Slovenská akadémia vied

Združenie slovenských jazykovedcov pri SAV

Slovak Academy of Sciences

Association of Slovak Linguists attached to the Slovak Academy of Sciences

Volume IV

Recueil linguistique de Bratislava

Proceedings of the Symposium on Algebraic Linguistics held 10—12 February 1970 at Smolenice

Vydavateľstvo Slovenskej akadémie vied
Publishing House of the Slovak Academy of Sciences
Bratislava 1973

Recueil linguistique de Bratislava is regularly published by the Association of Slovak Linguists attached to the Slovak Academy of Sciences. In this volume (IV) the papers are published which was presented to the International Symposium on Algebraic Linguistics, held 10—12 February 1970 at Smolenice. The Symposium have been prepared by the Linguistic Institute Ľudovít Štúr's of the Slovak Academy of Sciences.

The main topic of the Symposium was the problem of linguistic models. The papers have been ordered into some fields: the general problems of modelling, the algebraic views in model theory, the application of algebraic models to the linguistic data, the grammatical and semantic problems solved by the aid of models.

Contents

Programme of the Symposium on Algebraic Linguistics

Tuesday February 10th 9.00—12.30
Chairman: Mr. László Kalmár
Assistants: Mr. Jozef Ružička, Mr. Ján Horecký
Secretary: Mrs. Klára Buzássyová

Jozef Ružička, Bratislava
Einführungsrede
Ján Horecký, Bratislava
The Role of the Models in Linguistic Studies
Aleksander Ludskanov, Sofia
Quelques remarques sur l'emploi des termes "modèle" et "formalisation" et sur leurs relations dans les travaux linguistiques contemporains
Yves Gentillhomme, Besançon
La proportion linguistique et la notion de groupe
Marie Těšitelová, Praha
Zum Modellieren in der Linguistik vom quantitativen Standpunkt
Ludmila Uhlířová, Praha
On the quantitative description of systemic relations in language
Jan Průcha, Praha
Problems of Generative Model in Psycholinguistics

Tuesday February 10th 15.00—18.30
Chairman: Mr. Bernard Vauquois
Assistants: Mr. László Dezső, Mr. Petr Sgall
Secretary: Mr. Ján Bosák

László Kalmár, Szeged
On a Measure of Divergence of a Context-free Language from Finite State Languages
Miroslav Novotný, Brno
On Some Connections between the Generative and Analytic Models of Languages

Karel Čulík, Praha
On Conditional Context-free Grammars for Programming and Natural Languages
Adrian Birbanescu, Bucureşti
Kalgren's Decision Grammars
Eva Hajičová, Jarmila Panevová, Petr Sgall, Praha
Meaning of Tense and its Recursive Properties

Wednesday February 11th 9.00—12.30
Chairman: Mr. Rudolf Růžička
Assistants: Mr. Miroslav Novotný, Mr. Viliam Schwanzer
Secretary: Mr. Ján Horecký

Solomon Marcus, Bucureşti
Un modèle mathématique integral de l'oeuvre dramatique
Liana Schwartz, Bucureşti
Étude algébrique comparative de la structure syntaxique et sémantique des variants d'un texte poétique
Alexandru Cărăuşu, Iaşi
A Note on Ambiguity of Context-sensitive Languages
Mihai Dinu, Bucureşti
Un modèle marcovien de l'influence à distance dans les langues naturelles

Wednesday February 11th 15.00—18.30
Chairman: Mr. Solomon Marcus
Assistants: Mr. Gunnar Jacobson, Mr. Karel Čulík
Secretary: Mrs. Klára Buzássyová

Otto Sechser, Praha
Generative Grammars and Document Retrieval Languages
Eduard Fiodorovič Skorochoďko, Kyjev
Modelirovanije jazyka v sviazi s zadačami informacionnogo poiska
Victoria Hopârteanu, Ileana Lascu, Dan Mârza, Maria Ţenchea, Cluj
Problèmes de typologie verbale à l'aide de la théorie des graphes
Emese Kis, Cluj
La structure algébrique des adverbes des langues romanes
Jürgen Kunze, Walter Priess, Berlin
Wortformenklassensysteme und ihre Optimierung
Gerda Klimonow, Berlin
Die grammatischen Konfigurationen im Modell der Abhängigkeitsgrammatik
Constantin V. Craciun, Bucureşti
Sur quelques problèmes d'analyse algébrique contextuelle

Thursday February 12th 8.30—11.30
Chairman: Mr. Aleksander Ludskanov
Assistants: Mrs. Emese Kis, Mr. Ján Horecký
Secretary: Mr. Ján Bosák

Stephan Ylan Solomon, Bucureşti
The n-derivative of a Partition
Gabriel Orman, Braşov
Quelques résultats concernant les ensembles homologues
Klára Buzássyová, Bratislava
On the Definition of the Word-forming Paradigm
Ján Oravec, Bratislava
Das System der slowakischen Präpositionen
Zdeněk F. Oliverius, Praha
A Model of Morphemic Description of Russian Words
László Dezső, Budapest
Topic—comment in Child Language and in Diachronic Typology
Bernard Vauquois, Grenoble
Réalisation de modèles d'analyse et de synthése de langues naturelles

Sehr verehrter Herr Vorsitzender, verehrte Damen und Herren!

1. Vor dem Anfang des wissenschaftlichen Programmes unseres Symposions möchte ich Sie, verehrte Damen und Herren, herzlich willkomen heißen, und zwar zuerst im Namen des Präsidenten der Slowakischen Akademie der Wissenschaften und dann im Namen der veranstaltenden Institution — des Sprachwissenschaftlichen Institutes der Slowakischen Akademie der Wissenschaften.

Herr Štefan Schwarz — der Präsident der Slowakischen Akademie der Wissenschaften — ist Professor der Mathematik an der Technischen Hochschule in Bratislava. Er hat viel Verständnis auch für unsere Probleme: er wollte sogar persönlich an unserer Tagung teilnehmen — leider, es gibt verschiedene wichtige Dinge anderer Art, die er als Präsident der Akademie zu erledigen hat. Darum müssen wir seine Abwesenheit entschuldigen.

Professor Štefan Schwarz ist der Meinung, daß auch die Linguistik nur durch Anwendung gewisser mathematischen Methoden einen beachtenswerten Grad von Exaktheit erreichen kann. Gott sei Dank — es gibt nicht nur *andere solche* Mathematiker, sondern auch *andere* Mathematiker auch bei uns, welche die Möglichkeiten der mathematischen Methoden jedenfalls hochschätzen, aber doch auch die Grenzen dieser Möglichkeiten empirisch betasten. Diese sind für die Entwicklung der Linguistik im Allgemeinen von größerer Bedeutung, denn sie helfen uns nicht nur die Methodologie unserer Arbeit weiter zu entwickeln, sondern auch die methodologische Buntheit unserer Wissenschaft aufzubewahren. Wir haben nämlich noch eine Menge von Problemen und von konkreten Aufgaben zu lösen, die nur mit anderen — älteren, aber gut erprobten — Methoden zu bewältigen sind.

2. Diese Tagung wird vom Sprachwissenschaftlichen Institut der Slowakischen Akademie der Wissenschaften veranstaltet. Darum wäre es vielleicht geeignet und besonders für unsere Gäste — die zum ersten Mal bei uns zu Gast sind — wahrscheinlich auch vom Nutzen einige Informationen über diese Arbeitsstelle anzuführen.

Unser Institut wurde im Jahre 1943 gegründet: es gehört also zu den ältesten Instituten der Slowakischen Akademie der Wissenschaften. Es entwickelte sich aber ziemlich langsam als ein komplexes Forschungsinstitut, das sich vorwiegend mit der Problematik der slowakischen Sprache beschäftigte. Im Vordergrund seiner Aufgaben

steht auch heutzutage die Problematik der slowakischen Schriftsprache, wodurch auch die Hauptlinien unserer theoretischen Bemühungen gegeben sind.

Das Sprachwissenschaftliche Institut ist auch vom Standpunkt der Struktur der Slowakischen Akademie und ihrer Institute gesehen ein mittelgroßes Institut: es hat insgesamt 50 Angestellte. Die wissenschaftliche Basis des Institutes ist in vier Abteilungen aufgeteilt.

Die Abteilung für die slowakische Schriftsprache ist die größte: hier werden einige lexikologischen und grammatischen Aufgaben bearbeitet und auch die Probleme der Sprachkultur theoretisch und auch applikativ gelöst. Die zweite Abteilung des Institutes befasst sich mit allen Problemen der slowakischen Mundarten. Und die dritte Abteilung hat die Entwicklung der slowakischen Sprache zu erforschen.

In diesen drei Abteilungen wurden bisher einige fundamentale Fragen der slowakischen Sprache tüchtig bearbeitet und auch große Werke davon veröffentlicht. Ich führe nur das wichtigste an: das sechsbändige Wörterbuch — *Slovník slovenského jazyka* (1959—1968), die große Morphologie — *Morfológia slovenského jazyka* (1966) und den monumentalen Sprachatlas — *Atlas slovenského jazyka* $I_{1,2}$ (1968—69). Alle diese Werke wurden (ebenso wie viele Monographien und Sammelbände aus unserem Institut) auch von ausländischen Slawisten mit Freude und Lob aufgenommen.

Bei dieser Gelegenheit ist aber eben die vierte Abteilung unseres Institutes vom größten Interesse — die Abteilung für mathematische Linguistik. Das ist zwar die jüngste und auch die kleinste Abteilung unseres Institutes, die jedoch unter Leitung von Professor Ján Horecký schon bisher gute Erfolge erreicht hat. In dieser Abteilung konzentrieren sich unsere Bemühungen auf theoretischem Gebiet: Prof. Ján Horecký ist nämlich der führende Methodologe der slowakischen Linguistik überhaupt, das bedeutet nicht nur im akademischen Institut, sondern auch an beiden slowakischen Universitäten (in Bratislava und Prešov). Das beweist auch sein Scriptum *Einführung in die mathematische Linguistik*.[1]

In der Abteilung für mathematische Linguistik wird in diesen Jahren besonders die Frequenz der Morpheme und grammatischer Formen erforscht: diese Arbeitsgruppe interessiert sich also um einige Probleme der statistischen Linguistik. Freilich auch einige Probleme der algebraischen Linguistik wurden schon mit erfreulichem Erfolg erörtert: hier sollte ich mehrere Aufsätze von verschiedenen Autoren anführen, z. B. aus der Phonologie der slowakischen Schriftsprache.[2]

Diese kleine Arbeitsgruppe will auch Kontakte mit anderen ähnlichen Arbeitsgruppen und Arbeitsstellen aufnehmen: in diesem Sinne ist auch das heutige Symposion zu betrachten. Hoffentlich wird auch dieses Unternehmen mit Erfolg gekrönt werden.

3. Das Sprachwissenschaftliche Institut der Slowakischen Akademie der Wissenschaften behauptet sich im Rahmen der slowakischen Linguistik als die einzige verhältnissmässig gut aufgebaute linguistische Arbeitsstelle. An der Universität in Bratislava gibt es nur ein phonetisches Laboratorium, das für physiologische Untersuchungen gut ausgerüstet ist, sonst gibt es leider an den Hochschulen keine Arbeitsgruppe, die sich ausschliesslich der wissenschaftlichen Arbeit widmen könnte. Die pädagogische Tätigkeit beansprucht fast alle Zeit der Linguisten, die auf dem Gebiet der Slawistik, Slowakistik, Romanistik und Germanistik an unseren Hochschulen angestellt sind. Trotzdem kann das hohe Niveau mancher Aufsätze und Bücher dieser Autoren mit Genugtuung bestätigt werden. Ich führe nur das Frequenz-Wörterbuch an, das im vorigen Jahr erschienen ist.[3]

Es ist freilich kein Wunder, daß eben die Erforschung der slowakischen Sprache am erfolgreichsten ist: es wäre eigentlich ganz leicht schöne Erfolge auf dem Gebiet der Phonetik, Phonologie, Grammatik und Lexikologie der Schriftsprache und auch der Mundarten, auch aus der Geschichte der slowakischen Sprache anzuführen. Das würde uns aber ziemlich weit führen, denn die angeschnittene Problematik ist nicht nur thematisch und methodologisch sehr bunt, sondern sie wird auch von circa hundert Leuten bearbeitet (worduch die große Menge der Aufsätze auch gegeben ist).

Hoffentlich wird es nicht überflüssig sein zu bemerken, daß nur ein Bruchteil der ganzen Produktion methodologisch sozusagen ein bischen altmodisch ist: damit ist nich nur das Sammeln des historischen, dialektologischen und anderen Materials gemeint, sondern auch einige kompilatorische Werke, die meistens als Handbücher nur der Schule dienen. Der größte Teil der linguistischen Produktion in der Slowakei hat — meines Erachtens — gutes Niveau: besonders die Theorie der sog. klassischen Prager Schule der dreißiger Jahren hat dazu bedeutend beigetragen. Wir arbeiten meistens mit klassifikatorischen und komparativen Begriffen, indem wir die Sprache als ein integriertes System von zeichenartigen Elementen betrachten. Nur in kleinerem Masse machen wir Gebrauch von quantitativen Begriffen, obwohl wir glauben, daß jedes Sprachelement prinzipiell meßbar ist. In der letzten Zeit werden auch in der slowakischen Linguistik erfolgreiche Versuche gemacht, neben den induktiven auch deduktive Methoden und Verfahren anzuwenden.

Die dialektische Methode führt uns dabei zur Dämpfung der Gefahr, die Sprache und die Spracherscheinungen als ein in sich geschlossenes und isoliertes System zu betrachten: sie ermöglicht uns die Grenzen des Sprachsystems zu passieren, damit

---

[1] Úvod do matematickej jazykovedy. Bratislava 1969.

[2] Einige Beispiele: HORECKÝ, J.: The Evaluation of three-member Consonant Clusters. Asian and African Studies, Bratislava 1965, S. 112—122; Sabol, J.: Štvorčlenné konsonantické skupiny v slovenčine. Slovenská reč, *34*, 1969, 30—33; Findra, J.: Frekvencia foném v ústnych prejavoch. Jazykovedný časopis, *19*, 1968, S. 84—95; Bosák, J.: Frequency of Phonemes and Letters in Slovak. Jazykovedný časopis, *16*, 1965, S. 120—135.

[3] MISTRÍK, J.: Frekvencia slov v slovenčine. Bratislava 1969.

uns auch die Relationen und Übergänge zwischen Nachbarsystemen sichtbar werden, was sowohl in der synchronen, als auch in der diachronen Forschung große Bedeutung hat. Die dialektische Methode ermöglicht uns auch das Gleichgewicht zwischen dem System und seinen Teilen zu halten, was besonders in der synchronen Beschreibung der Sprache vom großen Nutzen ist.

4. Meine Damen und Herren!

Gestatten Sie mir noch diese Worte zu sagen:

Ich danke Ihnen, daß Sie unsere Einladung angenommen haben und dadurch unsere Bemühungen zu unterstützen bereit sind. Fühlen Sie sich hier in den Räumen unserer Akademie gut, angenehm — wie zu Hause!

J. Ružička

The Role of Models in Linguistic Studies

JÁN HORECKÝ, BRATISLAVA

In the theory of modelling which becomes an important methodological tool in scientific inquiries an already classical list elaborated by Youen Ren Chao [1] is known in which about thirty various concepts of the term and also of the notion model are given. They are mainly the conceptions of linguists, but some philosophical conceptions can be found there too. It is remarkable, however, that these conceptions vary more in linguistic papers than in philosophical ones and that even the same author uses the term model in more meanings. It is perhaps because the terms of a developing branch of science are not made quite clear at the beginning. For instance Ch. Hockett uses the term model both as a system and description or as a way of using language. Similarly Z. Harris considers the model as a pattern of language functioning system, a special kind of grammar, a theory of structure and also as an interpreted or partly interpreted system of signs. By N. Chomsky the model is also conceived as a conception of language structure, as a theory or simply as a grammar.

Though there are many questions that are not clear in using the term model and probably there are also some unprecisenesses, it is remarkable that the notions structure, system and theory are often used in operating with the term model. It is the influence of the general development in the methodology of sciences as well as of the linguistic structuralism. In this way the linguistic conception of model becomes closer to the philosophical conception of model.

It is also necessary to say that like in philosophy in linguistics there is often not the question of defining the notion of model in general but only of explaining an individual conception of this notion as used in individual studies. In this sense it is possible to add also for instance the notion of S. K. Šaumian to the above-mentioned list by Chao. S. K. Šaumian identifies the model with the theory: the model is a theory with concrete contents in the form of pattern used as analoga for unobservable elements. S. Marcus [2] considers as model any formation different from an original and having some common isomorphical features.

Examples of various conceptions of model in philosophical papers are possible to be found for instance in the book *Logic, Methodology and Philosophy of Sciences* [3].

In this situation we can welcome the attempt of Peter Hartmann (though it is

still premature, as it seems) to systemize various conceptions of model [4]. He divides the models into two groups: in formal sciences there are always deductive models that can be taken for the realisation of a construed theory, whereas in empirical sciences (we can perhaps speak about theoretical-empirical sciences, as S. K. Šaumian does) there are inductive models that can be characterized as general conceptions intuitively achieved.

Since linguistics is not taken for formal science the models of inductive type will be mostly used in it. P. Hartmann himself, however, admits that the models of deductive type can be also used in linguistics (as e.g. the generative model of N. Chomsky), but these models are regularly resulting not from a linguistic attitude but from an extra-linguistic one. It is necessary to answer the question whether N. Chomsky does consider the model or the grammar as generative. It appears that the model as a formal element (i.e. a logical model in Revzin's [5] conception) cannot generate sentences nor describe their structure. Only a theory can do it. We can say that the model is not generative, but the grammar which is generative can be represented by a model.

According to P. Hartmann linguistics is, unlike other sciences, in a better situation because it can suitably make use of both processes, inductive and deductive, and thus it can also use both types of models.

There is also the question whether any intuitively construed general conception can be taken for a model. Since according that also the conception of phoneme as a distinctive element or of phoneme as a bundle of distinctive features could be considered as model. It is clear that a model like that has no qualities of models as summed up e.g. by Ju. D. Apresian [6]. It is evident that the phoneme as a model cannot be taken for a mechanism functionally similar to the object, because it does not form a system and it has no explanatory power.

As it is shown here there are still many problems that are not clear. In solving them it is necessary especially to investigate whether linguistics belongs to the formal or empirical (or theoretical empirical) sciences and which part of language, which qualities of language can be modelled.

At first sight it seems that linguistics does not belong to the formal sciences. Especially when examining the texts or the sets of texts it can be considered as an empirical or even taxonomic science. It is natural that only the ascertaining of occurrence of certain elements (phonemes, morphemes, words, syntagms) is not sufficient for a science, even when examining texts empirically. In using such a method we could not speak about modelling — similarly it is not possible to speak about modelling in such pure taxonomic sciences as descriptive zoology or botany are. We can explain nothing in finding out the occurrence of elements only. So we cannot build up any integrated theory only on the basis of the occurrence of some elements and their distribution in certain positions. Along with ascertaining the occurrence also relations among the elements under investigation have to examined (though they are still relative, not oppositional relations) and gradually an inductive theory has to be

construed. Using this theory the investigated elements can be evaluated. For instance, when examining the sounds we can build up certain theory (e.g. a descriptive theory) of phoneme on the basis of which we can decide what a phoneme is or by which allophones the phoneme is formed. But in using such a theory the system of phonemes cannot be explained.

In this connection it is necessary to remember the paper of R. C. Lewontin [7]. He is right when saying that the scientific knowledge is not based only on the corpus of data, but on the structured set of data. It is this structure or this structuration, not only the observable data, which are to be investigated by a science. On the base of this structure, but simultaneously also for explaining it a theory is drawn up. But on the base of this structure (as Lewontin says) the model can be construed.

It is possible to show that a theory is really construed by examining the qualities that any theory must possess. They are especially the explanatory power and the predictability. An other question arises here, whether the model can be construed directly on the base of the structure of observed data. If that were possible, esperanto would have be acknowledged as a model of one or some natural languages. It would be perhaps the Black's analogical model [8].

It is more suitable to say that a theory is drawn up in order to explain the given structure and the effectiveness of that theory can be verified not only by confrontation with observed data, but also by construing a model. This procedure is inevitable when the structure of data cannot be examined directly.

This is in accordance with the thesis of S. Marcus that the immediate models of language cannot be construed. They can be construed only by the aid of the theory: it is not a language phenomenon that is modelled, but its description (i.e. non mathematical model). In this sense every linguistic model can be considered as an interpretation of a theory. Or by other words, the validity of a theory can be verified by the interpretation of the model. In this connection the model of grammatical gender can be remembered.

But there is another question also, whether the deductive and inductive theories are not to be differed and according to them also the deductive and inductive models. The problem of the theoretical empirical science to which perhaps linguistics belongs (at least the structuralist one) is to be solved in connection with this question.

The second problem is, what is the object of modelling. From that point of view the typology given by Ju. D. Apresian is quite sufficient [9]. He distinguishes, as it is known, the models of speech action, the investigator's models and the models of linguistic description (metatheories). The most important are the models of speech action, the other two types are only auxiliary with regard to them.

Some other distinguishing elements must be added to the typology given by Ju. D. Apresian. It depends on whether only texts are examined and modelled or only the system of language or both of them. Here is the question of difference between langue and parole, incidently also of language. It is obvious that the texts can be

examined directly, whereas langue cannot be at all. Language can be investigated as a black box: we can see the input and output data, but we cannot observe immediately how the speech action takes place. It is not clear also what relation between functioning of a system and the speech action holds. Language can be considered as a mechanism that produces the texts on the base of the langue. In the input some elements must be that have been achieved by examining the already existing texts, but somewhere also the rules must be according to which also yet not observed, but still correct (grammatical) texts are produced. A theory can be construed about the situation in the language (considered as a system). Now, it is the question if this theory can be deductive or inductive. In any case this theory can be modelled by the aid of the set theory.

If it is a deductive theory following relations are valid:

language $\not\rightarrow$ model
↓
theory $\rightarrow$ model

If it is a inductive theory similar relations are valid, but the base is a corpus of texts:

corpus of texts $\not\rightarrow$ model
↓
theory $\rightarrow$ model

In both cases the model is the analogon of the theory, not of the language or of the corpus of texts.

There is another question rised by R. B. Braithwaite [10]: what for are the models like that and can they advance our knowledge? According to the modelists (as represented by Campbell) it is useful to represent a theory by a model in order to make easier the understanding of unknown theoretical notions. This situation can arise when the isomorphism holds between the hypotheses of a theory and of a model, but when the known notions are in the model. According to the contextualists (as represented by Braithwaite, Quine and others) the modus of functioning of theoretical notions in a deductive theory is given by interpreting the calculus expressing the theory or by the place of the notions in this calculus.

When applying those theories to linguistics the situation is not so clear as it could seem in Braithwaite's interpretation. In linguistics both inductive theories in the case of investigating the corpus of texts and deductive theories about the system can be construed. Both theories can be hardly compared. But any of them can be represented by the aid of a model and those models can be compared. If the hypothesis is correct that something what is not in the langue cannot either appear in the corpus of texts, or that the system of langue can be concluded from the state in the corpus of texts

then we can presume that the models would be isomorphic. This can be demonstrated on the category of case.

