Complex Features in Description of Chinese Language

Feng Zhiwei
Institute of Applied Linguistics
Chinese Academy of Social Sciences
51 Chaoyangmen Nanxiaojie
100010 Beijing, China

Abstract

In this paper, the similarity of "multi-value label function" and "complex features" is discussed. The author especially emphasizes the necessity of complex features for description of Chinese language.

§1 Multiple-Value Label Function

The phrase structure grammar (PSG) was used extensively in the parsing of natural language. A PSG can be expressed by a tree graph where every node has a correspondent label. The relationship between the node x and its label y can be described by a mono-value label function L:

$$L(x) = y$$

For every value of node x, there is only one corresponding value of label y.

In 1981, we designed a multilingual automatic translation system FAJRA (from Chinese to French, English, Japanese, Russian, German). In 1985, we designed two automatic translation systems GCAT (from German to Chinese) and FCAT (from French to Chinese). In system FAJRA, we must do automatic analysis of Chinese, in system GCAT and FCAT, we must do automatic generation of Chinese language. We found that the linguistic features of Chinese expressed by the mono-value label function of PSG is rather limited. In the automatic analysis of Chinese language, PSG can not treat properly the ambiguity, the result of analysis often brings about a lot of ambiguous structures. In the automatic generation of Chinese language, the generative power of PSG is so strong that a lot of ungrammatical sentences are generated. It is the chief drawback of PSG. In order to overcome this drawback of PSG, in system FAJRA, we proposed a multiple-value label function to replace the mono-value label function, and in systems GCAT and FCAT, we further improved this approach.

A multiple-value label function can be described as below:

$$L(x) = \{y_1, y_2, \ldots, y_n\}$$

whereby a label x of tree can correspond to several labels $\{y_1, y_2, \ldots, y_n\}$. By the means of this function, the generative power of PSG was restricted, the number of ambiguous structures were reduced, and the drawback of PSG was efficiently overcome.

Beginning with the augmented transition network (ATN) concept and inspired by J.Bresnan's work on lexically oriented non-transformational linguistics, the lexical functional grammar (LFG) framework of J. Bresnan and R. Kaplan was evolved. Simultanously, M. Kay devised the functional unification grammar (FUG), G. Gazdar, E. Klein and G. Pullum proposed the generalized phrase structure grammar (GPSG). Implementation of GPSG

at Hewlett－Packard led C. Pollard and his collegues to design the head－driven phrase structure grammar (HPSG) as a successor to GPSG. In all these formalisms of grammars, the complex features are widely used to replace the simple feature of PSG and it is an improvement of PSG. Therefore, The concept of complex features is very important for present development of computational linguistics.

In the systems FAJRA, GCAT and FCAT, the values of labels must be the features of language, so the multiple－value labels must be also the complex features of language. In fact, the concept of multiple－value labels and the concept of complex features is very similar. Historically, all these concepts are the results of separate researches in computational linguistics for improvement of PSG. Thus we can take our multiple－value labels as complex features.

The famous linguist De Saussure (1857－1913) had pointed out in his <Course in General Linguistics> : "Language in a manner of speaking, is a type of algebra consisting solely of complex terms" (p122, English version, 1959). He takes the flexion of "Nacht : Nächte" in German as the example. The relation "Nacht : Nächte" can be expressed by an algebraic formula a/b in which a and b are not simple terms but result from a set of relations －－ complex terms. The complex terms of Nacht are: noun, femmine gender, singular number, nominative case, its principal vowel is "a". The complex terms of Nächte are: noun, femmine gender, plural number, nominative case, its principal vowel is "ä", its ending is "e", the pronunciation of "ch" changes from /x/ to /ç/. De Saussure said: "But language being what it is, we shall find nothing simple in it regardless of our approach; every where and always there is the same complex equilibrium of terms that mutually condition each other" (P122,E. v., 1959). So called "complex terms" mentioned here by De Saussure is nothing but the "complex features" or the "multiple－value labels" in computational linguistics. After all, De Saussure is a scholar with a foresight. He has proved himself to be a pioneer of modern linguistics. The concept of "complex terms" of De Saussure has served as a source of inspiration for us to proposed the concept of "multiple－value labels". However, the property of Chinese language is more important than the concept of De Saussure in the developmemnt of our "multiple－value labels" (or "complex features"), because without the "multiple－value labels" (or "complex features") we can not adequately describe the Chinese language in the automatic translation.

§2 Necessity of Complex Feature for Description of Chinese Language

If there is a necessity of complex features in description of English, then this necessity is more obvious for description of Chinese.

