001, Beijing, Commercial Press.
4. Ben Ari et al., Translation ambiguity rephrased, Proceedings of the 2nd International Conference on Theoretical and methodological Issues in Machine Translation of Natural Languages, 1988.
5. Isabelle and Bourbeau, TAUM-AVIATION: its technical features and some experimental results, Computational Linguistics, 11(1), 1985.