The existing theories about the models of case have been construed on the base of texts, i.e. inductively. I. I. Revzin, for instance, defines the case as a set of words allowed in one class of case contexts having the same capacity. As there are more such contexts in a corpus, several cases exist in a given language. It is possible to determine in which case a form of word belongs to the given case (even when it is an indeclinable word). So a field of cases arises or a space filled up with single cases. In the Revzin's model there is nothing said about the relations among the members of this space.

The deductive procedure in determining the case is based on the hypothesis that a verb action goes from one substance (agens) to another substance (patiens), i.e. that it attains the substances in the intentional space [11]. These intentional or attained substances are ordered in such a way that the action attains some of them fully, some only partly, some others peripherically, some of the substances are attained before other ones, some only after, sometimes also two substances in the same part of space can be attained. One can suppose that the arrangement of both the Revzin's space of cases and our intentional space are corresponding.

We do not want, however, to say that the above-mentioned spaces are the model of the system of cases. It is expressed only metaphorically. But it is obvious that these spaces are to be interpreted by an integrated theory construed by a deductive method from the basic axioms (e.g. that the verb action goes from one substance to other one) and primitive notions (as the intentional space, to attain etc. are) for the intentional space and by an inductive method on the base of the case forming contexts for the case field.

In any case it is a theory which is represented by the aid of a model. It is of no use then to decide whether one of the models is closer to the original, i.e. to the system than another one and to proclame an asymptotic development in the chain of models. It holds that the original for a model is neither the language nor the system, but the theory about that language or about the system.

REFERENCES

[1] CHAO, YUEN REN: Models in Linguistics and Models in General, Logic, Methodology and Philosophy of Sciences. California, Stanford, 1962. Cf. also ŠAUMIAN, S. K.: Strukturnaja lingvistika. Moskva 1965, pp. 80—82.

[2] MARCUS, S.: Introducere in lingvistică matematică. București 1966, pp. 71—73.

[3] Teorie modelů a modelování. Praha 1967. There are papers translated from the Logic, Methodology and Philosophy of Sciences as well as many other papers.

[4] HARTMANN, P.: Modellbildungen in der Sprachwissenschaft. Studium Generale, *18*, H. 6, pp. 364—379.

[5] ZINOVIEV, A. A.—REVZIN, I. I.: Logičeskaja modeľ kak sredstvo naučnogo poňatija, Voprosy filosofii, 1960, pp. 101—114.

[6] APRESIAN, JU. D.: Sovremennyje metody izučenija značenij i nekotoryje problemy strukturnoj lingvistiki. Problemy strukturnoj lingvistiki. Moskva 1963, pp, 102—110.
[7] LEWONTIN, R. C.: Models, Mathematics and Metaphors. Synthese, *15*, 1963, pp. 22—244; Teorie modelů a modelování, pp. 69—91.
[8] BLACK, M.: Models and Metaphors, Studies in Language and Philosophy, 1962.
[9] APRESIAN, JU. D.: Idei i metody sovremennoj strukturnoj lingvistiki. Moskva 1966, pp. 99—100.
[10] BRAITHWAIT, R. B.: Models in the Empirical Sciences, Logic, Methodology and Philosophy of Sciences. Teorie modelů a modelování, pp. 289—298.
[11] Cf. MIKO, F.: Rod, číslo a pád. Bratislava 1962, pp. 92—94. This is a restricted notion of intention. For the original interpretation of the intention cf. RUŽIČKA, J.: Valencia slovies a intencia slovesného deja. Jaz. časopis, *19*, 1968, pp. 50—56.

Некоторые замечания в отношении моделирования в языкознании*

АЛЕКСАНДЕР ЛЮДСКАНОВ, СОФИЯ

Лингвисты хорошо помнят все сомнения, возражения, отрицания и обвинения, которыми многие авторы встретили лет 15—20 тому назад первые теоретические предложения и практические попытки применения метода моделирования и в частности математического моделирования в языкознании. Подобные возражения и отрицания встречаются и в наши дни, может быть лишь в немного менее острой форме и при другом обосновании. Обыкновенно сторонники „новых методов" отвечают своим оппонентам, ссылаясь на общие тенденции развития современных наук, в том числе и гуманитарных, и на целый ряд фактов и логических аргументов, которые обосновывают возможность и плодотворность применения метода моделирования и в языкознании. Такая линия ведения дискуссии имеет, разумеется, свои основания. Однако к проблеме можно, а на мой взгляд и необходимо, подойти с другой стороны и поставить следующий вопрос: не существует ли как в теоретических суждениях о моделировании в лингвистике, так и при практических попытках, применения этого метода в конкретных языковедческих исследованиях, невыясненных положений, противоположных и ошибочных мнений, которые, с одной стороны, бы препятствовали успешному практическому применению моделирования в языкознании и вели бы к неудовлетворительным результатам, а с другой — в силу именно этих неудовлетворительных результатов и неясностей давали бы реальные аргументы в руки противников применения этого метода?

К сожалению, следует признать, что в языкознании дело обстоит именно так. Это положение вещей является результатом в значительной степени недостаточной теоретической ясности в отношении сути моделирования в нематематических областях, несмотря на наличие довольно богатой, но в основном фрагментарной литературы по этим вопросам. Восполнение этого пробела является несомненно одной из насущнейших научных задач в этой области. Выполнение этой задачи предполагает ряд детальных исследований и, конечно, далеко выходит за рамки одного доклада. Поэтому

---

* Настоящее изложение представляет собой русскую версию доклада автора Quelques remarques sur la notion de modèle en linguistique.

лишь некоторыми предварительными уточнениями, необходимыми для дальнейшего изложения.
здесь я поставлю себе лишь следующие ограниченные и предварительные цели: в начале будет дано краткое описание сути метода эвристического моделирования в языкознании (§ 1); а затем на этой основе будут выделены и проанализированы фазы реализации этого типа моделирования (§ 2); в заключении будут сделаны некоторые выводы и уточнения.

Само собой разумеется, что объем настоящего изложения позволяет наметить лишь некоторые проблемы и высказать лишь некоторые суждения только в самых общих линиях, не имея возможности в большинстве случаев даже ссылаться на соответствующую литературу. Его целью не является установление неких истин последней инстанции, а в основном изложение точки зрения автора и создание основы для дискуссии и размышления.

§ 1

В этом параграфе после некоторых вступительных замечаний и уточнений (I) мы опишем в самых общих чертах суть моделирования в языкознании (II) и выделим некоторые его основные типы (III). Дополнительные ограничения предмета наших рассуждений будут сформулированы в ходе самого изложения.

I. Не будучи в состоянии останавливаться на истории вопроса (см. напр. [1 и 2]), отметим, что в наши дни метод моделирования вышел далеко за рамки так наз. традиционных областей его применения и превратился в общенаучное достояние, в могучее орудие в руках ученых в деле познания и практического преобразования мира.

Все более широкое применение этого метода привлекло к нему (и продолжает привлекать) пристальное внимание не только математиков, физиков, химиков, биологов, но и философов, логиков, историков, языковедов и пр. И может быть обстоятельство, что современная теоретическая литература по проблемам моделирования принадлежит перу представителей столь различных областей и является одной из существенных причин наличия в ней столь различных определений, мнений и взаимоисключающихся взглядов почти по всем основным проблемам в этой области (напр. о соотношении моделирования в математических и нематематических областях, о его соотношении с другими частными методами научного познания и с основными категориями гносеологии; по проблемам определения самой сути этого процесса и понятия модели; о возможных оригиналах и моделях и типах логической связи между ними; о наглядности моделей, их познавательной роли, области применения и пр.). Не имея, конечно, возможности останавливаться на всех этих вопросах, я ограничусь здесь

1. В отличие от некоторых авторов, которые или рассматривают модели лишь в строгом логико-математическом смысле (см. напр. [3]), или придерживаются взгляда, что ввиду различия в степени их абстрактности по отношению к оригиналу, модели в области математики принципиально отличаются от моделей во всех других областях (см. напр. [4]), а также в отличие и от тех авторов, которые считают, что оригиналом модели в математике могут быть только теории (аксиоматично-дедуктивные системы; см. [5]), мы будем исходить из более широкой концепции Софийской математической школы, изложенной напр. в [6] и считать, что модель в математике может функционировать, так сказать, в двух направлениях: пусть S — некоторая математическая структура; она, с одной стороны, может быть моделью (М) некоторых фактов действительности или других менее абстрактных математических структур а с другой — реализацией, интепретацией некоторых более абстрактных математических структур или формальных систем.

Исходя из этого по признаку большей или меньшей степени абстрактности модели в отношении к оригиналу можно выделить два типа моделирования как в математике, так и в нематематических областях: *первый*, при котором модели более абстрактны, чем оригиналы (этот тип моделирования характерен для нетрадиционных областей применения этого метода и в первую очередь для языкознания) и *второй*, при котором модель менее абстрактна, чем оригинал (напр. интерпретация некоторых структур Маркова при помощи фактов языка). Предметом дальнейшего изложения будет только первый из этих двух типов моделирования.

2. Метод моделирования применяется, как в теоретических областях, в качестве способа приобретения новых знаний, так и в прикладных областях, в качестве приема установления самого рационального и действенного способа применения на практике приобретенных теоретических знаний. Первый из этих двух типов моделирования мы будем называть *эвристическим*, а второй условно, *прагматическим*. В дальнейшем мы будем говорить лишь об эвристическом моделировании (ЭВМ).

3. Довольно часто (в результате непосредственного влияния кибернетики) под моделированием понимают воспроизведение некоторых умственных деятельностей человека. Несмотря на то, что реализация и экспериментирование моделей на ЭВМ является весьма желательным, под моделированием мы будем понимать определенный тип научной умственной деятельности человека, а не машинную реализацию ее элементов. Кроме этого некоторые авторы выделяют в языкознании особый класс моделей — так наз. действенные модели, — т. е. „устройства, способные понимать заданные предложения и строить предложения по заданному значению“ [7]. На наш взгляд под моделью следует понимать не устройство (конечно, это

зависит от того, что понимается под устройством; см. напр. правильное замечание в [8]), а описание, которое при представлении его в соответствующей форме может быть реализовано данным устройством.

II. Имея в виду указанные ограничения и уточнения, мы попытаемся теперь привести краткую характеристику самой сути эвристического моделирования в языкознании, исходя из следующих положений: что „модель следует всегда рассматривать как отображение. Вопрос состоит в том, что отображается и как выглядит функция отображения“ [2]; из того, что моделирование является способом приобретения новых знаний, и из того, что приобретение любых научных знаний (в том числе и методом моделирования) по необходимости предполагает отвлечение от всего богатства явлений и переход к их сущности.

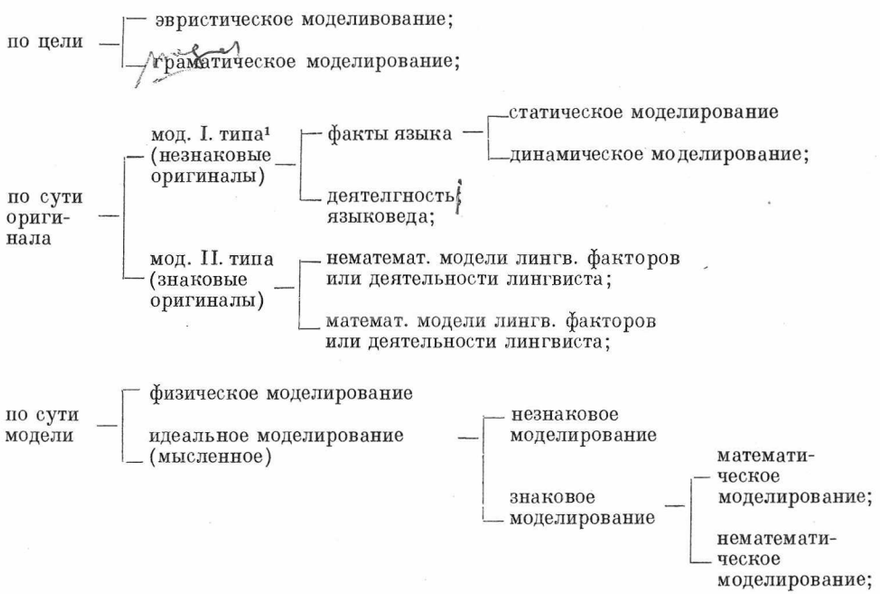
В целях нашего изложения можно выделить два способа приобретения новых научных знаний: непосредственный и опосредствованный. *Непосредственный* подход бывает только эмпирическим: при нем новые знания получаются в результате анализа свойств *самого интересующего нас объекта* (скажем А). *Опосредствованный* подход может быть эмпирическим или логическим. При *эмпирическом* опосредствованном подходе новые знания получаются в результате анализа не самого объекта исследования (А), *а некоторого его заменителя* (модели — см. ниже), скажем, (В). При логическом опосредствованном подходе новые знания получаются также не в результате непосредственного анализа объекта исследования (А), а на основе установления *его реляций, зависимостей по истинности* с некоторой совокупностью уже приобретенных знаний, в отношении, скажем, (В). Зависимость мыслей по истинности называется логичностью [10]. Так, напр. если мы знаем, что А истинно, то тем самым мы приобретаем знание и в отношении $>$ А (не-А), (т. е. мы уже знаем, без его непосредственного анализа, что не-А не может быть истинным).

Как видно из сказанного, общее и самое важное при опосредствованном подходе заключается в том, что новые знания получаются не в результате анализа самого объекта исследования (А), а его заменителей (В) и их реляции. Так, вместо того, чтобы анализировать естественный поток звуков речи (А), можно анализировать, напр., некоторую представляющую его Марковскую цепь (В); вместо того, чтобы анализировать реальные распределения некоторых единиц в тексте (А), можно анализировать, напр., их Пуассоновское распределение (В); вместо того, чтобы анализировать непосредственно (да это и вообще невозможно) сам механизм языка (langue) как систему, порождающую речь (А), можно анализировать в качестве его упрощенного представления некий автомат, скажем, с конечным числом состояний (В); или вместо того, чтобы анализировать непосредственно некие свойства данного естественного языка (А), можно анализировать

некий искусственный язык, скажем, безконтекстный язык (В).

На основании этого в первом приближении можно сказать, что суть моделирования первого типа заключается в том, *что при нем новые знания приобретаются не в результате непосредственного анализа объекта исследования* (*который обыкновенно недоступен непосредственному наблюдению*), *а в результате опосредствованного анализа сознательно созданной* (*или подобранной*) *модели.* Но опосредствованный характер приобретения новых знаний при моделировании является только, так сказать, его первым „дифференциальным“ признаком. Второй отличительный признак заключается в реализации, экспериментировании и эвентуальном изменении моделей.

III. Не будучи в состоянии останавливаться на проблемах классификации (и ее основах), мы выделим лишь некоторые типы моделирования в языкознании в целях ограничения дальнейшего изложения:



[1] Это разграничение можно принять, имея в виду следующую оговорку: строго говоря, при исследовании фактов языка (как и любых других) мы обыкновенно оперируем не самими фактами, а их второсигнальными отражениями в нашем сознании; следовательно, опять-таки строго говоря, и оригиналы первого типа моделирования являются знаковыми; введенное подразделение, может быть и нуждающееся в терминологической коррекции, подчеркивает факт, что в этом случае оригиналы не являются знаковыми моделями или математическими структурами.

Как следует из изложенного выше в II, целостная реализация процесса эвристического моделирования в любой области, в том числе и в языкознании, предполагает *две основные фазы*: создание модели и ее реализацию и экспериментирование. Их рассмотрению и посвящены обе части — A и B — этого параграфа.

A. Анализируя первую фазу реализации указанного типа моделирования, мы введем сперва понятие „модель-объект" (I), рассмотрим более подробно некоторые его характерные свойства (II), а затем охарактеризуем и другие звенья этой фазы (III).

I. Обыкновенно молчаливо принимается, что процесс моделирования сводится к созданию моделей. Однако, анализ его первой фазы показывает, что она протекает в значительной степени более сложно и предполагает переход через известное промежуточное звено. Мы постараемся показать наличие и суть этого промежуточного звена, исходя из анализа трех примеров.

1. Возьмем сперва один из самых распространенных в нетрадиционных областях, в том числе и в языкознании способ построения динамических моделей (напр., все действенные модели семантического синтеза советских авторов — см. напр. [8]), известный под названием „черного ящика" (black box) (см. напр. [11]), механизм которого сводится к следующему. Пусть имеется некоторая функциональная система со скрытым внутренним механизмом, в отношении которого мы ничего положительного не знаем, и представим себе, что мы хотим исследовать закономерности функционирования этой системы: для достижения этой цели мы можем абстрагироваться от ее внутреннего механизма, и мыслить ее в качестве черного ящика и концентрировать наше внимание только на состояниях входа и выхода системы и на соотношениях между ними. В результате такого подхода мы можем представить эту систему, которая в данном случае является оригиналом (0) нашей познавательной процедуры, весьма упрощенно в качестве упорядоченной пары переменных вход-выход ⟨B, B′⟩. Затем это упрощенное, схематизированное представление[2] может быть описано в некотором коде и в некоторой логике (напр., нематематическими или математическими средствами). Это описание и будет моделью нашей системы. Графически рассмотренный процесс можно представить так:

O — □ — M

[2] Здесь и в подобных контекстах термин „представление" употреблен в значении совокупности наших знаний в отношении данного объекта.

Здесь важно подчеркнуть, что между оригиналом (O) и его моделью (M) появляется промежуточное звено, обозначенное в нашем примере посредством квадратика.

2. Теперь рассмотрим в общем виде весьма часто встречающиеся в современном языкознании исследования некоторых свойств единиц на графемическом уровне естественных языков, напр. частоты появления и распределения графем (или их сочетаний), установление энтропии и пр. Во всех этих случаях оригинал исследования, т. е. поток устной речи или точнее его графемический перевод, мыслится весьма упрощенно, схематизированно и идеализированно в отвлечении от их акустических и смыслоразделительных свойств, в виде последовательности случайных событий, при которой последующее событие находится в некоторой зависимости от предшествующих. Затем это упрощенное представление описывается посредством некоторой математической структуры, напр. Марковской цепи, которая и представляет собой модель соответствующего оригинала. Как видно, и в этом случае между оригиналом и моделью появляется промежуточное звено, которое можно графически представить, как и в приведенном выше примере.

3. В последнюю очередь рассмотрим построение некоторой синтаксической модели (формальной грамматики), напр., безконтекстной НС-грамматики (типа Н. Чомского, см. напр. [12]). Оригиналом в данном случае является язык (langue), т. е. скрытый от непосредственного наблюдения механизм, находящийся в голове человека и порождающий и распознающий языковые сообщения. Затем это общее представление упрощается, схематизируется и идеализируется, в результате чего рассматривается порождение лишь изолированных предложений (т. е. исследователь отвлекается от проблем гиперсинтаксиса и связанной речи) и, во-вторых, рассматривается лишь функционирование т. н. синтаксического компонента (т. е. исследователь отвлекается и от проблем семантики). Затем это упрощенное, схематизированное и идеализированное представление скрытого механизма языка описывается при помощи определенной математической структуры — конкатенативной или ассоциативной системы. Это описание и представляет собой соответствующую синтаксическую модель. Как видно, и в том случае между оригиналом моделью появляется промежуточное звено.

Анализ этих трех примеров позволяет утверждать, что в ходе реализации первой фазы рассматриваемого типа моделирования в языкознании между оригиналом и моделью появляется промежуточное звено. И именно это промежуточное звено мы предлагаем (см. напр. [13]) называть МОДЕЛЬЮ-

ОБЪЕКТОМ и ввести это понятие в теорию моделирования в нетрадиционных областях.[3]

II. На основании изложенного понятие модели-объекта, в качестве конструктивного звена в ходе реализации первой фазы первого типа моделирования можно определить следующим образом: *модель-объект представляет собой одно из возможных упрощенных, схематизированных и идеализированных мысленных представлений некоторого оригинала (более абстрактное, чем этот последний), которое создается (или заимствуется) сознательно и целенаправленно в ходе реализации первой фазы процесса моделирования рассматриваемого типа.* Среди существенных характерных признаков моделей-объектов, отмеченных в определении, отметим особо следующие:

1. В отличие от моделей, которые могут быть и физическими, модели-объекты могут быть только *мысленными*; они представляют собой идеальное предвидение именно того вида, о котором говорит Маркс в его известном сравнении архитектора и пчелы.

2. Модель-объект должно быть *упрощенным, идеализированным и схематизированным* (такое отвлеченное от некоторых свойств представление называется иногда формализованным)[4] представлением оригинала по той простой причине, что мы или не знаем всех свойств оригинала и стремимся применить метод моделирования, чтобы получить опосредствованным путем сведения именно об этих свойствах, или же потому, что одновременное исследование всех свойств является слишком сложным или даже порой практически невозможным.

3. *Сознательный* характер построения модели-объекта заключается, конечно, не в общем сознательном характере этой деятельности, так как любое научное познание имеет сознательный характер, а в том, что она строится ,,предумышленно" именно в качестве необходимого звена процесса моделирования как частного метода исследования. В связи с этим находится и то, что было названо *целенаправленностью* построения модели-объекта: обыкновенно исследователь стремится получить не новые знания вообще, а конкретные новые знания в отношении данной проблемы (напр., существует ли в языке ядро и трансформы) или же знания, потверждающие

[3] Появление модели—объекта в качестве конструктивного промежуточного звена можно проиллюстрировать не только анализом любых моделей в языкознании и других гуманитарных областях, но и на примере моделирования в физике, химии, биологии и пр. (см. напр. [14]).

[4] Отметим, что в современной лингвистической литературе (да и далеко не только в ней) в отношении таких понятий, как формализирование и формализм существует еще больше неясностей, чем в связи с понятиями модели и моделирования. Эти проблемы являются предметом другой работы автора.

некоторые его конкретные предположения (напр., о том или ином распределении ,,тематических" слов в терминологии Гиро). Таким образом цель исследования поставлена предварительно и сознательно, ввиду чего и модель-объект должна подчиняться этой цели.

4. Но самое важное свойство модели-объекта, которое лежит в основе всей специфики моделирования и отличает его от всех других частных методов научного исследования и в особенности от метода аналогии,[5] заключается в ее *множественном характере.* В результате того, что модель-объект отражает только некоторые (а может быть и предполагаемые) свойства и реляции оригинала, варьируя эти свойства и реляции для одного оригинала можно построить не одну, а в принципе много моделей-объектов (при одном и том же оригинале-языке, напомним все те модели-объекты, которые лежат в основе, напр., грамматик с конечным числом состояний, валентностных, аппликативных и предсказуемых грамматик).

5. В качестве последней характерной черты отметим, что тип связи между оригиналом и моделью-объектом при этом типе моделирования может быть только *приближением.*

III. Выяснение сути модели-объекта позволяет предложить и приближенные определения оригинала и модели. *Под оригиналом мы будем понимать то, что описывается моделью-объектом, а под моделью — некоторое представление или описание модели-объекта.*

Характеристикой и классификацией оригиналов здесь мы вообще не в состоянии заниматься. Представление или описание модели-объекта может быть физическим или знаковым. В результате физического представления получаются так наз. физические модели или аналоги, на которых мы тоже не будем останавливаться, несмотря на то, что и они имеют свое место в языкознании (напр., в фонетике, методике преподавания языков и пр.). Знаковое описание модели-объекта дает знаковые модели, которые могут быть нематематическими и математическими. Среди нематематических знаковых моделей для языкознания имеют особое значение так наз. логические и трансформационные модели. Математическая модель — при первом типе моделирования — это описание модели-объекта при помощи некоторых математических средств, объектов, структур (напр., математическая модель морфологического анализа получается в результате описания соответствующей модели-объекта при помощи математической структуры, эквивалентной автомату с конечным числом состояний). В зависимости от

[5] Метод аналогии сводится к тому, что ,,из сходства некоторых признаков двух или более предметов или явлений действительности делается вывод о сходстве других признаков этих предметов или явлений" [15].