The reasons are as following:

1. There is not one－to－one correspondence relation between the phrase types (or parts of speech) and their syntactic functions in the Chinese sentences.

Chinese is different from English. In English, the phrase types generally correspond to their syntactic functions. For example, we have NP + VP －－> S in English, whereby NP corresponds to subject, VP corresponds to predicate, and S is a sentence, becoming a construction"

subject + predicate". There is a one-to-one correspondence between the phrase types and their syntactic functions. In Chinese, the structure "NP + VP" can form a sentence, so we have also NP + VP --> S. E. g. in phrase "小王/咳嗽" (little Wang coughs), "小王" (little Wang) is a NP, "咳嗽" (to cough) is a VP, forming a construction "subject +predicate". However, in many cases, the NP doesn't correspond to the subject, the VP doesn't correspond to the predicate. For example, "程序/设计" (progrmming) in Chinese, "程序" (program) is a NP, "设计" (to design) is a VP, but this NP is a modifier, and this VP is the head of the structure "NP + VP". The structure "NP + VP" can not form a sentence, but form a new noun phrase NP1: NP + VP --> NP1. So this noun phrase becomes a construction "modifier + head". Similar example are: "语言/学习" (language learning), "物理/考试" (physics examination), etc. In these phrases, the phrase types "NP + VP" can not form a construction "subject + predicate", but form a construction "modifier + head". In this case, the "NP + VP" is a syntactically ambiguous structure, the simple features "NP + VP" can not distinguish the differences between the construction "subject + predicate" and the construction "modifier + head". We must use the complex features to describe these differences.

The structure "NP + VP" which forms a construction "subject + predicate" can be formularized as following complex feature set:

$$\left[\begin{bmatrix} K = NP \\ CAT = N \\ SF = SUBJ \end{bmatrix} + \begin{bmatrix} K = VP \\ CAT = V \\ SF = PRED \end{bmatrix}\right]$$

whereby K is the feature of the phrase type, NP and VP are the values of this feature; CAT is the feature of the parts of speech, N and V are the values of this feature; SF is the feature of syntactic function, SUBJ and PRED are the values of this feature. With the complex features, the structure "NP + VP" which forms a construction "modifier + head" can be formularized as following:

$$\left[\begin{bmatrix} K = NP \\ CAT = N \\ SF = MODF \end{bmatrix} + \begin{bmatrix} K = VP \\ CAT = V \\ SF = HEAD \end{bmatrix}\right]$$

whereby MODF and HEAD are the values of the feature SF.

Obviously, the structure "NP + VP" is ambiguous, there are two syntactically different constructions included in this structure, their phrase types are identical, but their syntactic functions are different. In order to adequately discribe the differences of them, we have to use the complex features in deed.

In English, we have VP + NP --> VP1, the VP corresponding to the predicate, the NP to the object, and VP1 is a new verb phrase, it is a construction "predicate + object". In Chinese, the structure "VP + NP" can form a new verb phrase, so we have also VP + NP --> VP1. For example, "讨论/问题" (to discuss a problem) in Chinese, "讨论" (to discuss) is a VP, "问题" (problem) is a NP, forming a new verb phrase "predicate + object".

However, in many cases, VP doesn't correspond to predicate, NP doesn't correspond to object. For example, "出租 / 汽车" (taxicab) in Chinese, "出租" (to hire) is a VP, "汽车" (automobile) is a NP, but this VP is a modifier and this NP is the head of the structure "VP + NP". This structure can not form a new verb phrase, it forms a new noun phrase, so we have: VP + NP ---> NP1. The syntactic function of this noun phrase is a construction "modifier + head". The similar examples are: "研究 / 方法" (approach for reseach), "学习 / 制度" (regulation for study), "开放 / 政策" (policy of openning the door). The phrase type structure "VP + NP" can not form a construction "predicate + object", but forms a construction "modifier + head". In this case, the structure "VP + NP" is syntactically ambiguous. The simple features "VP + NP" is not enough to distinguish the differences between the construction "predicate + object" and the construction "modifier + head". We must use the complex features to describe these differences.

The structure "VP + NP" which forms a construction "predicate + object" can be formularized as following complex feature set:

$$\left[\begin{bmatrix} K = VP \\ CAT = V \\ SF = PRED \end{bmatrix} + \begin{bmatrix} K = NP \\ CAT = N \\ SF = OBJE \end{bmatrix}\right]$$

whereby PRED and OBJE are the values of feature SF.

The structure "VP + NP" which forms a construction "modifier + head" can be formularized as following complex feature set:

$$\left[\begin{bmatrix} K = VP \\ CAT = V \\ SF = MODF \end{bmatrix} + \begin{bmatrix} K = NP \\ CAT = N \\ SF = HEAD \end{bmatrix}\right]$$

whereby MODF and HEAD are the values of feature SF.