характера этих математических средств математические модели могут быть детерминистическими и стохастическими, теоретико-множественными, алгебраическими, логическими и пр. Отметим, что при математическом моделировании в широком смысле слова возможны два подхода: модель-объект описывается непосредственно в алгоритмической форме (напр., первые алгоритмы для МП) или же сперва описывается при помощи некоторой математической структуры, которая затем может быть представлена в алгоритмической форме (напр., трансформирование синтаксических деревьев описывается сперва при помощи теории графов, а затем эта модель представляется в алгоритмической форме).

Приведем и краткую характеристику некоторых основных свойств моделей при рассматриваемом типе моделирования.

1. Не трудно показать, что модели (в том числе и математические), в качестве знакового представления модели-объекта, должны обладать всеми теми свойствами, которыми обладают и модели-объекты: они являются более абстрактными, чем оригиналы, представляют их упрощенное, схематизированное и идеализированное описание и обязательно должны иметь сознательный, целенаправленный и множественный характер. Отметим также, что в отличие от довольно распространенного мнения модели в предложенном понимании (речь идет о знаковых моделях) в принципе не должны обладать наглядностью.

2. Связь между моделью-объектом и моделью может быть или приближением или гомоморфизмом, а между некоторыми элементами и изоморфизмом (см. напр. [16 и 13]).

3. В конце следует подчеркнуть самое важное свойство математических моделей: построение данной математической модели является предпосылкой превращения данной содержательной задачи, принадлежащей данной предметной области в математическую задачу, подлежащую решению математическими методами и реализации на ЭВМ.

На основании всего изложенного первую фазу реализации рассматриваемого типа моделирования в языкознании графически можно представить следующим образом:

$0 \approx \text{МО} \approx /\hat{=}/ \equiv \text{М}.$

В. Вторая фаза осуществления рассматриваемого типа моделирования включает следующие звенья: обработка и реализация модели (I), ее экспериментирование и формулирование выводов (II), оценка и эвентуальное изменение модели или отказ от нее (III).

I. Обработка сводится к анализу, к исследованию модели. При нематематическом моделировании обработка осуществляется логическими методами, а при математических моделях — при помощи существующих математических методов или при помощи нового математического аппарата (именно

таким образом математическое моделирование в области лингвистики содействует обогащению и развитию самой математики). Обработка математической модели при помощи математических методов имеет не только все преимущества математического исследования, и может привести и к установлению новых свойств этой структуры (модели). А в силу существующих отношений между моделью, моделью-объектом и оригиналом мы имеем основания полагать, что новое свойство, обнаруженное в модели, должно существовать и в оригинале. Но не только это. Предположим, что некоторые объекты, оригиналы описаны при помощи некоторой математической структуры, которая является их моделью. При этом наши объекты будут обладать и всеми свойствами, которые проистекают из системы аксиом структуры. Таким образом несмотря на характер объектов, которые мы исследуем, и на то, из какой области науки или практики они были взяты, в дальнейшем уже нет необходимости исследовать свойства этих объектов методами данной науки или каким-нибудь другим способом — эти свойства уже предварительно известны, благодаря свойствам структуры. В результате всего этого целые области человеческого познания, связанные иногда с объектами, на которые наука впервые обращает свое внимание, оказываются предварительно завоеванными [17, с. 22]. И именно в этом заключается одна из форм проявления гносеологической, познавательной силы моделирования и математического моделирования в частности.

Затем (независимо от того, были ли обнаружены новые свойства при обработке модели или нет) начинается *реализация* модели. Так как в большинстве случаев модели в языкознании описывают какие-то механизмы или закономерности, то их реализация сводится к приписанию переменным модели некоторых значений и выведению из нее ,,искусственных“ фактов (напр., всех трансформ данной Т-грамматики). Это выведение может осуществляться или ,,вручную“ путем ,,умственного эксперимента“ (обыкновенно возможности такой реализации весьма ограничены) или после представления модели в алгоритмической форме — на ЭВМ (при этом возможно выведение недоступного человеку количества фактов, при полной объективности и точности операции, напр., всех отмеченных фраз данного языка и только их, или всех перифраз, выражающих заданный смысл в форме глубинной структуры или некоторого basic-языка). В этой возможности реализации моделей на ЭВМ заключается второе огромное преимущество моделирования.

II. После обработки и реализации модели следует ее *экспериментирование.* Как мы уже знаем, реализация модели позволяет получить ,,искусственные“ факты. Сопоставление ,,естественных“ фактов с ,,искусственными“ и есть то звено процесса моделирования, которое мы назвали экспериментированием модели или проверкой модели посредством практики,

на основании результатов которой делаются соответствующие выводы.

III. *Выводы*, сделанные в результате экспериментирования модели, и на основании которых будет проводиться ее оценка, условно можно поляризировать следующим образом: проверка практикой потверждает (1) или не подтверждает (2) модель.

1. Экспериментирование *подтверждает* модель тогда, когда ,,искусственные" факты совпадают в определенном проценте и аспекте с ,,естественными", причем в принципе модель должна позволять получить больше фактов, чем те, которые наблюдались при создании соответствующей модели-объекта; в этом, между прочим, заключается прогнозирующая, проспективная сила моделирования (см. напр. [18]). При этом положении вещей процесс моделирования заканчивается положительным результатом.[6]

2. Экспериментриование *не подтверждает* модели тогда, когда ,,искусственные" факты полностью (что бывает редко) или в определенном проценте несовпадают с естественными. В этом случае перед исследователем становится проблема следующего творческого решения: или изменить модель или отказаться от нее. Возможность (а иногда и необходимость) изменения является одной из самых характерных особенностей метода моделирования. Эта возможность, являющаяся логическим следствием множественного характера модели-объекта и самой модели, существует как в отношении модели-объекта, так и в отношении модели. При этом один из первых способов изменения модели сводится к изменению стоимостей, которые приписываются ее переменным. Если процент совпадания ,,естественных" и ,,искусственных" фактов весьма низок, то исследователь может придти к выводу о необходимости вообще отказаться от данной модели-объекта и модели; отметим, что установление неадекватности данной модели в отношении данного оригинала и необходимости отказа от нее является также определенным научным достижением (см. напр. [13]).

Такова вкратце суть второй фазы рассматриваемого типа моделирования в языкознании. Продолжая приведенное выше графическое представление первой фазы, в целости его можно представить так:[7]

---

[6] В связи с этим ставятся чрезвычайно важные в гносеологическом отношении проблемы: проблема степени достоверности (вероятностный или логический характер) полученных в результате моделирования новых знаний и проблема, которую некоторые авторы называют ,,превращением" модели в теорию, как и проблема места и роли модели между гипотезой и теорией.

[7] О — оригинал; МО — модель-объект; М — модель; Об — обработка и реализация модели; Э — экспериментирование модели и ее проверка практикой; В — выводы; возможные изменения модели-объекта и модели показаны стрелками; символ $\approx$ означает приближение, символ $\widehat{=}$ гомоморфизм, а символ $\equiv$ изоморфизм.

$$О \approx МО \approx | \mathrel{\widehat{=}} | \equiv М — Об — Э — В.$$

Мы постарались описать в общих линиях процесс эвристического моделирования первого типа в языкознании. Как было указано, кроме первого типа моделирования, при котором оригинал не является знаковой системой (моделью, структурой), существует и второй тип моделирования, при котором оригинал является или некоторой моделю или математической структурой. Однако не следует полагать, что этот второй тип моделирования возможен только в математике. Напротив он возможен и в нетрадиционных областях, в том числе и в лингвистике. Описание этого типа моделирования выходит за рамки настоящего изложения и мы отметим лишь то, что оригиналом при нем является математическая модель некоторого лингвистического оригинала, полученная в результате реализации первого типа моделирования. Реализация этого второго типа моделирования в языкознании составляет часть предмета той области, которую называют математической лингвистикой, в отличие от применения первого типа моделирования в лингвистике, который предоставляет собой лишь математическое моделирование естественных языков.

На основании изложенного можно сделать некоторые выводы и уточнения.

1. О наличии и применении метода моделирования в языкознании можно говорить только тогда, когда реализованы все его фазы и их соответствующие звенья, несмотря на то, что и самостоятельная реализация его первой фазы имеет определенные преимущества в сравнении с традиционным описанием.

2. Представляя собой средство опосредствованного получения новых знаний, метод моделирования предполагает, очевидно, известные аналогии между существенными (с точки зрения данного исследования) свойствами и реляциями оригинала и модели, однако, как было указано выше, он отличается в принципе от метода аналогии в науке, ввиду чего не представляется обоснованной постановка знака равенства между ними.

3. Не следует ставить знак равенства и между методом моделирования и методом формализации. Очевидно, что создание модели-объекта и самой модели предполагает известное отвлечение, абстрагирование, т. е. формализирование в смысле представления в обобщенном виде; однако, такое формализирование является необходимой предпосылкой любого способа научного познания, но не покрывается с другими частными методами научного исследования в их целости. Еще более строго следует проводить различие между моделированием, в том числе и математическим моделированием первого типа и формализацией в строгом логико-математическом

смысле слова (см. напр. [19] и [3]). Необоснованным является и довольно распространенное среди лингвистов мнение, что простое введение некоторых количественных данных, графов, матриц и пр. является уже математическим моделированием, формализацией, математизацией и пр. науки о языке.

4. Также не следует смешивать эвристическое моделирование с предвидением. В то время как предвидение представляет собой основу любой сознательной человеческой деятельности, как в науке, так и в повседневной жизни (см. напр. [6]), не любое предвидение есть моделирование, которое должно удовлетворять изложенным выше условиям. То же самое следует подчеркнуть и о соотношении моделирования и схематизирования, упрощения и идеализирования. Эти приемы представляют собой необходимые компоненты моделирования, но не покрываются с ним.

5. Наглядность не является необходимым элементом знаковых моделей, конечно, за исключением той, которой обладают означающие стороны любого знака (см. напр. [13]).

6. Поскольку при первом типе моделирования модель всегда является огрубленным и упрощенным представлением оригинала, постольку и знания, полученные в результате такого моделирования, не описывают и не могут описывать оригинал во всем богатстве его свойств. Эта „неполнота" получаемых новых научных знаний является причиной одного из постоянных обвинений в адрес моделирования в языкознании — обвинения в упрощении. Но: „кто знает иной способ понимания сложных вещей, чем понимание посредством их упрощения?" ([18], стр. 110).

7. В связи с оценкой приобретаемых в результате моделирования знаний следует отметить и следующее чрезвычайно важное обстоятельство. В тех (поистине немногочисленных) работах лингвистов, в которых оценка результатов моделирования в языкознании основывается на некотором фактическом и логическом анализе, а не на утверждениях a priori и предвзятых идеях, обыкновенно рассуждения ведутся следующим образом. „Искусственные" факты, порождаемые моделью, сравниваются с тем или иным естественным языком и это сравнение бывает обыкновенно отрицательным: модели не описывают всего богатства естественных языков. Но при этом упускается из виду следующее. Модель является описанием модели-объекта при помощи некоторых средств. Из этого явственно следует, что сила, эффективность, адекватность и пр. модели зависят по крайней мере от двух факторов: от уровня и степени адекватности модели-объекта или точнее лингвистических концепций, на которых она построена, и от мощности, адекватности и пр. описания, принятого для модели. Следовательно, адекватность и богатство знаний, полученных в результате моделирования, зависит от двух факторов, а не только от одного!

8. Как следует из изложенного, метод эвристического моделирования в языкознании (как и в любой другой области) имеет гносеологическую силу: он позволяет получить новые научные знания опосредствованным путем. В связи с этим следует особо остановиться на следующем вопросе. В последнее время все чаще встречается утверждение, что моделирование и в частности математическое моделирование, не только возможное, но и необходимое в области прикладного языкознания, неприменимо и недопустимо в области теоретического языкознания. Не будучи в состоянии обсуждать подробно этот взгляд в настоящем изложении, необходимо указать со всей определенностью его полную ошибочность. Задача теоретических дисциплин — языкознания в том числе — именно в приобретении новых знаний, к чему и ведет метод моделирования. Кроме этого одной из самых важных задач современного языкознания — исследовать механизмы языка, благодаря конкурирующему действию которых он осуществляет свою основную функцию — коммуникативную. А эти механизмы недоступны непосредственному наблюдению и, следовательно, единственным методом научного познания является в данном случае метод моделирования. Современная практика моделирования в языкознании показывает несостоятельность и утверждений, что моделирование неприменимо к диахронии, к историческому развитию языка, в области диалектологии и пр. (см. напр. [20]).

9. В конце отметим ошибочность и мнения, в силу которого моделирование, и в особенности математическое моделирование, противопоставляется всем другим методам исследования языка и в первую очередь традиционным. В действительности такого противопоставления и взаимного исключения нет, а существует непосредственная преемственность к взаимозависимость: ведь создание модели-объекта необходимой для реализации процесса моделирования, основывается на всех наших знаниях о языке, приобретенных в ходе исторического развития языкознания любыми методами, в том числе и традиционными. Следовательно, чем глубже наши знания о языке, приобретенные любыми способами, тем более адекватные результаты будет давать и метод моделирования. А из этого следует, что успешное развитие науки о языке предполагает преемственность и сотрудничество, а не взаимное отрицание „старых" и „новых" методов.

ЛИТЕРАТУРА

[1] ROSENBLATH, A.—WIENER, N.: The role of models in science. Philosophy of Science, 1945, Vol. 12, No 4; см. также
BRAITHWAITE, R. B.: Models in the empirical sciences. В сб. Logic, methodology and philosophy of science, No 6; BRAITHWAITE, R. B.: Scientific explanation. New York 1960.

[2] ШТОФФ, В. А.: О роли моделей в познании. ЛГУ, 1963; см. также ШТОФФ, В. А.: Гносеологические функции модели. ВФ, 1961, № 12.
[3] BADIOU, A.: Le concept de modèle. Théorie. Cours de philosophie pour scientifiques IV, Paris, François Maspero 1968.
[4] KLAUS, G.: Kybernetik in philosophischer Sicht. Berlin 1961.
[5] ШАУМЯН, С. К.: Структурная лингвистика. Москва 1965.
[6] а) ИЛИЕВ, Л.: Математика как наука о моделях. Материалы международной конференции по математическому моделированию, Варна 1970.
б) СЕНДОВ, Бл.: Общи принципи на математическото моделиране. В сб. Икономика, математика и кибернетика, Варна 1968.
[7] АПРЕСЯН, Ю. Д.: Экспериментальное исследование семантики русского глагола. Москва 1967.
[8] ЖОЛКОВСКИЙ, А. К.—МЕЛЬЧУК, И. А.: К построению действующей модели языка „смысл↔текст". Машинный перевод и прикладная лингвистика, вып. *11*, 1969.
[9] FREY, G.: Symbolische und ikonische Modelle. Synthese, vol. XII, 1960, No 2/3.
[10] СПАСОВ, Д.: Симболна логика, София 1969.
[11] BUNGE, M.: A general Black Box Theory. Philosophy of Science, *30*, 1963, 346.
[12] CHOMSKY, N.: Three models for the description of language. IRE Transaction on information Theory, IT *2*, 1956 (русский перевод в Кибернетический сборник, № 2).
[13] BUNGE, M.: Les concepts de modèle. L'Age de la science, 1968, No 3, Dunod, Paris.
[14] СЕНДОВ, Бл.—ЦАНЕВ, Р.: Об одной модели дифференциации клеток на основе существования эпигенетического кода. Материалы международной конференции по математическому моделированию. Варна 1970, см. также LIEB, E. H.—MATTIS, D. C.: Mathematical Physics in One Dimension. New York, Academic Press 1966.
[15] АНДРЕЕВ, И. Д.: О методах научного познания. Москва 1964.
[16] ASHBY, W.: An introduction to Cybernetics. London 1957.
[17] ИЛИЕВ, Л.: О некоторых вопросах научного познания и использования его результатов. Международный симпозиум Управление, планирование и организация научных и технических исследований, Москва 1968.
[18] FRANK, Ph.: Philosophy of Science. Prentice-Hall 1957.
[19] KOEENE, S. C.: Introduction to Metamathematics, New York — Toronto 1952 (русский перевод, Москва 1967).
[20] AFENDRAS, E. A.: Mathematical Models for Balkan Phonological Convergence. International Conference on Computational Linguistics COLING, Stockholm 1969. SMITH, R. N.: Automatic Simulation of Historical Change. International Conference on Computational Linguistics. COLING, Stockholm 1969.; SKALMOWSKY, N. — OVERLEKE, M. van: Computational Analysis of Interference on the Lexical Level. COLING, Stockholm 1969; WOOD, G. R.: Dialectology by Computer. COLING, Stockholm 1969.

La proportion linguistique et la notion de groupe. Essai pédagogigue

YVES GENTILHOMME, BESANÇON

Dans cet article, l'auteur, en se fondant sur un exemple de modèle mathématique, le groupe des permutations, et la notion de proportion en linguistique, présente un mode d'enseignement des cadres formels logico-mathématiques à des étudiants linguistes de formation littéraire.

1. Avant propos

Le problème que l'auteur se propose d'aborder ici concerne bien et la linguistique et l'algèbre, cependant le but poursuivi est, peut-on dire, inverse de celui que se posent les linguistes-algébristes-logiciens.

En effet, ce que l'on recherche habituellement c'est un modèle formel susceptible de décrire certains faits de langue. Le modèle vise à l'exhaustivité, à la précision, à la rigueur, à l'économie, à la simplicité. Un écart par rapport aux faits observés, l'incapacité de rendre compte de certains résultats sont considérés comme une faiblesse du modèle proposé.

A un niveau d'abstraction plus élevé, on élabore des systèmes formels généraux, susceptibles d'être utilisés, avec plus ou moins de succès, comme modèles linguistiques dans des situations particulières.

Dans cette étude, l'objectif poursuivi est tout autre. Il se pose en termes de pédagogie: comment enseigner de façon efficace les cadres formels aux linguistes de formation littéraire?

Par efficace, il faut entendre: tel que, par la suite, ces linguistes se servent effectivement des connaissances acquises.

„Savoir par coeur n'est pas savoir", disait Montaigne dans ses *Essais*.

Il nous importe peu que nos élèves sachent ou non réciter par coeur le cours que nous leur avons exposé, qu'ils soient capables de dominer toutes les subtilités des démonstrations mathématiques. Le but poursuivi ici n'est pas de former de futurs mathématiciens, mais des linguistes, armés de certains rudiments de l'appareil logico-mathématique *naïf*, qu'ils pourront utiliser dans la pratique quotidienne de

leurs recherches. Cet appareil doit leur permettre d'exprimer avec plus de clarté, de simplicité, de précision, de rigueur, les propriétés des langues qu'ils analysent à des fins, pratiques ou théoriques, diverses: enseignement, traitement en machine, traduction ... Il importe qu'ils apprennent à découvrir les limites de validité d'un modèle, la notion d'approximation et, de ce point de vue, un modèle inadéquat peut parfois rendre, pédagogiquement parlant, plus de services qu'un modèle où il n'y a rien à redire.

Il faut bien se persuader que la recherche n'est pas seulement le fait de quelques génies, c'est une oeuvre collective à laquelle prennent part aussi bien des êtres d'élite capables d'être à la fois de bons mathématiciens, de bons logiciens et de bons linguistes, que ces êtres moins éclectiques que sont, par exemple, des linguistes méritant mais qui n'accèdent que lentement, à la suite d'efforts considérables, aux exposés des théoriciens formalistes et inversement.

Les uns comme les autres ont leur contribution à apporter. Un dialogue doit pouvoir s'établir entre les diverses tendances. S'il est rentable d'enseigner à certains les mathématiques comme à des mathématiciens, l'expérience montre que, dans d'autres cas, les étudiants ou bien enregistrent les théories sans en tirer aucun parti dans la pratique, ou bien sont incapables de les assimiler et abandonnent les cours.

Il nous appartient de supprimer les cloisons par un effort pédagogique dynamique qui crée la motivation pour les mathématiques, qui assouplisse la présentation des matières de façon à les rendre assimilables, qui montre comment on peut faire usage de notions théoriques même dans le travail routinier de la recherche.

2. Corpus et programme

Passons aux exemples concrets.

Soit à enseigner la notion de groupe. Si je présente, répétons-le, cette notion dans le cadre mathématique traditionnel comme étant une certaine structure d'ensemble douée de certaines propriétés, je ne serai pas suivi par mon auditoire. Les étudiants les plus appliqués, et en un certain sens les plus passifs, apprendront par coeur la définition imposée et la réciteront le jour de l'examen pour obtenir une bonne note. Les étudiants doués d'un sens critique se demanderont pourquoi je leur enseigne cette notion et la devise:

„Apprends, apprends toujours, quand tu seras grand tu verras bien pourquoi" ne leur suffira pas.

Pour me mettre en accord avec leur façon d'appréhender le monde, je déduirai cette notion comme résultant d'une certaine expérience linguistique.

Soit le corpus:[1]

1. *Père avare, fils prodigue.*

2. *Le texte est à la langue ce que le procès est au système.*
(d'après Hjelmslev).

3. *Le terme „conductance" se forme à partir de „conductivité" comme „résistance" à partir de „résistivité".*

4. *Seulement ces passe-temps courants et somme toute assez anodins sont jeux à nous ce que sont d'aimables jeux de société aux jeux sanglants du cirque.*
(Nathalie Sarraute, *Martereau*)

5. *Vous ne publiez pas de courrier du coeur, certes, mais un carnet du jour: c'est la même chose, le carnet du jour étant au salon ce que le courrier du coeur est à l'office. Tout est affaire de catégorie sociale.*
(Jacques-Arnaud Penent, *Un printemps rouge et noir*)

6. *Votre rayon, si je puis dire, ce n'est pas le noir, c'est le rose, ce n'est pas le crime, c'est la sucrerie. A mes yeux, il n'y a pas de différence: dans les deux cas, il s'agit de commerce. On ne vous lit pas, on vous consomme.*
(ibidem)

7. *Cinq hommes ont révolutionné notre temps: Kierkegaard qui a découvert l'angoisse, Marx la faim, Freud le sexe, Einstein la relativité, Curie la radioactivité.*
(A. Malraux)

8. *Qui veut aller loin ménage sa monture.*

9. *L'italique est à l'imprimerie ce que le gros plan est au cinématographe.*
(S. Antonio. „Faut-il vous l'envelopper?")

10. *„Ce que se disent ses [Nathalie Sarraute] personnages est à ce que se disent les héros de Joyce, de Virginia Woolf ou de Faulkner ce que les mouvements surpris par la physique moléculaire sont aux figures de la physique macroscopique, ce que l'eau souterraine est aux cristallisations de la surface".*
Gaëtan Picon, *L'Usage de la lecture*, 276—280
(Mercure de France, éd. 1961)

11. *L'adverbe est au verbe ce que l'adjectif est au substantif. Il en résulte que, quand on change un substantif en verbe, il faut parallèlement changer l'adjectif en adverbe.*
L. Tesnière, *Eléments de syntaxe structurale*, 63.

---

[1] Tout discours-objet sera transcrit en caractères italiques par opposition au métadiscours écrit en caractères droits.

3. Les 24 permutations

Considérons la première proposition de la citation (11) de Tesnière, en portant notre attention sur la position des quatre mots significatifs que j'appellerai *termes*:

adverbe, verbe, adjectif, substantif

Il ne nous appartient pas ici de juger de la vérité du contenu significatif de cette assertion, examinons-la en quelque sorte de l'extérieur et voyons ce qui se passe quand on modifie l'ordre des 4 termes. Parmi les 24 propositions que l'on obtient, certaines paraissent avoir quelque chose de commun avec la citation proposée, d'autres donnent nettement l'impression d'être des contre-sens, par rapport à la première, peut-être même d'une façon plus catégorique des non-sens (cf. liste, ci-dessous).

Une telle conclusion ne s'impose pas pour n'importe quelle proposition. Ainsi, les 24 permutations des quatre termes:

Malgré les remontrances de Joséphine, Alphonse, toujours, ronfle bruyamment

conduisent à des phrases toutes compréhensibles, véhiculant sensiblement la même information, susceptibles d'être prononcées avec des intonations de voix appropriées, bien que certaines, surtout sous la forme écrite, écorchent quelque peu nos oreilles.

La situation est encore différente avec:

Jean boude Marie Dupont depuis leur dispute; qui boude qui? qui porte le nom de Dupont?

1234	(1)	l'*adverbe* est au *verbe* ce que l'*adjectif* est au *substantif*
1324	(2)	l'*adverbe* est à l'*adjectif* ce que le *verbe* est au *substantif*
1342	(3)	l'*adverbe* est à l'*adjectif* ce que le *substantif* est au *verbe*
1432	(4)	l'*adverbe* est au *substantif* ce que l'*adjectif* est au *verbe*
1423	(5)	l'*adverbe* est au *substantif* ce que le *verbe* est à l'*adjectif*
1243	(6)	l'*adverbe* est au *verbe* ce que le *substantif* est à l'*adjectif*
2143	(7)	le *verbe* est à l'*adverbe* ce que le *substantif* est à l'*adjectif*
2134	(8)	le *verbe* est à l'*adverbe* ce que l'*adjectif* est au *substantif*
2314	(9)	le *verbe* est à l'*adjectif* ce que l'*adverbe* est au *substantif*
3214	(10)	l'*adjectif* est au *verbe* ce que l'*adverbe* est au *substantif*
3124	(11)	l'*adjectif* est à l'*adverbe* ce que le *verbe* est au *substantif*
3142	(12)	l'*adjectif* est à l'*adverbe* ce que le *substantif* est au *verbe*
3412	(13)	l'*adjectif* est au *substantif* ce que l'*adverbe* est au *verbe*
4312	(14)	le *substantif* est à l'*adjectif* ce que l'*adverbe* est au *verbe*
4132	(15)	le *substantif* est à l'*adverbe* ce que l'*adjectif* est au *verbe*
4123	(16)	le *substantif* est à l'*adverbe* ce que le *verbe* est à l'*adjectif*
4213	(17)	le *substantif* est au *verbe* ce que l'*adverbe* est à l'*adjectif*

2413	(18)	le *verbe* est au *substantif* ce que l'*adverbe* est à l'*adjectif*
2431	(19)	le *verbe* est au *substantif* ce que l'*adjectif* est à l'*adverbe*
2341	(20)	le *verbe* est à l'*adjectif* ce que le *substantif* est à l'*adverbe*
3241	(21)	l'*adjectif* est au *verbe* ce que le *substantif* est à l'*adverbe*
3421	(22)	l'*adjectif* est au *substantif* ce que le *verbe* est à l'*adverbe*
4321	(23)	le *substantif* est à l'*adjectif* ce que le *verbe* est à l'*adverbe*
4231	(24)	le *substantif* est au *verbe* ce que l'*adjectif* est à l'*adverbe*

4. Notations condensées

Il n'y a pas besoin d'être mathématicien pour se rendre compte qu'une telle liste est fastidieuse à écrire et à lire, aussi des notations „laconiques", où l'on ne retient que l'essentiel, sont-elles les bienvenues. On peut, par ex., se contenter d'inscrire seulement les 4 termes sur lesquels on porte son attention:

„ *adverbe* „ *verbe* „ *adjectif* „ *substantif* „

ou plus brièvement:

adv., verb., adj., subst.

ou encore on peut nommer les termes par leurs numéros d'ordre dans la première proposition:

1 2 3 4

C'est ce qui a été fait dans la colonne de gauche de la liste ci-dessus: Cette dernière sténographie où, à la limite, on fait abstraction de la nature concrète des termes eux mêmes, et où l'on ne s'occupe que de l'ordre dans lequel ils sont rangés, a l'avantage de pouvoir s'appliquer à d'autres citations ou fragments de citation, contenant les mêmes termes ou des termes (mots ou suites de mots) différents.

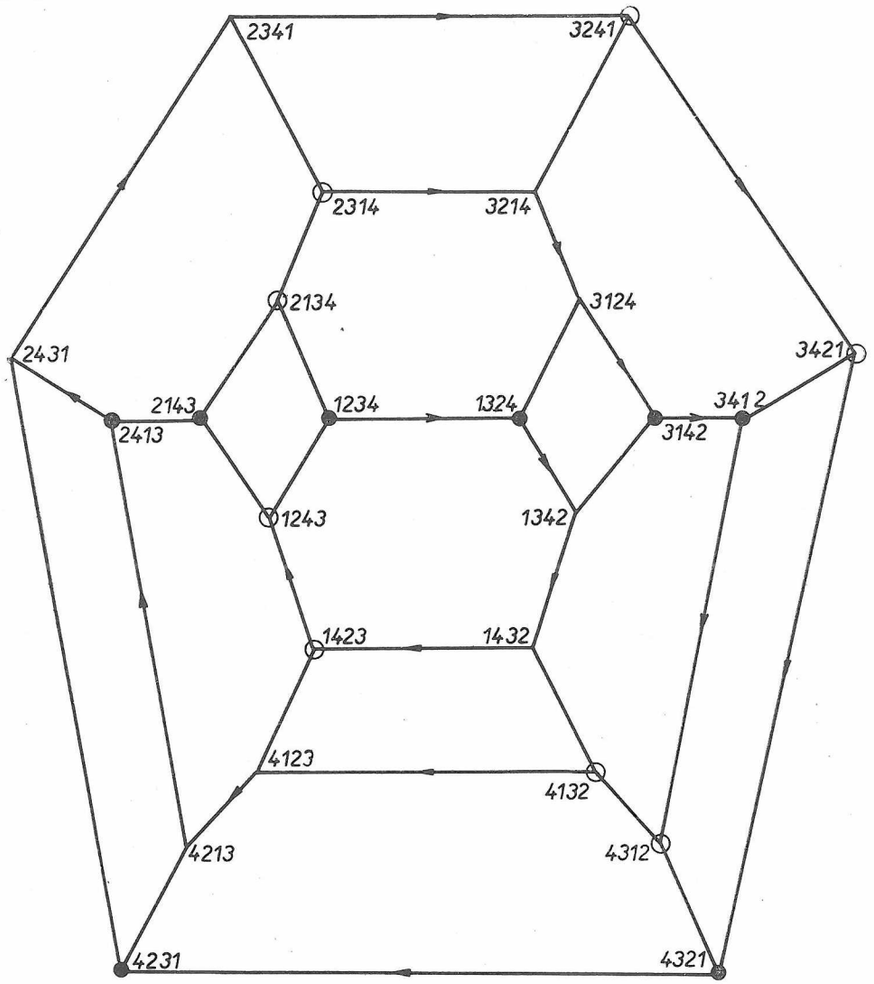
(11) ... *on échange un substantif en verbe ... l'adjectif en adverbe.*
(1) *Le texte est à la langue ce que le procès est au système.*
(6) ... *Ce n'est pas le noir, c'est le rose, ce n'est pas le crime, c'est la sucrerie.*
(7) *Kierkegaard a découvert l'angoisse, Marx la faim,* etc ...

5. Le polyèdre des inversions

Si l'on désire donner une image plus „tangible" de la liste des 24 permutations en rapprochant les propositions qui ne diffèrent que par une inversion de termes voisins, on dessinera le réseau ci-contre, réseau qui peut être considéré comme la projection dans le plan d'une figure de l'espace plus harmonieuse, un polyèdre semi-régulier dont les facettes sont des pentagones réguliers et des carrés (voir fig.).

Sur ce réseau, les sommets associés aux propositions ayant sensiblement le même sens que 1234, à savoir: 1324, 2143, 2413, 3142, 3412, 4231 et 4321, ont été renforcés.

En suivant les flèches à partir de 1234 ou a un exemple de génération successive des 24 propositions en n'inversant chaque fois que deux occurrences voisines, et cela sans produire deux fois une même proposition. Notons que certains chemins s'engagent mal et mènent à des impasses, on se trouve bloqué avant d'avoir engendré toutes les propositions.[2]



[2] A titre de digression généralisante on peut faire le rapprochement avec le problème classique du linguiste qui visite diverses bibliothèques situées dans diverses villes, mais qui se trouve aux

Remarquons également qu'aucun chemin ne permet de produire les 8 propositions comparable à 1234 sans donner naissance en plus à des propositions parasites absurdes du point de vue de la théorie de Tesnière.[3]

6. La proportion linguistique

Si l'on relit la liste des permutations avec un peu d'attention, on s'aperçoit que les huit propositions retenues sont proches quant au sens parce qu'elles maintiennent toutes un certain équilibre, les quatre termes, disons, une certaine proportion que nous noterons, par analogie avec les proportions arithmétiques,

$$(adverbe\ /\ verbe) \sim (adjectif\ /\ substantif)$$

— les parenthèses et le signe „∼“ (équivalent à) nous rappelant qu'il ne s'agit là que d'une ressemblance[4] — ou plus brièvement, avec les mêmes remarques que tout à l'heure:

$$(1/2) \sim (3/4)$$

---

prises avec cette double difficulté: d'une part, il n'existe pas de moyens de locomotion entre toutes les bibliothèques, d'autre part, il ne veut pas passer deux fois par la même ville. Pourquoi ne pas signaler au passage que de tels problèmes peuvent être traités mathématiquement et prononcer à ce propos le nom d'Hamilton?

[3] Insistons sur l'illusion qu'il y aurait à croire qu'un schéma, comme celui-ci, est «assimilé» sans entraînement approprié. Pour s'en convaincre essayer de faire découvrir le schéma plus simple correspondant aux 6 permutations de: *Jean ronfle bruyamment* ou bien *Jean boude Marie* et l'on verra les difficultés qui ne manqueront pas de se poser.

Signalons au passage qu'on peut imaginer de nombreux exercices en fusionnant certains sommets *il ment comme il respire* (1 = 3), de même dans comme *on fait son lit on se couche*; *tu seras forgeron comme ton père*; *tel père tel fils ... tel petit-fils*; *oeil pour oeil, dent pour dent*; *ne fais pas à autrui ce que tu ne voudrais pas qu'on te fît à toi même* interprété:

([*toi*]/ *autrui*) (on / toi). Mais ceci dépasse le cadre de notre exposé.

[4] En effet, la notation de rapport: (numérateur / dénominateur) désigne ici simplement un couple ordonné d'objets linguistiques que le locuteur sent comme liés d'une certaine façon, la nature précise de la liaison n'étant pas prise en considération. Elle peut être, par exemple, relativement simple, interne à une proposition, morphologiquement assez homogène, comme la relation entre prédiqué et prédicat: *Les parents boivent, les enfants trinquent*; ou entre deux prédiqués: cf. 3, 7, 11; ou au contraire complexe, dépassant le cadre de la proposition, non homogène comme dans:

curieux réflexe qui dans l'ouest de la France *engage souvent l'argent à se caser à gauche comme pour s'excuser en quelque sorte d'exister, tandis que la pauvreté vote volontiers à droite parce que ça fait honorable.*

(H. Bazin «Qui j'ose aimer»), relation 6-aire où l'on peut, semble-t-il, dégager 3 rapports enchevêtrés:

(*argent / pauvreté*) ∼ (*gauche / droite*) ∼ (*pour s'excussr / ça fait honorable*) l'inversion du 2ème provoquant un effet de surprise recherché.

Ce fait est patent pour la proposition 3412, dont l'insertion dans la citation entière de Tesnière, paraît d'autant plus acceptable que l'on permute également dans la 2ème partie:

verbe et *substantif*, *adverbe* et *adjectif*.

D'où le pastiche:

„*L'adjectif est au substantif ce que l'adverbe est au verbe. Il en résulte que, quand on change un verbe en substantif, il faut parallèlement changer l'adverbe en adjectif*".

La proposition 4231 établit des rapprochements un peu différents: néanmoins elle reste tout à fait acceptable, et même apparemment on n'est pas obligé de modifier la fin de la citation.

Le substantif est au verbe ce que l'adjectif est à l'adverbe. Il en résulte que, quand on change le verbe en substantif, il faut parallèlement changer l'adverbe en adjectif.

Ainsi l'ordre 3412 de la première partie s'accomoderait peut-être mieux de l'ordre 2413 dans la seconde, tandis que l'ordre 4231 de la première ne semble pas exiger impérativement une modification de l'ordre initial 4231 dans la deuxième.

Mais laissons de côté, pour l'instant, le subtil problème de grammaire interphrastique établissant un lien entre l'ordre des termes de deux propositions qui, d'une certaine façon, se répondent.[5]

7. Deux transformations

Il y a au moins deux façons de concevoir la transformation faisant passer d'une permutation à une autre.

R) *Par réécriture* (ou substitution des termes). Un terme est remplacé par un autre, par ex. pour 1324 et 3412:

l'adverbe est à l'adjectif ce que le verbe est au substantif

l'adjectif est au substantif ce que l'adverbe est au verbe

[5] Il est facile d'imaginer de nombreux exemples de cette sorte. «*Jacques fait ses devoirs de français et de mathématique. S'il s'en sort à peu près avec le premier, le deuxième, par contre, lui donne du fil à retordre*». «*M. Dupont préfère le bourgogne au bordeaux. Serait-ce parce que l'un lui rappelle son pays natal et l'autre celui de sa belle-mère*»? etc. ...

Avec la notation sténographique cela donne:

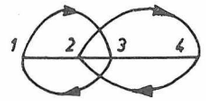
N°	des places	1	2	3	4	
N°	des termes	1	3	2	4	proposition initiale
		↓	↓	↓	↓	
N°	des termes	3	4	1	2	proposition transformée
N°	des places	1	2	3	4	

Les flèches indiquent la réécriture à effectuer.

Le terme	1	(*adverbe*)	se réécrit	3	(*adjectif*)	ou 1 → 3
Le terme	2	(*verbe*)	se réécrit	1	(*adverbe*)	ou 2 → 1
Le terme	3	(*adjectif*)	se réécrit	4	(*substantif*)	ou 3 → 4
Le terme	4	(*substantif*)	se réécrit	2	(*verbe*)	ou 4 → 2

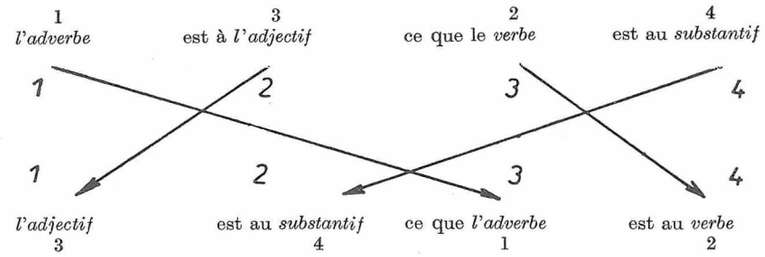
D'où la notation abrégée de la *transformation réécriture*.

R = (3142) en système sémiotique symbolique ou:



en un système sémiotique iconique.

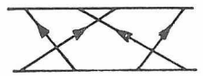
P) Par permutation des trous. Dans la proposition à 4 trous, obtenue en effaçant les 4 termes, on numérote les trous. Le terme qui se trouvait dans un certain trou vient occuper un autre trou.[6]



[6] Comparez au «jeu des 4 coins» avec à la place des coins, des arbres: un platane, un chêne, un marronnier, un bouleau; prennent part au jeu; André, Pierre, Jacques, Claude et Bernard. On peut décrire les diverses phases du jeu en ne parlant que des personnes qui se déplacent (le R-ien) ou des places occupées, c'est-a-dire des arbres (le P-ien).

En notation sténographique:

N°	des termes	1	3	2	4	
N°	des trous	1	2	3	4	proposition initiale

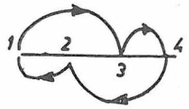


N°	des trous	1	2	3	4	proposition transformée
N°	des termes	3	4	1	2	

Les flèches indiquent le nouvel emplacement.

Le terme occupant le trou 1 vient dans le trou 3
Le terme occupant le trou 2 vient dans le trou 1
Le terme occupant le trou 3 vient dans le trou 4
Le terme occupant le trou 4 vient dans le trou 2

D'où la notation abrégée de la *transformation permutation.*



P = (3142) en un système sémiotique symbolique; ou en un système sémiotique iconique.

Voici les tableaux complets des deux transformations.

Réécriture des termes

ijkl \ i'j'k'l'	1234	1324	2143	2413	3142	3412	4231	4321
1234	(1234)	(1324)	(2143)	(2413)	(3142)	(3412)	(4231)	(4321)
1324	(1324)	(1234)	(2413)	(2143)	(3412)	(3142)	(4321)	(4231)
2143	(2143)	(3142)	(1234)	(4231)	(1324)	(4321)	(3413)	(2413)
2413	(3142)	(2143)	(4231)	(1234)	(4321)	(1324)	(3412)	(2413)
3142	(2413)	(3412)	(1324)	(4321)	(1234)	(4231)	(2143)	(3142)
3412	(3412)	(2413)	(4321)	(1324)	(4231)	(1234)	(3142)	(2143)
4231	(4231)	(4321)	(3142)	(3412)	(2143)	(2413)	(1234)	(1324)
4321	(4321)	(4231)	(3412)	(3142)	(2413)	(2143)	(1324)	(1234)

Les termes ijkl se réécrivent respectivement i'j'k'l' lorsqu'on applique l'instruction située à la croisée de la ligne et de la colonne correspondante.

Permutation des trous

ijkl \ i'j'k'l'	1234	1324	2143	2413	3142	3412	4231	4321
1234	(1234)	(1324)	(2143)	(3142)	(2413)	(3412)	(4231)	(4321)
1324	(1324)	(1234)	(2413)	(3412)	(2143)	(3142)	(4321)	(4231)
2143	(2143)	(3142)	(1234)	(1324)	(4231)	(4321)	(2413)	(3412)
2413	(2413)	(3412)	(1324)	(1234)	(4321)	(4231)	(2143)	(3142)
3142	(3142)	(2143)	(4231)	(4321)	(1234)	(1324)	(3412)	(2413)
3412	(3412)	(2413)	(4321)	(4231)	(1324)	(1234)	(2142)	(2143)
4231	(4231)	(4321)	(3142)	(2143)	(3412)	(2413)	(1234)	(1234)
4321	(4321)	(4231)	(3412)	(2413)	(3142)	(2143)	(1324)	(1234)

Les termes qu'occupaient les trous ijkl viennent se placer respectivement dans les trous i'j'k'l'.

8. Deux systèmes sémiotiques

On se trouve, en quelque sorte, en présence de deux „micro-métalangages globallement synonymiques" le „R-ien" et le „P-ien", utilisant les mêmes graphèmes 1, 2, 3, 4 avec des „symboles suprasegmentaux" {(,)} ces derniers jouant sur le „mot" entier. Cf. 1234 et (1234).

Le R-ien et le P-ien coïncident, pour tout ce qui concerne les mots non parenthésés; ils diffèrent partiellement pour les mots parenthésés. Pour ces derniers le catalogue des signifiants est le même, mais les signifiés associés sont distincts au point qu'il n'est pas possible d'établir de dictionnaire (correspondance biunivoque) entre les deux micrométalangages, car chacun d'eux structure la „réalité" à sa façon. Les signifiants identiques ne correspondent pas toujours aux mêmes couples de proposition associés. Ex.

Dans le R-ien et dans le P-ien l'instruction (2143) peut être utilisée dans huit phrases, respectivement colonne 1 et 2 ci-dessous:

2143 = (2143) 1234 ↔ 2143 = (2143) 1234
2413 = (2143) 1324 3142 = (2143) 1324

1234 = (2143) 2143 ↔ 1234 = (2143) 2143
1324 = (2143) 2413 4231 = (2143) 2413
4231 = (2143) 3142 1324 = (2143) 3142
4321 = (2143) 3412 ↔ 4321 = (2143) 3412
3142 = (2143) 4231 2413 = (2143) 4231
3412 = (2143) 4321 ↔ 3412 = (2143) 4321

Ainsi les signifiés, de même signifiant (2143), dans le R-ien et le P-ien, expriment dans huit contextes associés l'analogie ressentie par les „native speakers de la R-anie et de la P-anie" entre les ordres de termes de certaines propositions. Cependant, d'une obédience linguistique à une autre, cette analogie ne s'applique pas nécessairement aux mêmes couples de propositions. Dans 4 cas il y a coïncidence des deux points de vue. Dans 4 autres il n'en est rien. Bien plus, une assertion métalinguistique telle que:

2413 = (2143) 1324

qui est, pour une certaine syntaxe rudimentaire, „grammaticalement acceptable" dans les 2 langues, est attestée dans R-ien comme une phrase *normale*, alors qu'elle peut être considérée comme *anomale* de par son sens dans le P-ien. La raison, apparemment, doit en être recherchée dans une „structure profonde" ou, si l'on préfère, dans une syntaxe plus fine qui tient compte d'un certain contenu (ordre des termes) des propositions-objets.

Une seule instruction conserve le „même sens" dans tous les cas en R-ien et en P-ien c'est: „ne rien faire" (laisser les termes „comme ils sont" — R-ien, „où ils sont" — P-ien), instruction que nous nommerons *identité*:

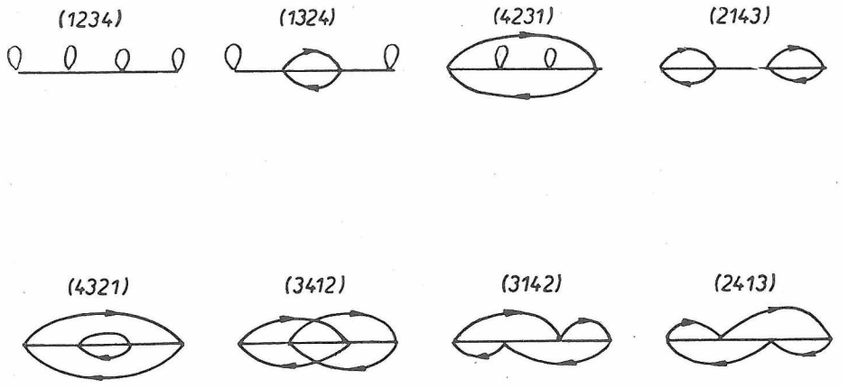
I = (1234) ou

↻	↻	↻	↻
i	j	k	l
i′	j′	k′	l′

9. Démarcation iconique

La figuration métalinguistique iconique proposée utilise les mêmes huit images en R-ien comme en P-ien et ne fait que démarquer les signifiants symboliques respectifs; voici ces 8 images accompagnées des 8 signifiants du R-ien ou du P-ien indifféremment:

Cependant le sens des images se lira de façon spécifique dans chaque idiome. Une même image ne voudra pas dire la même chose selon l'„obédience linguistique" dont le codeur-décodeur est tributaire. Un même signifiant ne sera pas associé au même signifié. Voir schéma page suivante.