Obviously we can not only use the simple feature to describe the structure "VP + NP", we have to use the complex features.

2. For the construction which have the same phrase type structure and the same syntactic function structure, its semantic relation may be different. So there is not simple one-to-one correspondence between the syntactic function and its semantic relation.

The values of semantic relation feature are the following: agent, patient, instrument, scope, aim, result, ... etc. In English, there is no much correspondences between the syntactic function of sentence element and its semantic relation. In Chinese, the correspondence is more complicated and sophisticated than in English.

In the construction "subject + predicate" with corresponding phrase type structure "NP + VP", the subject may be agent, but it may also be patient, or instrument, etc. For example, in the sentence "我/ 读了" (I have read), the subject "我" (I) is the agent, but in the sentence "书/ 读了" (the book has been read), the subject "书" (book) is the patient, and the verb "读了" (to read) doesn't change its form, it always takes the original form. In most European languages, if the subject is the agent, then the verb must take an active form, and if

the subject is patient, then the verb must take the passive form. However, In Chinese, the verb always keeps the same form no matter whether the subject is an agent or a patient. In an overwhelming majority of cases, the passive form of verb is seldom or never used in Chinese. By this reason, in the automatic information processing of Chinese language, it is not enough to only use the features of phrase types and syntactic functions, we must use the features of semantic relations, thus the features for description of Chinese language will become more complex.

In the structure "NP + VP", if the syntactic function of NP is subject, and the semantic relation of NP is agent, then the complex features of this structure can be formularized as following:

$$\left[\begin{array}{l}\left[\begin{array}{l}K = NP\\CAT = N\\SF = SUBJ\\SM = AGENT\end{array}\right] + \left[\begin{array}{l}K = VP\\CAT = V\\SF = PRED\end{array}\right]\end{array}\right]$$

whereby SM is the feature of semantic relation, AGENT is the values of feature SM.

In the structure "NP + VP", if the syntactic function of NP is subject and the semantic relation of NP is patient, then the complex features of this structure can be formularized as following:

$$\left[\begin{array}{l}\left[\begin{array}{l}K = NP\\CAT = N\\SF = SUBJ\\SM = PATIENT\end{array}\right] + \left[\begin{array}{l}K = VP\\CAT = V\\SF = PRED\end{array}\right]\end{array}\right]$$

whereby PATIENT is the value of feature SM.

In the strcture "VP+NP" with corresponding syntactic construction" predicate+object", the object may be a patient, but it may also be an agent, or an instrument, or a scope, or an aim, or a result, . . . , etc. In the sentence "擦/窗子" (to wipe the window), "擦" (to wipe) is the predicate, "窗子" (window) is the object, and its semantic feature is the patient. The complex features of this sentence can be formularized as following:

$$\left[\begin{array}{l}\left[\begin{array}{l}K = VP\\CAT = V\\SF = PRED\end{array}\right] + \left[\begin{array}{l}K = NP\\CAT = N\\SF = OBJE\\SM = PATIENT\end{array}\right]\end{array}\right]$$

whereby PATIENT is the value of feature SM.

But we have also the following sentences where the semantic relation of object is not the patient. There are many very interesting phenomena in Chinese language: In the sentence "死了/父亲" (the father died), the object "父亲" (father) is the agent of the verb "死了" (to die). This structure can be formularized as following comlex feature set:

$$\left[\begin{bmatrix}K=VP\\CAT=V\\SF=PRED\end{bmatrix}+\begin{bmatrix}K=NP\\CAT=N\\SF=OBJE\\SM=AGENT\end{bmatrix}\right]$$

whereby AGENT is the value of feature SM.

In the sentence "吃/大碗" (to eat with a big bowl), the object"大碗" (big bowl) is the instrument of the verb "吃" (to eat). This structure can be formularized as following complex feature set:

$$\left[\begin{bmatrix}K=VP\\CAT=V\\SF=PRED\end{bmatrix}+\begin{bmatrix}K=NP\\CAT=N\\SF=OBJE\\SM=INST\end{bmatrix}\right]$$

whereby INST is the value of feature SM.

In the sentence "考/数学" (to examine in mathematics), the object"数学" (mathematics) is the scope of the verb "考" (to examine). This structure can be formularized as following complex feature set:

$$\left[\begin{bmatrix}K=VP\\CAT=V\\SF=PRED\end{bmatrix}+\begin{bmatrix}K=NP\\CAT=N\\SF=OBJE\\SM=SCOPE\end{bmatrix}\right]$$

whereby SCOPE is the value of feature SM.