10. Instructions et prédicats. Symétrie et réciprocité

Les versions iconiques des métalangages font mieux ressortir les deux types de „mots", que nous pouvons interpréter dans les deux systèmes soit comme des *instructions* permettant d'obtenir une proposition, à partir d'une autre en conservant l'harmonie de la proportion, soit comme des prédicats à deux places précisant qu'il existe une relation sémantique particulière entre certaines propositions (ce qui est formel au niveau de la langue-objet — ordre des termes — devient sémantique au niveau du métalangage).

En effet, les 6 premiers prédicats sont *symétriques*. On le voit sur l'image et on le constate sur les tableaux.

Ex., dans le R-ien on a:

(4231) 3142 = 3412 et (4231) 3412 = 3142

ce qu'il est loisible de noter à la manière des logiciens:

(4231) [3142, 3412] = (4231) [3412, 3142]

autrement dit ces prédicats se conduisent comme le verbe réfléchi *se bouder* dans:
Pierre et Marie se boudent = Marie et Pierre se boudent
(*se boudent*) [*Pierre, Marie*] = (*se boudent*) [*Marie, Pierre*]

Par contre les deux derniers prédicats ne sont pas symétriques.

Ex., dans le R-ien on a:

(2413) 1324 = 2143 ≠ (2413) 2143 = 4231 [7]

[7] On a utilisé le signe ≠ au lieu de ≠ car il s'agit, somme toute, d'un méta-métalangage, différent du métalangage.

cependant les images nous incitent à examiner:

$$(3142)\ 2143 = 1324 \not\equiv (3142)\ 4231 = 2143$$

On s'aperçoit alors que:

(2413) [1324, 2143] tandis que (3412) [2143, 1324]

et vice versa. C'est-à-dire ces deux prédicats se comportent comme le verbe transitif à la voix active ou passive.

disputer et *être disputé(e) par* dans:

Pierre dispute Marie et *Marie est disputée par Pierre*, (*dispute*) [*Pierre, Marie*] et (*est disputée par*) [*Marie, Pierre*], ce qui est bien différent de:

Marie dispute Pierre et *Pierre est disputé par Marie*

Si l'un des prédicats est considéré comme étant „à la voix active", l'autre sera „à la voix passive" et réciproquement. Nous dirons que les prédicats, ou les instructions correspondantes, sont *réciproques* l'une de l'autre.

On remarquera que les tableaux des instructions en R-ien et en P-ien coïncident lorsqu'on passe d'une instruction symétrique à une autre symétrique ou bien d'une instruction non-symétrique à une non-symétrique. Les instructions sont inversées lorsqu'on change de catégories d'instructions.

Ainsi, dirons nous, il existe un „sous-glossaire" international, d'instructions commun aux deux métalangages.

Ex. La comparaison des deux quadruplets 2143 et 3412 s'établit à l'aide du même prédicat (4321) en R-ien comme en P-ien.

Pour les deux instructions internationales (1234), (4321) le glossaire bilingue devient inutile, elles veulent toujours dire la même chose dans ces deux idiomes.

Les 4 instructions (1324) (2143) (3412) et (4231) ne sont „internationales" que pour certains contextes, à savoir pour ceux qui ne sont formés que des mots 1234, 1324, 2143, 3412, 4231 et 4321 ou que des mots 2413 et 3142. Elles sont réparties de façon différente pour des contextes „mixtes".

11. Bilan des notations utilisées

Faisons le bilan des symboles métalinguistiques utilisés:[8]

Une séquence de quatre graphèmes (chiffres) est susceptible de désigner:

a) Une proposition contenant quatre termes rangés dans un certain ordre, chacun d'eux étant „nommé" à l'aide d'un numéro.

Ex. 1342

b) Une proportion linguistique (structure de la proposition considérée à un certain point de vue)

Ex. $(1/3) \sim (4/2)$

c) Une instruction permettant de passer d'une phrase à une autre, exprimée dans le métalangage R ou dans le métalangage P.

Ex. $(1324)_R \not\equiv (1324)_P$

Cette même notation pouvant être conservée pour désigner des prédicats à deux places.

Sur les 24 quadruplets, huit seulement admettent comme invariant la proportion linguistique.

Si l'on pose que l',,harmonie proportionnelle" entre les signifiés de la langue objet appartient au contenu (Hjelmslev) de la langue objet, l'expression

$$(1/3 \sim (4/2)$$

décrit la forme du contenu de la langue objet.

L'ordre des termes de la proposition:

1342

ne concerne que les signifiants et relève de la syntaxe de la langue objet dans ce qu'elle a de formel. On peut dire que 1234 décrit la forme de l'expression de la langue objet.

Pour les métalangages ces deux formes constituent des contenus, identiques pour le R-ien et le P-ien.

Cependant le R-ien et le P-ien diffèrent par la façon de décrire les rapports analogiques de la structure des signifiants de la langue objet. Lorsqu'il s'agira d'écrire des phrases métalinguistiques en posant comme invariant l'harmonie proportionnelle les assertions en R-ien et P-ien du point de vue formel ne vont différer que dans certains cas, du point de vue contenu elles voudront toujours dire des choses différentes.

---

[8] L'expérience pédagogique montre que les étudiants ont beaucoup de peine à se faire au symbolisme de ces deux notations. Pour être suivi il importe de couper l'exposé théorique par de nombreux exercices de pure technique.

On pourra opérer avec des séquences très courtes, puis plus longues comme les célèbres variations du Bourgeois-Gentilhomme sur:

Belle Marquise vos beaux yeux me font mourir d'amour, et rechercher quelles sont celles qui donnent des résultats relativement acceptables pour un francophone.

12. Analyse fine du corpus[9]

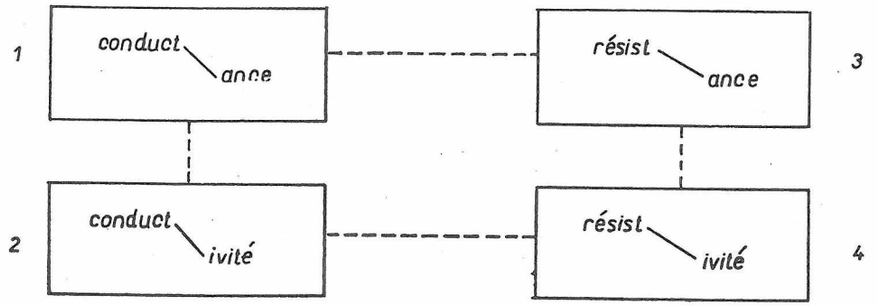
Reprenons le corpus donné au début.

On constate que certaines citations se comportent comme la citation (11) de Tesnière:

(2) (*texte / langue*) ~ (*procès / système*)

(3) (*conductance / conductivité*) ~ (*résistance / résistivité*)

En effet c'est le mode de comparaison selon la procédure des carrés de Greenberg:



Dans (4) les quatre termes qui interviennent sont de nature assez homogène, il s'agit toujours de „distractions":

1 = *ces passe-temps courants et somme toute assez anodins,*
2 = *nos jeux à nous,*
3 = *aimables jeux de société,*
4 = *jeux sanglants du cirque.*

Il n'en est pas toujours de même, et la nature du rapport marqué par la barre de fraction symbolique peut intervenir de façon plus ou moins coercitive.

L'exemple (6) conduit aux 8 proportions:

I) *Le noir est au rose comme le crime est à la sucrerie*
II) *Le crime est à la sucrerie comme le noir est au rose*
III) *Le rose est au noir comme la sucrerie est au crime*

[9] Ce paragraphe constitue une ébauche de typologie des «proportions linguistiques» et en particulier des métaphores en forme de proportion. C'est donc un point de départ pour toute une nouvelle étude qui déborde le cadre strictement pédagogique que nous nous sommes assigné.

IV) *La sucrerie est au crime comme le noir est au rose*
V) *Le noir est au crime comme le rose est à la sucrerie*
VI) *La sucrerie est au rose comme le crime est au noir*
VII) *Le rose est à la sucrerie comme le noir est au crime*
VIII) *Le crime est au noir comme la sucrerie est au rose*

Dans les 4 premières propositions les termes („numérateur" et „dénominateur") de chaque rapport subsistent, seule leur position à l'intérieur des rapports ou l'emplacement des rapports varie. Autrement dit la comparaison s'effectue sur les mêmes objets, entre les couleurs d'une part et *crime* et *sucrerie* d'autre part.
Il n'en est pas de même pour les 4 derniers énoncés.

Néanmoins toutes ces permutations semblent faire état en gros, de la même image stylistique.

La même remarque peut être faite à propos de l'exergue, la modification de la nature des rapports est tolérable.

(*course à pieds / autres sports*) ~ (*géométrie / autres sports*)
(*course à pieds / géométrie*) ~ (*autres sports / autres sciences*)

Par contre dans (9) nous tolérons, par ex:

(*les gros plan / cinématographe*) ~ (*imprimerie / italique*)

alors que:

(*imprimerie / cinématographe*) ~ (*italique / gros plan*)

nous laisse plutôt „rêveurs".

Dans (5) le rapprochement de *carnet du jour et* de *salon* n'est pas fortuit. Il y a là un effet de style voulu de sorte que les proportions ne sont pas toutes rigoureusement équivalentes.

L'exemple (1) pose des problèmes pour ce qui est d'énoncer des propositions à ordre des termes modifiée.

Comment dire:

(*avare / père*) ~ (*prodigue / fils*)?

On pourrait imaginer des restructurations morpho-syntaxiques:

L'avarice va au père comme la prodigalité au fils et dans ces conditions les 8 sentences proverbiales sont concevables.

(8) demande une remise en ordre préalable des concepts de la proportion:

(*aller / loin*) ~ (*monture / ménager*).

Il faut pas mal d'imagination pour bâtir les huit sentences proverbiales correspondantes!

Examinons encore cet extrait d'Omar Khayyam, dans la traduction de Franz Toussaint:

Une telle odeur de vin émanera de ma tombe, que les passants en seront enivrés.

Une telle sérénité entourera ma tombe, que les amants ne pourront s'en éloigner.

Là aussi on sent l'harmonie d'une proportion:

$$\left(\frac{\textit{L'odeur de vin émanant de ma tombe}}{\textit{les passants seront enivrés.}}\right) \sim \left(\frac{\textit{La sérénité entourera ma tombe}}{\textit{les amants ne pourront s'éloigner}}\right)$$

Cependant au niveau de la chaîne écrite ou parlée les permutations sont difficilement réalisables car l'expression même possède une structure poétique qu'il n'est pas sans danger de perturber.

13. Les rapports successifs multiples

Les exemples (7) et (10) font intervenir plus de 2 rapports.

$$\left(\frac{\textit{Kierkegaard}}{\textit{angoisse}}\right) \sim \left(\frac{\textit{Marx}}{\textit{faim}}\right) \sim \left(\frac{\textit{Freud}}{\textit{sexe}}\right) \sim \left(\frac{\textit{Einstein}}{\textit{relativité}}\right) \sim \left(\frac{\textit{Curie}}{\textit{radioactivité}}\right)$$

$$\left(\frac{\textit{Ce que se disent ses personnages [Nathalie Sarraute]}}{\textit{ce que se disent les héros de Joyce, de Virginia Woolf ou de Faulkner}}\right)$$

$$\sim \left(\frac{\textit{ce que se disent les mouvements surpris par la physique moléculaire}}{\textit{les figures de la physique macroscopique}}\right)$$

$$\sim \left(\frac{\textit{ce que l'eau souterraine}}{\textit{les cristallisations de la surface}}\right)$$

Notons (7) en sténo:

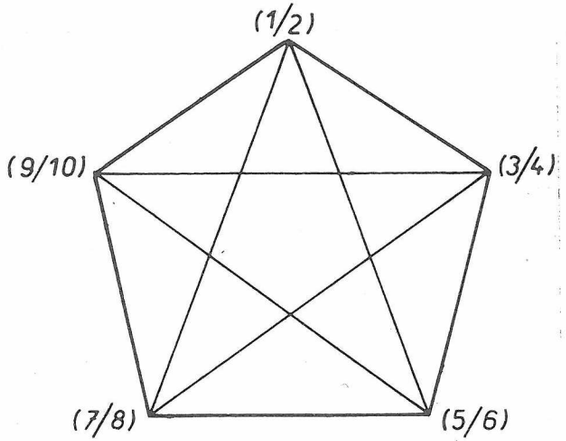
$$(1/2) \sim (3/4) \sim (5/6) \sim (7/8) \sim (9/10)$$

ce qui donne $5 \times 4 = 20$ proportions. A propos de chacune d'elle on peut faire les remarques précédentes.

Cependant toutes les permutations des termes ne sont plus permises pour des raisons formelles: il faut tenir compte de touts les rapports à la fois.

Les sommets du pentagone complet, ci-dessus, figurant les rapports, chaque arête étant dédoublée en 2 arcs opposés, chaque arc correspond à une proportion. En conservant les rapports tels qu'ils sont exprimés dans la citation, l'ordre d'énonciation des rapports ayant relativement peu d'importance, il y aura autant de pastiches possibles que de trajets passant une fois et une seule par tous les sommets, soit $5! = 120$, y compris la citation originale.

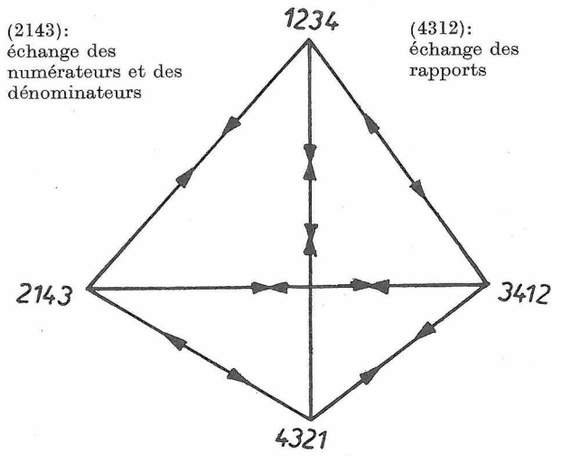
Si on inverse un rapport, on doit inverser également tous les autres, ce qui donne encore 120 autres pastiches nécessitant apparemment une restructuration plus ou moins poussée de la phrase. Il est peut-être gênant de commencer par dire: *cinq hommes* ... et puis de continuer *l'angoisse a été découverte par Kierkegaard*. On pourrait ajouter par ex.: *ainsi* l'angoisse ... et appuyer sur les noms propres.



14. Proximité sémantique

L'analyse fine du corpus nous a conduit à partager les 8 permutations possibles d'une proportion en deux classes.

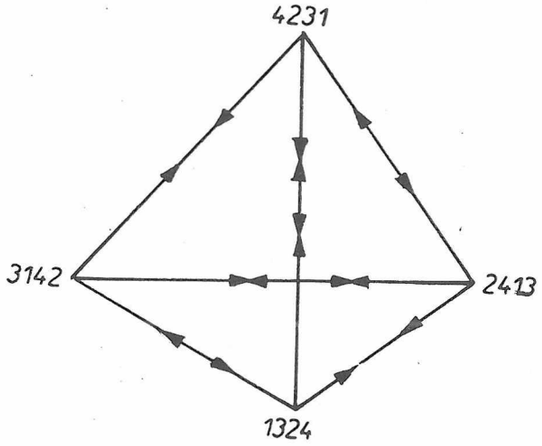
Dans la première les termes intervenant dans un même rapport ne changent pas, seuls peuvent être échangés les „numérateurs" avec les „dénominateurs" correspondants ou bien les rapports entre eux. Nous dirons que nous avons 4 propositions sémantiquement à peu près équivoisines, c'est pourquoi nous les placerons, de façon symbolique, aux quatres sommets d'un tétraèdre régulier.



(4321): double échange

Les sous-tableaux des réécritures et des permutations sont identiques.

R, P	1234	2143	3412	4321
1234	(1234)	(2143)	(3412)	(4321)
2143	(2143)	(1234)	(4321)	(3412)
3412	(3412)	(4321)	(1234)	(2143)
4321	(4321)	(3412)	(2143)	(1234)



Les 4 mêmes séquences de graphèmes interviennent et comme ordres de termes dans les phrases et comme 4 instructions de transformations.

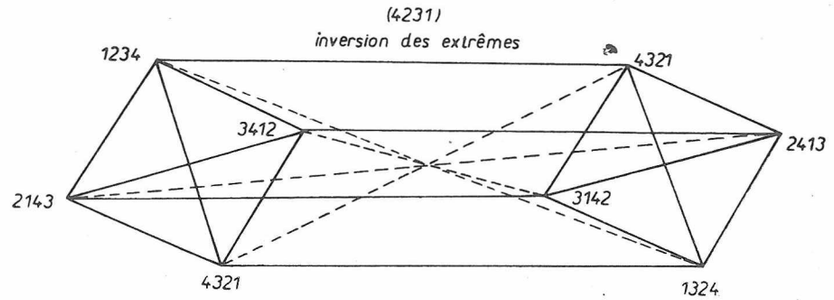
Des remarques analogues peuvent être faites pour les quatre permutations restantes.

Les quatre nouvelles propositions, significativement légèrement décalées par rapport aux quatre précédentes, sont également, pour leur contenu, à peu près équivoisines entre elles.[10] Quand on passe d'une quelconque des quatre premières propositions à une quelconque des quatre autres il y a toujours „brisure" de deux rapports.

Le symbolisme géométrique de l'équidistance ne peut plus être poursuivi de façon concrète lorsqu'on envisage à la fois les huit propositions. Cependant il est loisible d'imaginer des images plus ou moins suggestives. Par exemple, les deux tétraèdres étant placés de manière à se déduire l'un de l'autre par translation, on peut faire correspondre ainsi les propositions ne différant que par inversion des termes extrêmes; les propositions différant par inversion des termes moyens étant symétriques deux à deux par rapport à un centre. Il est évident que ces images n'ont pas de portée théorique mais servent uniquement à soutenir l'attention.

[10] De façon plus fine, on peut établir un ordre partiel de proximité sémantique à l'intérieur de l'ensemble des 4 permutations, ce que nous avons esquissé à l'aide des 3 types de doubles flèches ⟨—⟩, ⟩—⟨ et ×—× et qui répond à une certaine réalité sémantico-stylistique.

Les instructions sont figurées par des vecteurs joignant les sommets correspondants.



Les 4 instructions non marquées sur la figure sont du type non-symétrique. Ex. pour passer de 1234 à 2413 ou à 3142.

15. Modèles géométriques. Modèle des rotations de 45°

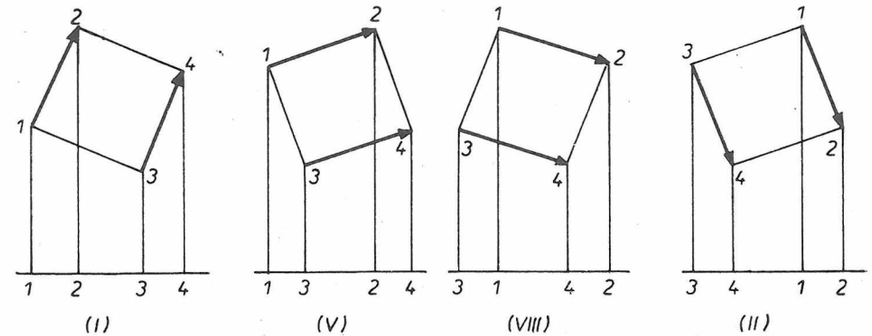
Il n'est pas difficile d'imaginer des modèles géométriques divers concrétisant les relations entre les huit instructions étudiées.

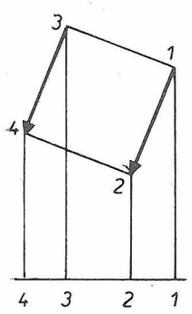
Les quatre termes étant associés aux extrémités de deux vecteurs équipollents formant un carré, par rotations successives de 45° dans le sens des aiguilles d'une montre, et en projetant les sommets sur un axe on obtient les 8 quadruplets (fig.).

Modèles des symétries-rotations

Toujours sur le même carré on peut effectuer soit des symétries par rapport aux médiatrices des côtés ou, par rapport aux diagonales, soit des rotations de +90°, —90° ou de 180° par rapport au centre, on obtient ainsi des carrés transformés. En lisant les sommets dans l'ordre où on les a marqués sur le carré initial, on obtient les 8 quadruplets (fig.).

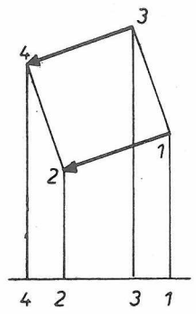
En comparant les modèles aux deux métalangages étudiés on se rend compte qu'ils démarquent le *P*-ien. Ainsi en reprenant les huit propositions traitées ci-dessus successivement à l'aide de l'instruction (2413) lue en *R*-ien et *P*-ien, on a avec le modèle des symétries-rotations (symétries par rapport à l'horizontale).





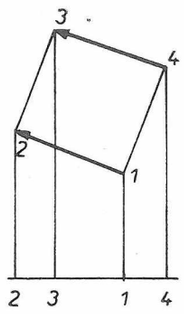
(IV)

rot. +180°



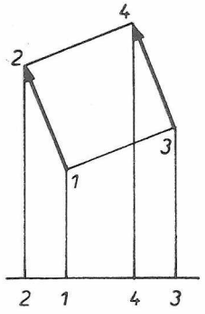
(VI)

rot. +135°
ou —225°



(VII)

rot. +90°
ou —270°



(III)

rot. +45°
ou —315°

Ces angles étant donnés à un multiple entier de 360° près. (ainsi —45° c'est aussi bien +315°, —405°, 675°, ...)

2 / 1	donne	4 / 3	1 / 2	3 ou 2413 / 4	1 / 3	donne	2 / 4	3 / 1	4 ou 1324 / 2
3 / 1	donne	4 / 2	1 / 3	2 ou 3142 / 4	4 / 3	donne	2 / 1	3 / 4	1 ou 4321 / 2
1 / 2	donne	3 / 4	2 / 1	4 ou 1234 / 3	2 / 4	donne	1 / 3	4 / 2	3 ou 2413 / 1
4 / 2	donne	3 / 1	2 / 4	1 ou 4231 / 3	3 / 4	donne	1 / 2	4 / 3	2 ou 3412 / 1

Pour nommer une transformation permettant de passer d'un ordre à un autre nous pouvons disposer de 3 langages formels:

1. Celui des permutations, des symboles.
2. Celui des symétries rotations laissant globalement invariant un carré.
3. Celui des rotations d'un multiple entier de 45°.

Entre ces trois langages on peut établir un dictionnaire:

I	(1, 2, 3, 4)	I	Rotation de 0° (à 360° près).
II	(3 4 1 2)	S_1	Rotation de —135°
III	(2 1 4 3)	S_2	Rotation de —90°
IV	(4 3 2 1)	R (180°)	Rotation de +180°
V	(1, 3, 2, 4)	S_3	Rotation de —45°
VI	(4, 2, 3, 1)	S_4	Rotation de +135°
VII	(2, 4, 1, 3)	R_+ (+90°)	Rotation de +90°
VIII	(3, 1, 4, 2)	R_- (—90°)	Rotation de —90°

16. Produit de transformation. Groupe

Que se passe-t-il si sur une proposition donnée on effectue plusieurs transformations successives?

Lorsqu'il s'agit de permutations, les deux modèles géométriques montrent que le résultat obtenu à l'aide de deux instructions revient à donner une seule instruction figurant dans la liste des instructions envisagées,[11] cela indépendamment du contexte.

D'une façon plus précise, on dira que l'ensemble des instructions-permutations conservant la proportion linguistique déterminent un *groupe*, au sens mathématique de ce terme, c'est-à-dire:

L'ensemble des instructions est structuré par une loi de composition associative, ici la composition de deux instructions.

Cet ensemble contient l'instruction identité: I.

A toute instruction on peut associer une instruction réciproque (cela peut être l'instruction elle-même) qui permet de revenir au point de départ.

Le produit de deux instructions est défini, c'est une instruction faisant partie de l'ensemble des instructions données.

Ci-dessous la table de composition exprimée dans le système sémiotique idéographique des Symétries-Rotations ainsi que dans le système sémiotique séquentiel initial.

Le carré encadré en trait gras contenant uniquement les symboles I, S_1, S_2 et R, constitue un sous-groupe du précédent et correspond aux instructions qui ne brisent aucun rapport (cf. § 14); le sous-groupe possède par conséquent une interprétation linguistique autonome.[12]

[11] Si T_1 et T_2 sont les deux transformations effectuées successivement sur l'objet x, l'usage est d'écrire $T_2(T_1(x)) = T_2T_1(x) = T(x)$ ou, en abrégé, $T_2T_1 = T$, T étant la transformation résultante.

[12] On pourrait à ce propos faire intervenir le groupe de toutes les permutations fournissant toutes les variantes de: *Alphonse, toujours, ronfle bruyamment* et montrer que le groupe associé aux

Tableau 1

1 ère transf. \ 2 ème transf.	I	S_1	S_2	R	S_3	S_4	R_+	R_-
I	I	S_1	S_2	R	S_3	S_4	R_+	R_-
S_1	S_1	I	R	S_2	R_-	R_+	S_4	S_3
S_2	S_2	R	I	S_1	R_+	R_-	S_3	S_4
R	R	S_2	S_1	I	S_4	S_3	R_-	R_+
S_3	S_3	R_+	R_-	S_4	I	R	S_1	S_2
S_4	S_4	R_-	R_+	S_3	R	I	S_2	S_1
R_+	R_+	S_3	S_4	R_-	S_2	S_1	R	I
R_-	R_-	S_4	S_3	R_+	S_1	S_2	I	R

Tableau 2

	1ère tran. \ 2 ème tran.	(1234)	(3412)	(2143)	(4321)	(1324)	(4231)	(4231)	(3142)
I_1	(1234)	(1234)	(3412)	(2143)	(4321)	(1324)	(4231)	(2413)	(3142)
S_1	(3412)	(3142)	(1234)	(4321)	(2143)	(3142)	(2413)	(4231)	(1324)
S_2	(2143)	(2143)	(4321)	(1243)	(3412)	(2413)	(3142)	(1324)	(4231)
R	(4321)	(4321)	(2143)	(3412)	(1234)	(4231)	(1324)	(3142)	(2413)
S_3	(1324)	(1324)	(2413)	(3142)	(4231)	(1234)	(4321)	(3412)	(2143)
S_4	(4231)	(4231)	(3142)	(2413)	(1324)	(4321)	(1234)	(2143)	(3412)
R_+	(2413)	(2413)	(1324)	(4231)	(3142)	(2143)	(3412)	(4321)	(1234)
R_-	(3142)	(3142)	(4231)	(1324)	(2143)	(3412)	(2143)	(1234)	(4321)

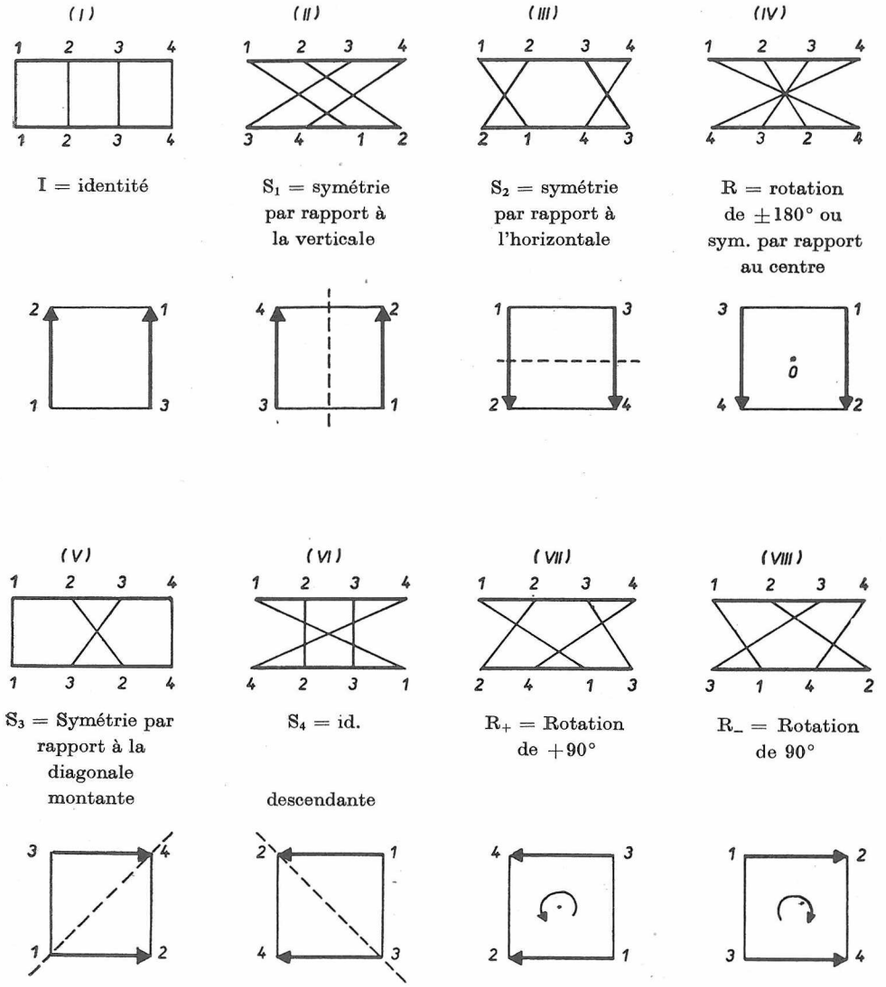
proportions en est un sousgroupe. Le problème peut être repris avec: *Belle Marquise vos beaux yeux me font mourir d'amour*. Quant au sousgroupe du groupe des 4 instructions on pourra le comparer à celui de la négation — interrogation classique.

[13] On comprend l'attachement qu'ont certains peuples pour l'écriture idéographique de leur langue.

[14] La Rotation de 180° est aussi une symétrie à centre.

Les notations géométriques idéographiques[13] beaucoup plus lisibles que les notations initiales, par séquence de graphèmes, mettent bien en évidence les deux sortes d'instructions: les instructions symétriques correspondent aux symétries[14] et les instructions réciproques aux deux rotations de 45° de sens contraire. De plus la composition de plusieurs instructions est immédiate dans bien des cas, ce qui n'est pas en notation séquentielle.

Dans le modèle des rotations, le sous-groupe est constitué par les rotations de 0°, — 135°, — 315°, — 180°.



On peut se demander si on a le droit de parler de groupe à propos des instructions de réécriture.

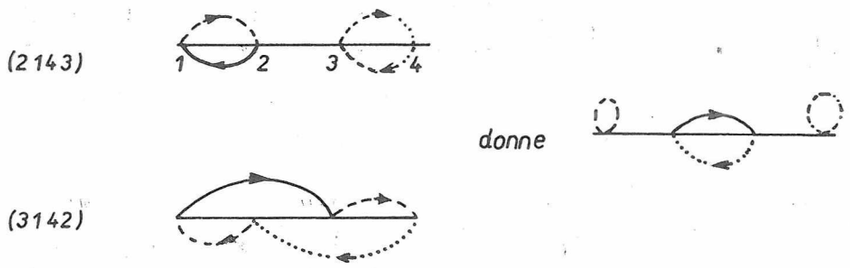
Il faut d'abord que la notion de produit de deux réécritures soit une nouvelle réécriture toujours définie et indépendante des objets sur lesquels on opère.

Faisons, à titre d'ex., un essai sur les deux instructions de réécriture (2143) et (3142):

(2143) 1234 = 2143	(3142) 2143 = 1324	(1324) 1234 = 1324
(2143) 1324 = 2413	(3142) 2413 = 1234	(1324) 1324 = 1234
(2143) 2143 = 1234	(3142) 1234 = 3142	(1324) 2143 = 3142
(2143) 2413 = 1324	(3142) 1324 = 3412	(1324) 2413 = 3412
(2143) 3142 = 4231	(3142) 4231 = 2143	(1324) 3142 = 2143
(2143) 3412 = 4321	(3142) 4321 = 2413	(1324) 3412 = 2413
(2143) 4231 = 3142	(3142) 3142 = 4312	(1324) 4231 = 4321
(2143) 4321 = 3412	(3142) 3412 = 4231	(1324) 4321 = 4231

On trouve chaque fois le même résultat: (1324). Il n'est pas impossible de faire tous les essais, mais c'est, pour le moins, long.

Or, on se souvient des schémas communs démarquant les signifiants du R-ien et du P-ien:



La succession des deux opérations, quels que soient les objets sur lesquels on opère, conduit aux mêmes dessins dans les deux systèmes sémiotiques iconiques, ou, si l'on préfère, la „syntaxe" demeure la même dans les deux langages.

Par conséquent, si l'un a une syntaxe de groupe, l'autre également.

Contre-exemples

Il importe de se rendre compte que n'importe quel ensemble structuré par une règle de réécriture ne constitue pas un groupe.

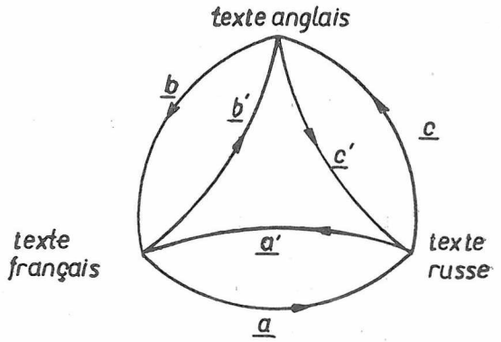
Ainsi une réécriture consistant à remplacer deux symboles par un seul peut fort bien dépendre du contexte.

Soit la proposition:

Il mange les parties charnues et crache les parties dures; le couple *parties dures* est susceptible d'être remplacé par *pépin, noyau, arête, os* selon ce que *il* mange. S'il est question de *pêches*, d'*abricots*, de *prunes*, d'*olives* ... ça ne peut être que *noyau* à l'exclusion des autres; pour une *pomme* on ne dira jamais (sauf pour un effet voulu) *les noyaux, les os* ou *les arêtes* bien que l'auditeur arrive à comprendre quand même et à restituer le terme approprié.

Contre-exemple. Si l'on envisage l'ensemble des instructions suivantes.

a = traduire un texte du français en russe
a' = traduire un texte du russe en français
b = traduire un texte du français en anglais
b' = traduire un texte de l'anglais en français
c = traduire un texte du russe en anglais
c' = traduire un texte de l'anglais en russe



La suite des opérations *ca* a un sens, elle veut dire qu'on traduit d'abord le texte du français en russe, puis du russe en anglais, c'est-à-dire, en définitive du français en anglais. Par contre *ac* est absurde quel que soit le texte du départ.

Ainsi l'ensemble a, a', b, b', c, c' muni de la loi de composition indiquée n'est certainement pas un groupe, car cette loi n'est pas toujours définie.

17. La fausse proportion

Soit la proposition à 4 termes

$$\underset{1}{\underline{Pierre}}\ \text{et}\ \underset{2}{\underline{Paul}}\ s'aiment\ \text{comme}\ \underset{3}{\underline{chien}}\ \text{et}\ \underset{4}{\underline{chat}}$$

qui fait penser à une proportion. En fait, il n'en est rien. Sont sensiblement équivalentes les propositions:

1234 1243 2134 2143

A la rigueur on peut admettre l'équivalence avec les précédentes:

3412 3421 4312 4321

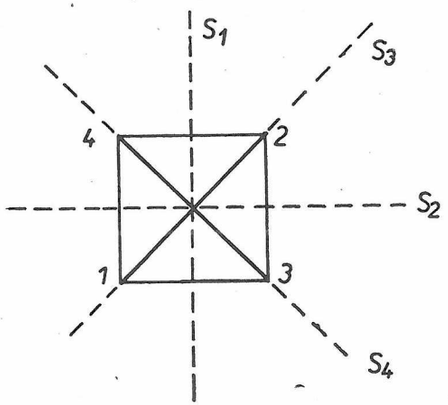
Constituent un contre-sens par rapport à la phrase donnée (permutation des moyens ou des extrêmes):

4231 3241 4132 3142

2413 1423 2314 1324

ainsi que les huit autres.

On pourra montrer que là encore on a affaire à un groupe, le modèle des symétries-rotations simule parfaitement ces propriétés à condition de numéroter en croix les sommets du carré:



La première série des quatres permutations correspond au sous-groupe I, S_3, S_4, R et la deuxième série se déduit de la première par S_1, R_-, R_+ et S_4 et correspond au même sous-groupe que la première.

Le schéma géométrique montre la parenté profonde qui existe entre la proportion et la fausse proportion, on passe de l'une à l'autre par une permutation, illicite dans les deux systèmes (1432), plus précisément: si (*Pierre* / *Paul*) ~ (*chien* / *chat*) constitue une fausse proportion (*Pierre* / *chat*) ~ (*chien* / *Paul*) en constitue une vraie, d'après les permutations autorisées, mais difficile à formuler sous forme de proposition intelligible. Ainsi semble-t-il certaines „structures profondes" de groupe pourront

être explicitées plus aisément en surface soit sous forme de proportion vraie soit sous forme de fausse proportion.

18. Harmonie et disharmonies

On a remarqué que sur les 24 permutations concevables, 8 conservaient la proportion.

Si l'on inverse un rapport sans inverser l'autre (2134) ou (1243) l'harmonie de la 1ère proportion est rompue, on a un contre-sens.

Le verbe est à l'adverbe ce que l'adjectif est au substantif.

On pourra en déduire 8 nouvelles propositions (cf. les points cerclés sur le réseau du § 5)

2134 3421 1243 4312

2314 1423 4132 3241

véhiculant la même fausse information. Elles s'obtiennent à partir des 8 proportions justes par réécriture du 1 en 2 et du 2 au 1, c'est-à-dire par l'instruction (2134) comprise en R-ien.

Si l'on permute les „numérateurs" (3214) ou les „dénominateurs" (1432), on obtient encore 8 nouvelles propositions (cf. les points non renforcés et non cerclés)

3214 1432 2341 4123

3124 2431 4215 1342

véhiculant une même information fausse, mais différente dans sa fausseté des précédentes. Là encore les instructions doivent être comprises en R-ien.

Ainsi par rapport à l'harmonie exprimée par une proportion, il existe deux disharmonies et deux seulement.

19. Post face

Le lecteur qui aura eu la patience de parcourir jusqu'au bout ces développements, sur le thème de la proportion linguistique saisira certainement l'esprit et la manière que l'auteur préconise pour l'enseignement efficace des cadres formels aux jeunes linguistes de formation littéraire. Il comprendra la raison de certaines comparaisons quelquefois un peu faciles entre les langues naturelles et des systèmes sémiotiques apparamment tout à fait différents. Certes, il semblerait que certains détours auraient pu être évités, un exposé traditionnel permet en moins de temps de présenter davantage de matière. Cependant connaître et savoir utiliser, être rigoureux et parvenir

à se faire comprendre constituent des dilemmes pour lesquels il faudra toujours se contenter d'un compromis. Ainsi la donnée d'un contre-exemple semble alourdir un exposé théorique, et cependant, humainement parlant, un contre-exemple a autant d'importance que l'exemple pertinent lui-même.

On remarquera sans doute que nombre de notions mathématiques qui auraient pu être évoquées à propos des exemples considérés n'ont pas été nommées. En effet l'auteur ne croit pas aux vertus d'une terminologie pédante. Mieux vaut utiliser correctement un mode de raisonnement que de prononcer un terme qui ne recouvre qu'une connaissance superficielle.[15]

Le soutien visuel, qu'offre la géométrie et auquel l'auteur a si souvent recours, tout comme les analogies entre les faits de langue et les faits de mathématique (et qui font sourire le spécialiste) pourront, et même souvent devront, être abandonnés dans une étude plus approfondie. Il convient de les considérer comme une étape pédagogique transitoire, indispensable pour certains esprits, explétive pour d'autres.

[15] A titre d'indication signalons que l'exposé des matières a demandé environ 6 heures de cours magistraux et de travaux pratiques pour être à peu près assimilé par un groupe d'étudiants littéraires. Il est raisonnable de prévoir des exercices de rappel pour une fixation plus profonde des notions acquises. Ce n'est qu'à ce prix qu'on évite le «verni de surface» qui ne constitue qu'un simulacre d'enseignement des mathématiques. Ce qui importe ce n'est pas tant la „quantité" de ce que l'on enseigne, que le mode de réflexion que l'on éveille en enseignant.

Zum Modellieren in der Linguistik vom quantitativen Standpunkt aus

MARIE TĚŠITELOVÁ, PRAHA

Die Theorie des Modellierens und der Modelle und ihre praktische Anwendung hat auf dem Gebiet der mathematischen Linguistik — wie bekannt — nicht überall dieselben Bedingungen; dies hängt vor allem damit zusammen, daß ihre beiden Bereiche, d. h. die algebraische und die quantitative Linguistik, verschiedenartig sind. Das Symposium der algebraischen Linguistik, das sich in erster Linie mit der Theorie des Modellierens und der Modelle befaßt, ist — meiner Meinung nach — zugleich das beste Forum dafür, dieses Problem auch vom quantitativen Standpunkt aus zu erörtern.

Für das Modellieren in der quantitativen Linguistik ist kennzeichnend, daß z. B. I. I. Revzin in seinem Buche *Modeli jazyka* [1] die Verbindung der Sprachmodelle mit dem Studium der statistischen Sprachstruktur zwar betont, aber den statistischen Modellen keine besondere Beachtung geschenkt hat. Ich halte es auch für wichtig zu unterstreichen, daß nach I. I. Revzin die statistischen Kriterien erst in dem Moment eine wichtige Rolle übernehmen, wo es sich um den *Übergang* eines abstrakten Modells zur Sprachverwendung handelt. Nur in diesem Sinne spricht er über das statistische Sprachmodellieren. Dies scheint mir kein Zufall sein. Beim Sprachmodellieren vom quantitativen Standpunkt aus gibt es manche Probleme; ihre Lösung ergibt, daß dem Modellieren auf dem Gebiet der quantitativen Linguistik eine andere Stellung zukommt als auf dem der algebraischen Linguistik. — In meinem kurzen Referat will ich eben *die speziellen Bedingungen* der Theorie des Modellierens und der Modelle in der quantitativen Linguistik diskutieren.

Welche Modelle kommen vom methodologischen Standpunkt aus in der quantitativen Linguistik zu Betracht?

Was die angewandte mathematische Methode betrifft, handelt es sich — wie in der algebraischen Linguistik — um *mathematisch-logische* Modelle [2], die auf der Zeichentheorie beruhen; vom linguistischen Standpunkt aus leisten uns die Begriffe des Strukturalismus eine große Hilfe, und zwar langue und parole und weiter signifiant und signifié. In der quantitativen Linguistik handelt es sich vor allem um *die statistischen Modelle* im weitesten Sinne des Wortes, und besonders diese Modelle werde ich in meinem Referat behandeln.

In Hinsicht auf den methodologischen Ausgangspunkt des Sprachmodellierens kann man — wie bekannt — entweder analytisch, oder synthetisch vorgehen. Mit anderen Worten: wir können entweder von den Fakten, Spracheinheiten ausgehen und dann die Regeln formulieren, mit deren Hilfe wir ein Sprachsystem bilden können, oder wir können vom System der Regeln, von einer Grammatik ausgehen und dann studieren, was für Reihen von Sprachsignalen (zum Beispiel Sätze) sie generieren können [3]. In der quantitativen Linguistik gehen wir meist *von der Analyse der Spracheinheiten* aus und bemühen uns, die Regeln zu formulieren, nach denen sich das Funktionieren der betreffenden Sprachelemente vollzieht. So können wir z. B. durch die statistische Analyse der Grapheme und Phoneme zu den Regeln über die Kombinationen dieser Elemente in der jeweiligen Sprache gelangen, durch die statistische Analyse der Wortformen oder der lexikalischen Einheiten zu den Regeln über das Funktionieren einzelner Einheiten oder ihrer Gruppen, z. B. Nominalgruppen oder Verbalgruppen, in verschiedenen Texten, Stilarten oder Sprachen u. ä. [4].

Was die bei dem Sprachmodellieren angewandte statistische Methode betrifft, nennen wir an erster Stelle Modelle, die auf den Begriffen *der Mengentheorie* aufgebaut sind. Hierher gehört vor allem das bekannte Modell der sowjetischen Mathematikerin Kulagina [5]. Dieses Modell hat I. I. Revzin in seinem obenerwähnten Buche *Modeli jazyka* weiterentwickelt. Den Ausgangspunkt des Modells von Kulagina bilden bekanntlich die sogenannten primitiven Begriffe, die weiter zergliedert werden und mittels der Definitionen und Theoreme in Beziehung gesetzt werden; bei der sogenannten linguistischen „Interpretation" werden die primitiven Begriffe mit den linguistischen konfrontiert.

Ein solches Modell, das auf den mengentheoretischen Begriffen aufgebaut ist, hat zur Voraussetzung, daß die Sprache, bzw. die Menge von Spracheinheiten homogen sei. Aber unsere Erfahrungen mit der Sprachanalyse und mit den Spracheinheiten zeigen ganz deutlich, daß es sehr schwierig ist, diese Voraussetzung im strengen Sinne des Wortes zu erfüllen. Nehmen wir als Beispiel den Wortschatz, d. h. eine Menge von Wörtern oder eine Menge von Wortformen, die in jedem Kontext vertauschbar wären, u. ä. Die Homogenität der Sprachelemente, die in konkreten Sprachen zur Verfügung stehen, stellt ein Problem sui generis und ein Hindernis im Sprachmodellieren dar. In der Praxis ist es manchmal schwer, eine Menge von Sprachelementen, die auch formal homogen wären, zu definieren. Nehmen wir z. B. den erwähnten Wortschatz als eine Menge von Wörtern, wo ein Wort als eine graphische Einheit (von einer Pause zu einer anderen) definiert wird. Wir erhalten eine Menge von Wörtern, die hinsichtlich ihrer Funktion und Semantik ganz verschieden sind; es genügt auf den Unterschied zwischen den sogenannten Strukturwörtern und Vollbedeutungswörtern, zwischen den Substantiven und Pronomina, Substantiven und Verben u. ä. zu verweisen. Hinzu kommen natürlich auch die Schwierigkeiten bei der Segmentation eines Textes in diskrete Elemente, die linear gereiht werden, in einer bestimmten Ordnung nachfolgen und zugleich zu verschiedenen Sprachebenen gehören.