In the sentence "考/研究生" (to examine in order to become a graduate student), the object "研究生" (graduate student) is the aim of the verb "考" (to examine). This structure can be formularized as follwoing complex feature set:

$$\left[\begin{bmatrix}K=VP\\CAT=V\\SF=PRED\end{bmatrix}+\begin{bmatrix}K=NP\\CAT=N\\SF=OBJE\\SM=AIM\end{bmatrix}\right]$$

whereby AIM is the value of feature SM.

In the sentence "考/满分" (to pass an examination and get an excellent marks). The object "满分" (excellent marks) is the result of the verb "考" (to examine). This structure can be formularized as following complex feature set:

$$\left[\begin{bmatrix}K=VP\\CAT=V\\SF=PRED\end{bmatrix}+\begin{bmatrix}K=NP\\CAT=N\\SF=OBJE\\SM=RESULT\end{bmatrix}\right]$$

whereby RESULT is the value of feature SM.

Thus we can see very clearly that only with the complex features the differences of these sentences can be revealed sufficiently.

3. The grammatical and semantic features of words play an important role to put forward the rules of parsing. These features can be used as the conditions in the rules, and

they must be included into the system of complex features. Thus the complex features is the basis to set up the rules for parsing of Chinese language.

In the structure VP + NP, if the grammatical feature of VP is intransitive verb, then the syntactic function of VP must be modifier, and the syntactic function of NP must be head. The syntactic function of this structure VP + NP will be "modifier + head". Thus we can have a rule which is described with complex features as following:

$$\begin{bmatrix} K = VP \\ CAT = V \\ TRANS = IV \end{bmatrix} + \begin{bmatrix} K = NP \\ CAT = N \end{bmatrix} \longrightarrow \begin{bmatrix} \begin{bmatrix} K = VP \\ CAT = V \\ TRANS = IV \\ SF = MODF \end{bmatrix} + \begin{bmatrix} K = NP \\ CAT = N \\ SF = HEAD \end{bmatrix} \end{bmatrix}$$

whereby TRANS represents the feature of transitivity of verb, IV represents intransitive verb, it is a value of feature TRANS.

This rule means: In the structure VP + NP, if the value of TRANS of VP is IV, then the syntactical function of VP can be given the value MODF, and the syntactical function of NP can be given the value HEAD.

The semantic features of words can also be used to put forward the rules of parsing. In the structure VP + NP, if VP is transitive verb, then we have to use the semantic features of NP to decide the value of syntactic function of this structure. Generally speaking, if VP is transitive verb, the semantic feature of NP is "abstract thing", or "title of a technical or professional post", then the syntactical function of VP is modifier, and the syntactical function of NP is head. For example, in the phrase "训练/目的" (purpose of training), the VP "训练" (to train) is a transitive verb, and the semantic feature of NP "目的"(purpose) is "abstract thing", we can decide that the syntactical function of "训练" (to train) is modifier, and the syntactical function of "目的" (purpose) is head. In the phrase "进修/教师" (high training teacher), the VP "进修" (to engage in advanced studies) is a transitive verb, the NP "教师" (teacher) is a title of professional post, thus we can decide that the syntactical function of VP "进修" (to engage in advanced studies) is modifier, the syntactical function of NP "教师" (teacher) is head.

Therefore we can have two follwoing rules for parsing:

(1) $$\begin{bmatrix} K = VP \\ CAT = V \\ TRANS = TV \end{bmatrix} + \begin{bmatrix} K = NP \\ CAT = N \\ SEM = ABS \end{bmatrix} \longrightarrow \begin{bmatrix} \begin{bmatrix} K = VP \\ CAT = V \\ TRANS = TV \\ SF = MODF \end{bmatrix} + \begin{bmatrix} K = NP \\ CAT = N \\ SEM = ABS \\ SF = HEAD \end{bmatrix} \end{bmatrix}$$

(2) $$\begin{bmatrix} K = VP \\ CAT = V \\ TRANS = TV \end{bmatrix} + \begin{bmatrix} K = NP \\ CAT = N \\ SEM = PRF \end{bmatrix} \longrightarrow \begin{bmatrix} \begin{bmatrix} K = VP \\ CAT = V \\ TRANS = TV \\ SF = MODF \end{bmatrix} + \begin{bmatrix} K = NP \\ CAT = N \\ SEM = PRF \\ SF = HEAD \end{bmatrix} \end{bmatrix}$$

whereby TV represents transitive verb, it is a value of TRANS, ABS represents abstract thing, it is a value of SEM, PRF represents title of a technical or professional post, it is another value

of SEM. On the basis of these complex features, two new values of feature SF can be deduced:

SF = MODF and SF = HEAD.