Es ist daher begreiflich, daß I. I. Revzin empfiehlt (o. c. in Anm. 1), in diesem Falle eher vom Studium der typologischen Differenzen als vom Sprachmodellieren zu sprechen.

Ich bin der Ansicht, daß man trotzdem die Sprachelemente so verteilen kann, daß sie die Bedingung der Homogenität erfüllen und daß man dann über das Sprachmodellieren sprechen darf. Als Beispiel könnte man das Problem der morphologischen Homonymie im Tschechischen anführen. Eine Menge von Wortformen, die eine gleiche Form bei verschiedenen Bedeutungen aufweisen, bietet uns die notwendigen Bedingungen dafür, ein Modell des Systems der morphologischen Homonymie zu bilden. Einen solchen Versuch habe ich in meiner Arbeit über die morphologische Homonymie im Tschechischen [6] unternommen. Es ist zu bemerken, daß zum Aufbau eines Modells im eigenen Sinne des Wortes natürlich noch ein Apparat von Theoremen und Beweisen nötig wäre, aber schon die einfache Formalisation hat gezeigt, daß sie für die Einordnung der grammatischen Elemente, die sonst den Eindruck einer Summe von zufälligen Sprachelementen machen, eine große Bedeutung haben kann. Für die Sprachen, die über eine reiche Morphologie verfügen, wie z. B. die slawischen Sprachen, könnte ein derartiges Modell der morphologischen Homonymie von großer Bedeutung sein. Es könnte ein Komplement zum Modell von Kulagina darstellen.

Die auf der Mengentheorie aufgebauten Modelle stellen in der quantitativen Linguistik eine Gruppe von Modellen dar, die auf einer Gesamtheit von Sprachelementen beruhen. Eine andere Gruppe von Modellen bilden die Modelle, die auf der Wahrscheinlichkeitsrechnung beruhen und die nicht nur die Sprachelemente, sondern auch *die Gesetzlichkeit* ihres Funktionierens in den Texten, in parole formalisieren können. Entsprechend dem Charakter der Wahrscheinlichkeitsrechnung beziehen sich die zuständigen Modelle *auf die Mengen* von Gesamtheiten und auf die Verteilung der einzelnen Sprachelemente in denselben Mengen. In der Praxis bedeutet das, daß wir nicht nur eine genug große Stichprobe von Sprachelementen haben müssen, sondern daß wir uns auch dafür interessieren müssen, *auf welche Weise* die Probenerhebung von Sprachelementen gemacht worden ist, welche Resultate wir mittels einer anders gemachten Stichprobe erreichen könnten, die z. B. denselben Umfang hätte, und welche Verteilung hier die einzelnen Elemente ausweisen könnten. Indem in den mengentheoretischen Modellen *die Homogenität* der einzelnen Einheiten ein großes Problem bedeutet, erscheint in den Wahrscheinlichkeitsmodellen als großes Problem die *Probenerhebung* des Sprachmaterials, des Korpus.

Es ist bekannt, daß die Statistik über verschiedene Methoden der Probenerhebung, wie z. B. eine zufällige, mechanische u. ä. verfügt [7]. In den letzten Jahren wurde besonders die Methode der Zufallsauswahl bevorzugt. Aber eine gründliche Analyse der durch diese Methode erreichten Stichproben hat gezeigt, daß diese Methode keineswegs als Universalmethode, wie es oft geschah, benützt werden kann [8]. Wenn wir z. B. eine Textanalyse zwecks der lexikalischen Statistik durchführen wollen,

erweist sich eine Zufallsauswahl als unzureichend. Mittels dieser Methode kann man nicht einmal den Wortschatz eines Textes oder eines Autors, desto weniger eine Menge von Texten modellieren. Die Ursache dieser Schwierigkeiten liegt natürlich auch in dem Wort, d. h. der Einheit der Textanalyse. Bei der lexikalischen Analyse spielt nicht nur die Zahl der bestimmten Einheiten, der lexikalischen Einheiten eine große Rolle, sondern auch ihre Proporzionalität, ja sogar auch ihre konkrete Realisation, d. h. die Frequenz der bestimmten Worteinheiten. Nach meiner Erfahrung kann man mittels der Zufallsauswahl nicht einmal die ersten 10 häufigsten Wörter (mit Rücksicht auf die Wortarten) in ihrer Reihenfolge (Rang) feststellen. Z. B. in den belletristischen Texten im Tschechischen lassen sich mittels der genannten Methode nur die ersten 2 häufigsten Substantive und Verben feststellen. Von den selteneren und seltenen Wörtern, die für den sogenannten Wortreichtum (darüber werde ich noch sprechen) von großer Bedeutung sind, lassen sich durch die Zufallsauswahl nur 28 % der Substantive und 27 % der Verben des ganzen betreffenden Wortschatzes feststellen.

Dieses Beispiel hat meines Erachtens gut gezeigt, wie große Bedeutung der passenden Methode der Stichprobenauswahl des Korpus beim Modellieren auf der Grundlage der Wahrscheinlichkeitsrechnung zukommt. Das zweite große Problem bei diesen Modellen in der quantitativen Linguistik ist die Verteilung der Wahrscheinlichkeitsdaten der einzelnen Sprachelemente. In diesem Zusammenhang erinnern wir an das sogenannte Zipfsche Gesetz [9], das für die nach der sinkenden Häufigkeit geordneten Worteinheiten gilt und durch die bekannte Formel

$$r \,.\, F = c$$

ausgedrückt wird, d. h. r Rang, Ordnungszahl einer bestimmten Worteinheit multipliziert durch F, d. h. Frequenz einer Worteinheit sind konstant. Es ist bekannt, daß dieses Gesetz weder für die häufigsten, noch für die seltenen Wörter gültig ist. Lassen wir jetzt vervollkommneren Formen dieses Gesetzes beiseite, wie z. B. die Bearbeitung Mandelbrot's [10], d. h. das harmonische kanonische Gesetz

$$p_r = P(r + p) - B,$$

oder die Bearbeitung Herdan's [11]. Soviel steht jedenfalls fest, daß die Reihe der Wörter, die nach der Frequenz geordnet sind, nicht eine einzige Verteilung aufweist, sondern daß wir zwei oder sogar drei Typen der Verteilungen voraussetzen müssen. So hat z. B. G. Herdan [12] drei Typen von Verteilungen vorausgesetzt: die binomische Verteilung (gültig für die häufigsten Spracheinheiten), die Poisson Verteilung mit Wiederholung (gültig für die Wörter mit der mittleren Häufigkeit) und die einfache Poisson Verteilung (gültig für die seltenen Wörter). Sowohl der Typus der Verteilung der Spracheinheiten als auch das Kriterium für eine passende Probenauswahl derselben müssen nach meiner Meinung erst entdeckt werden und sollen den Gegenstand einer „linguistischen Statistik" bilden. Ich

stimme hier Herdan (o. c. in Anm. 11) bei, daß Zipf mit seinem erwähnten Gesetz den Weg dazu kompliziert hat. Dies ist aber ein Problem für sich.

Weitere und komplizierte Probleme entstehen beim Modellieren mittels der Wahrscheinlichkeitsrechnung, besonders was — sagen wir — das technische Funktionieren betrifft, z. B. bei Zuordnung der Wahrscheinlichkeitsangaben, der Auswahl der Benennung beim Modellieren der stilistischen Bestandteile des Sprachkodierens u. ä.

Die Erfahrungen mit der Bildung der Sprachmodelle auf der Basis der Wahrscheinlichkeitsrechnung zeigen, daß wir dabei eher *ein System* sowohl der Elemente als auch der Verteilung der Wahrscheinlichkeitsdaten voraussetzen müssen. Eine Lösung bietet hier *die mathematische Statistik*, die sich bemüht die subjektiven Einflüsse beim Studium der Wahrscheinlichkeitsgesetzmäßigkeit auszuschalten. Indem die Wahrscheinlichkeitsrechnung die Modelle zur Beschreibung einiger Elemente und ihrer Gesetzmäßigkeit bildet, bietet uns mathematische Statistik Methoden, die das ganze System solcher Modelle mit der Wirklichkeit vergleichen. Es handelt sich besonders um die Theorie der Schätzung und die der Hypothesetests, Hypothesenprüfung, die die geprüften Modelle entweder annehmen oder ablehnen. Ich halte es für nützlich zu bemerken, daß die Mathematiker selbst zugeben, daß die Methoden der mathematischen Statistik nicht alle Probleme lösen können, besonders nicht diejenigen, die kompliziert sind, daß nicht alle Methoden als optimal bezeichnet werden können und die betreffenden Methoden nicht immer passend angewendet werden [14]. Damit hängen auch die Schwierigkeiten beim Aufbau der statistischen und der Wahrscheinlichkeitsmodelle zusammen. Dies aber bedeutet nicht, daß die Methoden der statistischen Linguistik nicht weiter entwickelt werden sollten. Es ist sogar mit Hinsicht auf den spezifischen Charakter des Sprachmaterials erforderlich, denn die Analyse der Beziehung zwischen den empirischen Theorien und den zugehörigen Daten setzt *eine Hierarchie der Modelle* von verschiedenen Typen voraus. So wird z. B. *in der Lexikalstatistik* der sogenannte *Wortreichtum* untersucht, d. h. die Menge von verschiedenen Lexikaleinheiten in den Texten von verschiedenen Autoren, Stilarten, Sprachen u. a. Abgesehen davon, daß der Termin (Wortschatzreichtum) nicht zutreffend ist, zeigt es sich, daß die Problematik, die die Struktur eines Lexikons darstellt, viel komplizierter ist, als daß sie mittels der bekannten Guiraud-Formel [15]

$$R = \frac{V}{\sqrt{N}} \quad \text{oder in der Variante} \quad R = \frac{V}{\sqrt{2N}}$$

erfaßt werden könnte.

Auf Grund meiner Ergebnisse auf dem Gebiet der Lexikalstatistik habe ich gezeigt [16], daß zum gegebenen Zweck eine einzige quantitative Charakteristik nicht ausreicht, sondern daß es notwendig ist, ein System von solchen Charakteristiken zu entwickeln, und zwar zumindest folgende drei:

1. *Umfang* des Wortschatzes (R), gegeben durch das Verhältnis der Zahl der lexikalischen Einheiten V und der Länge des betreffenden Textes N:

$R = 100 \frac{V}{80\ \%\ N}$ (80 %, falls wir alle Wörter mit Vollbedeutung in Betracht ziehen),

oder $R = 100 \frac{V}{70\ \%\ N}$ (70 %, falls wir als Wörter mit Vollbedeutung die Substantive, Adjektive, Verben und Adverbien in Betracht ziehen);

2. *Zerstreuung* des Wortschatzes (D) gegeben durch das Verhältnis der Zahl der Wörter mit Vollbedeutung mit der Frequenz 1—10 und des Wortschatzes V:

$$D \triangleq 100 \frac{\sum_{1}^{10} V}{V}.$$

3. *Konzentration* des Wortschatzes (K) als Komplement zur Zerstreuung, gegeben durch das Verhältnis der Länge des von den 10 häufigsten Wörtern gebildeten Textes und der Länge des Textes N:

$$K = 100 \frac{\sum_{1}^{10} N}{N}.$$

Die angeführten drei Charakteristiken können meines Erachtens vom quantitativen Standpunkt aus die Grundzüge des Wortschatzes zeigen und dadurch zur Formalisierung der Lexikalstatistik beitragen.

Den Ausgangspunkt dieses Studiums bilden — wie bekannt — die Zahlangaben in der Form von Tabellen oder experimentellen Kurven. Diese Tabellen und Kurven können wir in der Regel nicht direkt erklären, sondern mittels der mathematischen Formeln, die mehr oder weniger kompliziert sind. Dies ist vom linguistischen Standpunkt aus kein Ziel, sondern ein Mittel zur Entdeckung der neuen Eigenschaften der Spracheinheiten, bzw. zur neuen Erklärung der bekannten Eigenschaften u. ä. Es ist aber auch bekannt, daß ein Bild in der Form von Tabellen und Kurven in jede Theorie dieser Art etwas Unbestimmtes einführt. Mit anderen Worten: jede Zahl, die ein Maß ausdrückt, ist von einem empirischen Fehler beeinflußt. Deshalb wählen wir oder müssen wir unter den Formeln wählen, die natürlich mehr oder weniger genau sind. Darum müssen wir damit rechnen, daß die mittels dieser Methoden erhaltenen Resultate unserer Arbeiten nur approximativen Charakter tragen. Eines von den Grundproblemen liegt darin, daß wir immer mehrere Sprachelemente studieren, die in verschiedenen Relationen vorkommen.

Abschließend wollte ich noch die Frage nach dem Sinn der Modelle vom Standpunkt der quantitativen Linguistik aus kurz berühren.

In meinem Referat wollte ich zeigen, daß es sich bei den Sprachmodellen *um die Quantifizierung der empirischen Theorie*, nicht nur um die bloße Anwendung der statistischen Methoden an das Sprachmaterial handelt. Wir bilden eine Theorie, wobei wir den Beschreibungstermini Zahlen zuordnen. Wir erhalten dadurch eine Menge von empirischen Gesetzen, die unsere Kenntnisse über die Sprache, ihre einzelnen Elemente und ihre Beziehungen exakter machen und weiter vertiefen sollen.

LITERATUR

[1] REVZIN, I. I.: Modeli jazyka. Moskau 1961, bes. S. 16. — BERKA, K. — NOVÁK, P.: Výklad fonologických a gramatických pojmů pomocí teorie množin. SaS, *24*, 1963, S. 133 — 140.

[2] Teorie modelů a modelování. Prag 1967. — KIEFER, F.: Some Aspects of Mathematical Models in Linguistics. Statistical Methods in Linguistics (SMIL), 1964, Nr. 3, S. 8—26.

[3] SGALL, P.: Generativní popis jazyka a česká deklinace. Prag 1967.

[4] TĚSITELOVÁ, M.: K typologii slovanského slovníku z hlediska kvantitativního (na českém materiálu). Čs. přednášky pro VI. mezinár. sjezd slavistů v Praze 1968, S. 95—99. — KELEMEN, J.: Le problème des modèles statistiques lexicaux basés sur la relation "type-token" du point de vue espèces de styles de la langue littéraire. Cahiers de linguistique théorique et appliqués, *4*, 1967, S. 83—88.

[5] KULAGINA, O. S.: Ob odnom sposobe opredelenija grammatičeskich ponjatij na baze teorii množestv. Problemy kibernetiky, *1*, 1958, S. 203—215; vergl. SaS, *21*, 1960, S. 34—41.

[6] TĚŠITELOVÁ, M.: O morfologické homonymii v češtině. Prag 1966, S. 15.

[7] KÖNIGOVÁ, M.: K otázce statistického výběru v lingvistice. SaS, *26*, 1965, S. 161—168.

[8] Vergl. TĚŠITELOVÁ, M.: A propos du sondage aléatoire du point de vue linguistique. Actes du X^{e} congrès international des linquistes. Bucureşti 1970, S. 974—977.

[9] ZIPF, G. K.: Human Behavior and the Principle of Least Effort. Cambridge 1949.

[10] MANDELBROT, B.: Structure formelle des textes et communication. Word, *10*, 1954, S. 1—27. Vergl. Teorie informace a jazykověda. Prag 1964, S. 130—150.

[11] HERDAN, G.: Type-Token Mathematics. Gravenhage 1960, S. 38 ff.

[12] HERDAN, G.: Quantitative Linguistics. London 1964.

[13] NOVÁK, P.: K jednomu modelu stylistické složky jazykového kódování. SaS, *27*, 1966, S. 29—40.

[14] FABIAN, V.: Základy statistické metody. Prag 1963, S. 15—16.

[15] Vergl. GUIRAUD, P.: Les caractères statistiques du vocabulaire. Paris 1951.

[16] TĚŠITELOVÁ, M.: On the so-called Vocabulary Richness. Prague Studies in Mathematical Linguistics, *3*. Prague, Academia 1972, S.103—120.

On the Statistic in Syntax

LUDMILA UHLÍŘOVÁ, PRAHA

1. When P. Sgall has characterized the principles of the so-called algebraic linguistics in his book *Generative description of language and the Czech declination* [1], he has written there among others: "The point of the new conception can be really seen in the approach itself, that is, in the fact, that language system becomes the object of an explicit description, for the formulation of which certain mathematical device is chosen" (p. 7). He has stressed two points, first, the *explicitness* of the description, and second, the fact that algebraic linguistics, as it is represented by generative grammar, deals with the explicit description—or, explicit modelling—of *language system* (the concept of language system being understood either as a set of well-formed sentences, or in some other sense). The said thesis clearly reveals the differentia specifica of the algebraic linguistics as opposed to the other branch of mathematical linguistics, namely the quantitative (statistical) linguistics. Quantitative linguistics (or, statistical — the latter term being usually taken for narrower than the former) shares with the algebraic linguistics the feature of expliciteness: statistical observations provide an explicit, numerical basis for our intuitive linguistic observations. Or the other hand, the up-to-date experience has shown that the statistical linguistics is very often linguistics of text rather than a kind of the description of language system; it holds especially of the higher levels (strata) of the language system. At the same time it should not be forgotten that any quantitative description makes sense only if it is preceded by a structural analysis, and its significance increases proportionally to the depth of the structural analysis, on which it is based and from which it again results [2].

Various kinds of statistical investigation aim at a characterization of e.g. a text, group of texts, individual style, functional style etc. and they are usually done in such a way that absolute and relative frequencies concerning some linguistic *phenomena* (items), are calculated, and then some meaningful quantitative *relationships* are searched for among them. Only if such relationships are discovered and quantitatively described, one necessary condition is fulfilled (though not always the only and sufficient one) under which it is possible to speak about quantitative modelling, for which the statistical processing is only the material prerequisite.

The comparison with the algebraic linguistics leads to the question, whether it is possible to describe quantitatively the systemic relations among elements no different language levels. Let me start with several general remarks on the problem and then illustrate it on an example from the syntactic level.

First, as it is well-known from the theory and practice of the linguistic school of Prague, that the place of an element in the language system is defined by an ensemble of systemic relations (oppositions) to the other elements in the system. The frequency characteristics either may be in agreement with the position of the element in the system (elements belonging to the centre of a language system are high frequent and elements belonging to the periphery are low frequent), or, it may be in disagreement with it (an element on the periphery of the system preserves its existence owing to its high frequency of occurrence), or may be complicated in many other respects. The relationship between system (quality) and frequency (quantity) is neither straightlined nor simple in all cases [3].

Second, language is potentially infinite both generative and in the statistical senses. If the data are processed statistically, it is necessary to choose a representative sample from the infinite population, i.e. a homogeneous and sufficiently extensive one, and the unit of investigation must be clearly defined. In other words, the specific character of linguistic material must be respected. Statistical investigation may be performed either on an *inventory* of linguistic elements such that each of the elements is one unit, e.g. on vocabulary, or, on the contrary, directly on *texts*. In the first case, the so-called functional load, if Mathesius's term is used [4], e.g. the functional load of phonemes or groups of phonemes, word-formative exponents etc. is studied. The inventory of linguistic elements is considered in such cases as an approximation of the level of language system. Frequency characteristics which display a certain reasonable degree of stability with regard to it can be considered as frequency characteristics of phenomena and relationships in the language system. (However, at the same time an empirical question arises concerning the limits of the stability.)

In the second case occurrences in texts are studied. The relationship to the language system (or, to a certain level of it) is given only through the mediation of texts, and the answer of a question whether the texts represent the language system, remains always only relative ("to what degree"). The fact is substantial that it is the only reasonable and effective way of investigation of phenomena belonging to the syntactic level.

2. Below, I will try to illustrate some problems of statistical approach in syntax, taking as the basis of the illustration the relationship between the clause length and the syntactic structure of the clause.

The question is: what changes in the syntactic structure take place when the length of the clause increases. It is obvious that all of the syntactic elements in the clause are not equally frequent in clauses, and that when the clause length

increases, some syntactic elements occur in the clause more often than the others. Let us study statistically the frequency of occurrence of each of the clause elements in clauses of different lengths and by means of it let us empirically characterize the relationship between the syntactic structure and the length. The main notions of the linguistic analysis are as follows:

— clause — a syntactic unit with one predication nexus,

— syntactic structure of the clause — all syntactic elements explicitely present in the clause,

— clause length — the number of syntactic elements in the clause (number of words being irrelevant).

As the material four running texts of scientific style of present-day Czech were taken. The investigation consists of three steps.

2.1. In the first step the clauses were classified according to the kind and number of the following syntactic elements: subject (S), verbal predicate (P), verbo-nominal predicate (Pn) with copula (Sp), verbal attribute (D), object (O), adverbial (Ad), (noun) attribute (At), basis of the predicateless clause (J) and the group of the other low frequent syntactic elements of secondary importance, such as parenthetic elements etc. Because our material is a representative of written scientific texts, very long clauses, even of more than 30 syntactic elements, occur, though relatively rarely. Such extremely long clauses, the number of which is too small for the statistical processing, are excluded from the investigation. However, if we wish to process 90 % of clauses of the corpus, all clauses in the lenght interval from 1 to 12 or even 13 syntactic elements must be taken under investigation, as is shown in Table 1.

Now let us pay attention to the frequency distribution of each of the elements in clauses in the length interval from 1 to 13 elements. The detailed survey of numerical data is given in the Table 2. Let us first consider the last line of the Table 2, labelled with symbol x. This symbol indicates a fictive syntactic element which has the property of occurring just once in every clause, regardless to the length of the clause. The syntactic element x does not thus show any relationship to the clause length. On the contrary, all of the really existing syntactic elements manifest an apparent and close relationship to the clause length. The theoretical frequencies of x are similar only to the predicate group ($P + Pn$), which naturally follows from the definition of the clause as a unit with one predication nexus. The closest relationship to the clause length is manifested by the attribute. Whereas in the shortest clauses (cf. data for the lengths 1—3) the attribute either does not occur at all or is very seldom, in the clause of the length of 13 it makes 51 % of all syntactic elements. Also the group of verbal complements ($O + Ad$) increases with the clause length, but more slowly. A very low increase characterizes subject.