With the complex features, we got good results in machine translation systems FAJRA, GCAT and FCAT (see Annex).

BIBLIOGRAPHY

[1] Feng Zhiwei, multiple-branched and multiple-labeled tree analysis of Chinese sentence, <Journal of Artificial Intelligence>, No 2, 1983.

[2] De Saussure, Course in General Linguistics, English version, 1959.

[3] M. Kay, Parsing in functional unification grammar, in <Natural Language Parsing : Psychological , Computational and Theoretical Perspectives> , 1985.

Annex

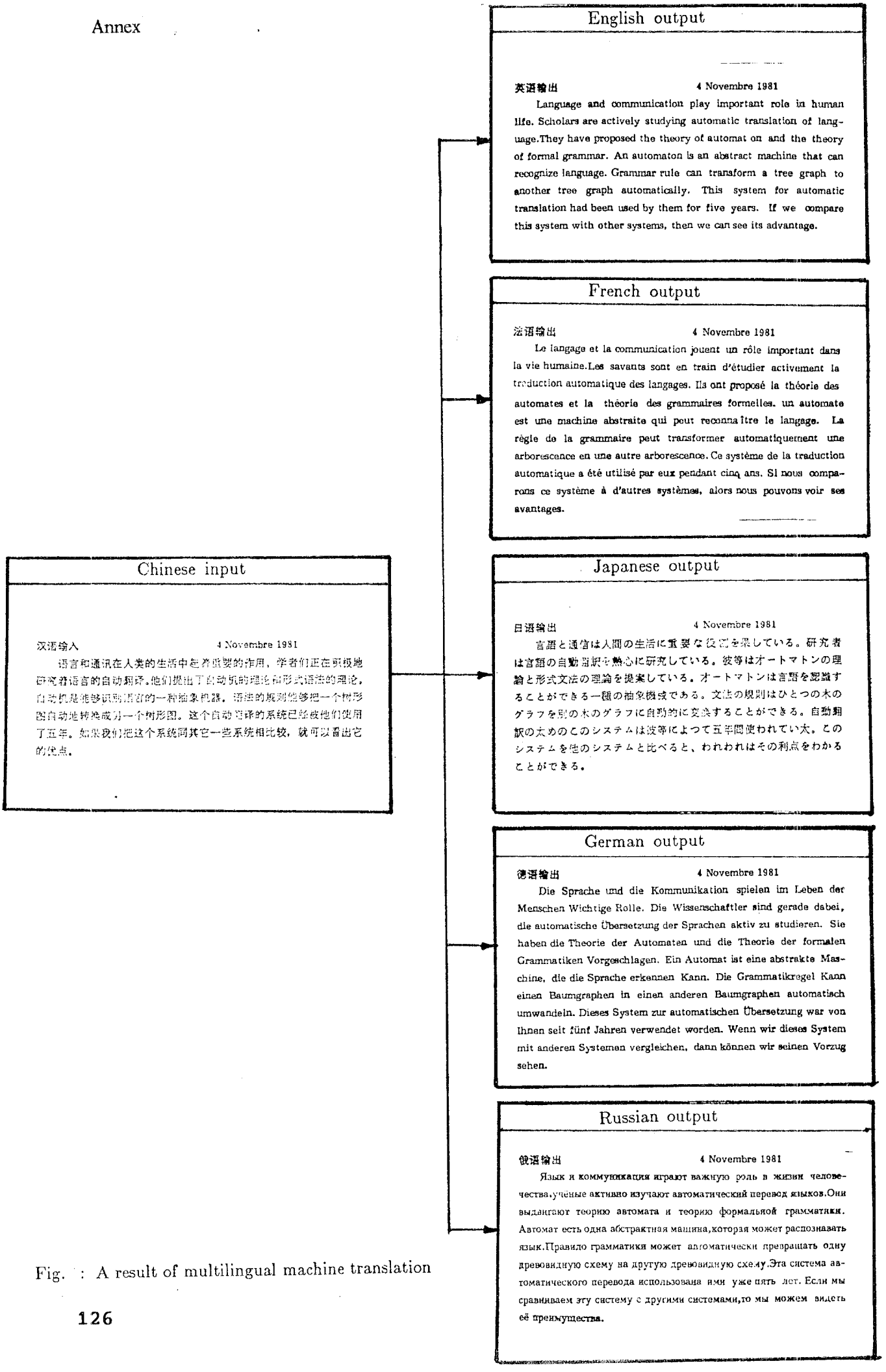

Fig. : A result of multilingual machine translation