2.2. The second step consists in reducing the clause in the following way. Only the syntactic elements, directly dependent on the predicate (P, or Pn), namely S, O, Ad, D and Av (the so-called inherent adverbial; a special type of adverbial — a qualitative

Table 1

The frequency distribution of clause lengths

Clause length	1	2	3	4	5	6	7	8	9	10	11	12	13
Absolute frequency	52	169	311	404	518	468	432	322	270	216	181	123	100
Relative frequency in %	1	4	8	10	13	12	11	8	7	6	5	3	3
Cummulative frequency in %	1	5	13	23	36	48	59	67	74	80	85	88	91

Table 2

The relative frequencies of syntactic elements in clauses of different lengths

Clause length	1	2	3	4	5	6	7	8	9	10	11	12	13
S	1.92	15.38	12.75	16.15	14.56	13.28	11.71	10.41	9.71	8.65	8.24	7.66	7.58
P + *Pn*	92.30	42.30	31.51	24.13	19.58	16.20	14.39	12.20	11.19	10.65	9.54	8.67	7.97
O + *Ad*	1.92	25.74	29.90	24.50	23.56	21.69	22.89	22.34	19.92	20.19	19.19	19.04	21.11
At	—	1.78	11.14	20.36	28.08	35.29	38.50	40.90	44.60	46.65	49.47	51.22	50.97
D	—	2.07	1.18	1.67	1.58	1.78	1.62	1.83	1.40	1.72	1.31	1.22	0.93
others	—	2.37	5.68	5.32	6.57	5.91	6.52	8.24	9.05	8.09	8.79	8.87	8.51
Sp	—	6.51	6.32	7.12	5.56	5.24	3.97	3.73	3.79	3.35	2.96	2.85	2.54
J	3.85	3.85	1.50	0.74	0.50	0.60	0.40	0.35	0.33	0.70	0.50	0.47	0.39
Σ	99.99	100.00	99.98	99.99	99.99	99.99	100.00	100.00	99.99	100.00	100.00	99.98	10.00
x	100.00	50.00	33.00	25.00	20.00	16.67	14.28	12.50	11.11	10.00	9.09	8.33	7.69

determiner of verb [6] — in the first step of our investigation it was included in the group of low frequent secondary syntactic elements), all other being excluded from the investigation. Thus we have eliminated those syntactic elements which play the most important role in the relationship of length — structure, i.e. those syntactic elements, the number of their occurrences increases most with the increasing clause length. The syntactic elements included are studied now in more detail, also with respect to the parts of speech (e.g. occurrences of *P*—verbal element are calculated separately from *Pn* — nominal element; *Ad* — nominal element separately from *Av* — adverb; etc.). The most of the "reduced" clauses consist of 1—5 syntactic elements (longer clauses are quite exceptional). The relationship of the structure and length manifests itself now in the following way, see Table 3. The influence of the clause length is most apparent in the case of adverbials. The longer the clause, the higher proportion of adverbials as compared with the relative frequencies of other syntactic elements. Whereas in clauses of the length one there is only 1 % of *Ad*, in clauses of the length two occurrences make 10 % and in the longest clauses even 36 % of occurrences of all syntactic elements. Also the relative proportions of verbal attribute and inherent *Av* slowly increase. The decrease of the predicate group is worth mentioning: *Pn* decreases much more with the clause length than *P* does; it means that longer clauses than 2 prefer verbal predicate rather than the verbo-nominal one. Probably it is so because the verbo-nominal predicate is not able intentionally to govern so many minimal syntactic elements as the verbal predicate does, and it is obvious that it is the number of nominal syntactic elements, which increases with the length of the clause. This statement may be formulated also in another way: if there is a high frequency of verbo-nominal constructions in a text, there is also a high frequency of short clauses. The contrary does not

Table 3

The relative frequencies of syntactic elements in reduced clauses of different lengths

Clause length		1	2	3	4	5
Relative frequency of syntactic elements in %	*S*	9	29	24	21	18
	P	33	22	25	21	17
	Pn	35	24	8	3	2
	O	—	12	17	15	13
	J	22	1	1	1	1
	D	1	1	3	4	4
	Av	—	1	3	6	8
	Ad	1	10	20	29	36
	Σ	100	100	100	100	100

hold, of course: the lack of *Pn* itself does not mean that there is a high frequency of long clauses in the text. If we now compare the distribution of clause lenghts in our four texts (see Table 4), we can see that the only differences among the texts are in

Table 4

The distribution of reduced clause lengths in absolute and relative frequencies

Clause length		1	2	3	4	5	Σ
Text W	abs.	38	204	428	235	58	963
	%	3.95	21.18	44.44	24.40	6.02	99.99
Text A	abs.	42	215	431	179	32	899
	%	4.67	23.91	47.94	19.91	3.56	99.99
Text K	abs.	54	347	373	180	32	986
	%	5.48	35.19	37.82	18.25	3.25	99.99
Text P	abs.	50	232	411	220	48	961
	%	5.20	24.14	42.76	22.89	4.99	99.98
	Σ	184	998	1643	814	170	3809

the distributions of clauses of the length 2, where there is an agreement in the distributions of clauses beginning with the length of 3. In the text *K* the number of clauses of the length 2 is the highest of all texts. If we now compare the frequencies of *Pn* in the texts with each other in the shortest clauses (lengths 1 and 2), we can observe (see Table 5) that it is highest in the text K (the text K dealing with philosophical reflections, W — university textbook on biology, A — handbook on electronics, P — psychological review).

2.3. If we — as the third step — take under consideration not only the number of syntactic elements in the clause, i.e. the clause length, but also their linear arrangement word order, it becomes apparent that the shortest clauses (of the length 2) are

Table 5

Frequency of *Pn* in clauses of different lengths in four texts

Text		W	A	K	P	∅
Frequency of *Pn* in %	clause length of 1	37	33	50	18	34.50
	clause length of 2	21	26	28	19	23.50

Table 6

The most frequent syntactic elements on the beginning and end of clauses of different lengths

Clause length		2	3	4	5
The most frequent syntactic element	on the beginning	*P* + *Pn* 46 %	*S* 47 %	*S* 49 %	*S* 51 %
	on the end	*P* + *Pn* 46 %	*O* 34 %	*Ad* 38 %	*Ad* 36 %

governed by other word-order statistical regularities and tendencies than longer clauses (of three or more syntactic elements), which is, last but not least, a consequence of different frequencies of syntactic elements in clauses of different lengths. E.g. a clause in a scientific text most often begins with subject (in 46 %) and most often ends with object (28 %); these data are the average data taken from the material as a whole. But if we take clauses of different lengths separately, the proportions substantially differ, as is shown in Table 6.

To sum up: If all average data of similar kind, e.g. the data concerning the average frequency of occurrence of syntactic elements, or syntagms, in texts (cf. the last column in Table 3) ought to be correctly interpreted and interrelated, it is important and necessary to know the distribution of the length of the syntactic units (clauses, or sentences) under investigation. The *factor of length* we consider as one of primary importance for the statistical investigation of syntactic structure, both of the statistical and linguistic (structural) reasons.

3. Conclusion. If one speaks about the expliciteness in the domain of statistical linguistics, then, last but not least, also the explicit formulation of statistical and linguistic conditions under which the investigation is undertaken and under which it holds, makes an important and unseparable part of it. The aim of this paper was to demonstrate that one of the important conditions for the study of the frequencies of syntactic elements, is to know the distribution of lengths of the chosen syntactic units.

In the present paper we have dealt with the relationship between the syntactic structure and the length of the clause, which is a kind of dependence relationship between two variables. We have not settled, which of the two variables is the dependent one and which is the independent one. To determine *the direction of the dependence* is not the matter of the statistical method or procedure, but of the linguistic definition. In our case the primary, independent variable is the syntactic structure and the dependent one is the length, because the syntactic elements by which the clause is expanded and the length increased, are dependent exclusively on the kinds of their explicit, or sometimes implicit (e.g. the implicit subject in Czech) syntactic determinants.

REFERENCES

[1] SGALL, P.: Generativní popis jazyka a česká deklinace. Praha 1967, p. 238.

[2] NOVÁK, P., TĚŠITELOVÁ, M.: Kvantitativní lingvistika. In: SGALL, P. et al.: Cesty moderní jazykovědy. Praha 1964, pp. 103—133.

[3] VACHEK, J.: Dynamika fonologického systému současné spisovné češtiny. Studie a práce lingvistické, vol. 8, Praha 1968, pp. 154.; DANEŠ, F.: The relation of centre and periphery as a language universal. TLP II, Praha 1966, pp. 9—21.; VACHEK, J.: On the integration of the peripheral elements into the system of language. TLP II, Praha 1966, pp. 23—37.

[4] MATHESIUS, V.: La structure phonologique du lexique du tchèque moderne. TCLP I, Praha 1929, pp. 67—84.

[5] KOPEČNÝ, F.: Základy české skladby. Praha 1962.

Problems of Generative Model in Psycholinguistics

JAN PRŮCHA, PRAHA

Psycholinguistics has become a fashion — says T. Slama-Cazacu in her *Introduction to Psycholinguistics* (1968). Whether we agree with this statement or not there is no doubt that the rapid development of this discipline with a far from long history has led to serious theoretical and methodological problems. Those connected with grammatical generative models will be the subject of the present contribution.

In the early sixties new theories gained ground in American psycholinguistics. Their authors were N. Chomsky on the one hand and some linguists and psychologists working with him on the other, in particular G. A. Miller.[1] The substance of these theories lies in a different conception of language and of the aims of linguistic theory than the one generally prevailing. Against the old dichotomy "langue—parole" Chomsky introduced the concepts of linguistic competence (LC) and linguistic performance (LP). As aims of linguistic theory he accordingly set up the production and exact description of two actually psycholinguistic models: a perceptual model and a model for acquisition of language (Chomsky, 1964 and his other works).

It is known that in his theory Chomsky postulates a substantial difference between LC models and LP models, the former describing the language mechanism (the speaker's knowledge of his language), with which man is equipped, without any explanation, however, of the functioning of the mechanism, the latter — the LP models (models of language users) — explaining the actual behaviour of the speaker and of the listener, i.e. the processes of producing and understanding messages. Chomsky's theories — taken as a whole and in their individual parts as well — contain some controversial and vague points, which from the psycholinguistic as well as purely linguistic points of view were subjected to discussion by a number of authors, e.g. Uhlenbeck (1963, 1967), Osgood (1963, 1968), Harmon (1967), Leontev (1967, 1968, 1969), Rommetveit (1968), Schlesinger (1967), Fromkin (1968) and others. The main points examined are the mutual relationship between the LC

[1] For psycholinguistic theory two studies resulting from this cooperation have been of the greatest importance, namely Chomsky and Miller (1963), Miller and Chomsky (1963).

and LP models, between the model of encoding and the model of decoding, the substance and the structure of the so-called competence, etc.

A great paradox in the development of the scientific theory still holds, namely that in spite of a strict distinction between LC and LP made by Chomsky the difference has been neglected[2] in numerous psycholinguistic investigations on behalf of a tendency to incorporate the generative (transformational) model as a basic element into the LP model.[3] This has conduced to an identification of the LC generative model with the LP model.

The growth and the interpretation of psycholinguistic models based on the generative principle have led to a number of experimental researches by which the "psychological reality"[4] of the models is put to the test. From a number of works published on the subject only some of the basic findings[5] crucial for the theory of psycholinguistics are presented. An example regarded as classic today is provided by a series of experiments carried out by G. A. Miller and his collaborators and opened by the study entitled *Some Psychological Studies of Grammar* (Miller, 1962). The results of these experiments (see especially Miller and McKean, 1964; Miller, McKean, and Slobin, 1962; Mehler and Miller, 1964; Mehler, 1963) are considered as a confirmation of the psychological reality of the generative model.

The methods of Miller's experiments are based on the assumption that a given psycholinguistic model can in principle be verified in the following manner: subjects are set a task with language material, certain results obtained indicating the way the task has been performed. The task in question in the original variant of Miller's experiment is the matching of sentences which are identical from the lexical point of view but differ in their syntactic structure. The experimental procedure consisted in

[2] In his book *Aspects of the Theory of Syntax* Chomsky said that this mixing of LC and LP was "a continuing misunderstanding" and on that account went as far as to contemplate changing the term "generate" (Chomsky, 1965, p. 9).

[3] The demand to include the generative model in the LP model was, of course, advocated by Chomsky himself (e.g. in 1964, pp. 112—113): "A perceptual model that does not incorporate a descriptively adequate generative grammar cannot be taken very seriously." This demand was being theoretically justified and developed in various works of the "transformationalists", e.g. Katz and Postal (1964) and others. A psychological formulation of this conception is given in the work of Miller, Galanter and Pribram (1960).

[4] It should be noted that in spite of the frequent use of the term "psychological reality" of transformational grammar, etc. the concept is not clearly defined and is actually intrinsically contradictory: how can a model (i.e. LP model) whose main part (i.e. LC model) is not to be interpreted in psychological terms constitute a "psychological reality"? Cf. also Osgood's remark: "It is one thing to use notions like "competence", "knowledge" and "rules" as heuristic devices, as sources of hypotheses about performance; it is quite another thing to use them as explanations of performance..." (Osgood, 1968, p. 505).

[5] For detailed survey of works on experimental verification of psychological reality of generative model see e.g. Leontev (1969), Fodor and Garrett (1966).

the presentation of two columns of sentences: the first contained kernel sentences (i.e. simple, affirmative, active, declarative), the second contained the same sentences with changed syntactic structures (e.g. negative, passive) and set in another order. The time required for a correct indentification of sentences differing in their syntactic structure is — according to Miller — dependent on the transformational complexity of the sentences.

The results of different variants of Miller's experiment (with exact chronometric measurements) have actually shown that 1. the transformation of a kernel sentence into a negative construction requires less time than the transformation of a kernel sentence into a passive construction, 2. the time required for sentences with both transformations (i.e. negative + passive) is roughly equal to the sum of intervals necessary for the completion of the individual transformations considered separately.[6] These results have been interpreted as showing the psychological reality of the generative model and have also served for basis of Miller's hypothesis that a) the more complex is the grammatical transformation the more time is required for performing it, b) any non-kernel sentence is remembered by first transforming it to its underlying kernel and then remembering the kernel plus an information (grammatical rule, memory tag) on the respective transformation.

Miller's experiment was repeated in various modified forms (every time with English material only) by other researchers and on the whole the results bore out Miller's hypothesis. There were other attempts to verify the psychological reality of the generative model by other means, the experiment of Savin and Perchonock (1965) being one of them. They studied the relationship between the transformational complexity of sentences and their storage requirements. Analogously to Miller's hypothesis it was assumed that the greater the syntactic complexity of the sentence (measured by the number of transformation rules necessary for deriving them from the kernel sentence) the greater the demands on storage. The results have confirmed that the smallest demands on storage is made by kernel sentences and that sentences requiring 2 transformations limit the storage more than sentences with 1 transformation. This has again been regarded as a proof that the so-called transformational history of the sentence significantly correlates with the facility with which the sentence is remembered and, therefore, as a proof of the correlation between the psychological process of decoding the sentence and the generative model.

Among other things it was found for instance by Mehler (1963) that kernel sentences were more easily remembered than other types of sentence, which implied that in the process of remembering sentences were not stored in their surface struc-

[6] "This is, in essence, what a transformational theory says: passive, negative, and passive-negative sentences contain all the same syntactic rules as do active-affirmative sentences, plus one or two more which increase their complexity and require a little more time for interpretation (or production)" (Miller and McKean, 1964, pp. 307—308).

ture but in some elementary form of deep structure together with a minimum of information on transformations necessary for producing the appropriate surface structure. Coleman (1964) found also that kernel sentences (active-verb constructions) were more easily understood by subjects than sentences with transformations (nominalizations and passives). Clifton, Kurcz and Jenkins (1965) confirmed the basic principle that a greater number of transformation rules required a greater number of operations.

Even though some of the data of recent experiments[7] support or at least do not refute the above results, the psychological reality of the generative model cannot be regarded as proved.[8] The statement is the outcome of contradictory results of other experiments on the one hand and of serious objections of theoretical criticism on the other. Relevant experimental data are presented first:

Miller's hypothesis of transformational decoding is called in question by Iljasov (1968) who carried out a research which with the greatest possible precision repeated Miller's basic experiment in all its three variations (as far as the procedure, the presentation of the material, the number and structure of sentences, the selection of subjects, etc. are concerned). The only difference actually was that Iljasov worked with sentences of Russian. In the experiment quite-insignificant differences in differential interval for sentences of various syntactic types were found, all this implying that Miller's hypothesis was not confirmed. At the same time a new variable was brought into the experimental verification of psycholinguistic models, namely the type of language, which had not been considered in the experiments carried out primarily with English material, one of the few exceptions being the research of Forster (1966) and Rommetveit et al. (1967) who worked with Turkish and Dutch, respectively.

Already before Iljasov Clark (1955) published his experimental finding that in the test called "Cloze Procedure" (where words systematically deleted from a text were to be replaced) another "semantic activity" was attributed to the actor and to

[7] Beside the above-mentioned experiments based on an older version of the generative model, in the last few years works verifying the psychological reality of the new version of the model as elaborated by Chomsky (esp. 1965) have appeared. These works are characterized by their concentration on the research into deep structures and into features of an associative and semantic nature, see e.g. Ammon (1968), Blumenthal (1967), Danks (1969), Marks (1967a, b), Martin (1969), Martin and Roberts (1966), Martin, Roberts and Collins (1968), Mehler and Carey (1967), Morris, Rankine and Reber (1968), Roberts (1968), Rosenberg (1968), Sachs (1967), Salzinger and Eckerman (1967), Slobin (1968), Taylor (1969).

[8] Cf. Bever's statement: "The problem with most of these experiments is that they were in large part devoted to the confirmation of the "psychological reality of TG". ... It is easy to show that phonological, surface, and deep structures of sentences have psychological reality (other than that inherent to the linguistic intuitions themselves). But the problem of exactly how these structures interact in actual psychological processes such as perception, short-term memory, and so on, is bewildering" (Bever, 1968, p. 490).

the object than in active sentences. Clark found at the same time that the beginnings of sentences with both grammatical constructions carried a smaller informational uncertainty than the endings, i.e. informational uncertainty was increasing with the linear course of the sentence from left to right. This testifies the inadequacy of Miller's hypothesis on the one hand and the existence of a sequential left-to-right processing of sentences[9] on the other.

In another experiment again Tannenbaum, Evans and Williams (1964) submitted to a test Miller's hypothesis of the dependency of time required for carrying out a given transformation on the very complexity of the transformation. Instead, however, of matching sentences the subjects directly generated them. This led to the significant finding that the amount of time necessary for generating certain sentence types was no function of their transformational complexity but was given by the frequency of their use.

Other experimental works, too (Forster, 1966; Marks, 1967a,b; Clark and Begun, 1968; Levelt, 1969) call into question the psychological reality of the generative model and appear to give support to the hypothesis of the left-to-right processing of sentences and of the semantic factor having a greater significance than the syntactic one in sentence decoding (Sachs, 1967; Taylor, 1969). The latest development is Epstein's (1969) refutation of the conclusions made by Savin and Perchonock (1965) that different syntactic types of sentence bring about different demands on storage.

Moreover, it is shown that in the experiments of Miller and his followers certain important factors have been neglected, such as the frequency of the occurrence of individual types of sentence with language users, the influence of the associative word structures, the sentence length and depth, etc.[10] It has been found for instance that in experiments where length (in number of words) and depth (according to the index of Yngve) of sentences have been controlled, the influence of the syntactic type on sentence decoding is negligible (Martin and Roberts, 1966).

Psychological reality of the generative model is opposed by direct experimental findings as well as by theoretical criticism. A detailed and convincing critical analysis of the experiments performed by Miller and his group was published by R. Rommetveit and A. A. Leontev. On the basis of his own experimental research, too, Rommetveit (1967, 1968, 1969) rejects as incorrect the investigation into sentences "in

[9] "These results present negative evidence for the notion that Ss generate passsive sentences simply by imposing a transformation on active sentences, as Miller's (1962) transformational model would imply; instead, the results argue for a sequential left-to-right generation of sentences" (Clark, 1965, p. 369).

[10] "The fact that one can show experimentally that some types of sentences are handled in some ways with greater difficulty than others, may mean that ease of handling is a function of the frequency with which the base structured sentences occur in everyday discourse" (Jonckheere, 1966, p. 87).

vacuo", isolated, deprived of their natural communicative context. Another objection is aimed at the assumption made in Miller's experiments that sentences with a given syntactic structure always are encoded and decoded in the same way under various conditions. This approach is regarded as erroneous by Rommetveit who emphasizes that in encoding and decoding a message the main part is played by semantic and pragmatic factors, the conditions of communication, the characteristics of the communicating persons, etc. This criticism originates from Rommetveit's entirely different approach to psycholinguistic models: whereas the main point of the psycholinguistic "school" of Chomsky—Miller is the investigation into the formal-syntactic properties of sentences, Rommetveit, on the contrary, advocates an investigation into the semantic and contextual aspect of sentences as a clue to an adequate study of the activity of the speaker and the listener. He is right in arguing that in the actual communication the semantic and syntactic aspects of the utterance are combined and are not realized as separate by the language user, whereas in Miller's experiments the subjects consciously concentrate on the syntactic element alone. And in all this only the so-called rote learning is involved, i.e. immediate and verbatim learning, which is not used in real communication.

Rommetveit often goes back to the views expressed by E. M. Uhlenbeck (1963, 1967) in his detailed appraisals of transformational theory. Together with Uhlenbeck he points to the discrepancy — in Chomsky's conception — between the LC of the ideal speaker-listener or in other words "the personification of la langue" on the one hand and the requirement that this competence represented by generative grammar should be incorporated into the LP model on the other hand. On the whole, Rommetveit characterizes the "post-Chomskyan" psycholinguistics endeavouring to verify the psychological reality of the generative model as a "détour strategy": "Instead of exploring the utterance in terms of available psychological evidence concerning word meanings, syntagmatic potentialities, and cognitive states, the researcher embarks upon a search for the psychological reality of assumed linguistic structures which are as yet very poorly understood" (Rommetveit, 1968, p. 216).

Rommetveit's conception is quite close to that of Leontev (1967, 1968, 1969), whose criticism is primarily concerned with the incorrectness of the experimental verification of the generative model and of the intrinsic shortcomings within the Chomsky—Miller conception. The main points of his criticism are in brief as follows: 1. In "transformational" psycholinguistics LC and LP models are constantly mixed; models produced by linguistics and a linguistic way of thinking directed towards the description of units and their properties and not towards the process involved are transferred into psycholinguistics. 2. The greatest shortcoming of the Chomsky—Miller model is that motivation and any "pre-grammar" stage in speech encoding are completely ignored. 3. The classic experiment of Miller and further experimental verifications of the generative model prove only the possibility of transforming sentences but not the actual way of generating them. 4. The generative model is

a theory of an exclusively unconscious use of language and does not include a description of various forms of conscious processes in speech activity. 5. The conclusions of the experimental verifications of the generative model cannot be generalized since they relate to one form of speech alone (monological, written form, isolated sentences with no context), whereas the psychological conditionality of the production of other forms of speech (especially the spoken form) is apparently very different.

A critical but also a very constructive and stimulating attitude towards the genetative model in psycholinguistics is adopted by C. E. Osgood (1963, 1968). His thoughts, in particular those advanced in his last work, have a more general impact on the psycholinguistic theory altogether. This primarily refers to his clear formulation of the difference between the language model and the LP model: The model of the speaker is a finite, a time dependent and a context-dependent device, whereas a model of the language (a grammar) is infinite, time-independent and relatively context-free device. In performance models of the speaker the semantic component is the central and the syntactic component operates on its output.

Another two properties characteristic of LP according to Osgood should be added to the above differences: 1. Language behaviour includes sequential processes and Markov-type transitional dependencies. 2. The LP models require selection among altenatives. Upon this Osgood agrees with Leontev and Rommetveit, who — beside other authors — also argue that in addition to other factors the LP model should include the probability—sequence mechanism and mechanism of selection playing a part primarily in operation at the semantic level of the encoding and decoding processes.

In *conclusion*, it may accordingly be stated:

1. In the contemporary theory of psycholinguistics the existing conceptions and models of the basic processes of language behaviour display considerable differences. The period of marked preference for the Chomsky—Miller generative model is succeeded by a strong wave of theoretical and experimental criticism of the model. Serious shortcomings of the model are being revealed on the one hand and pre-conditions for more adequate models of language behaviour are developing on the other. The interest is shifted again towards the probability-sequence processes in LP and away from the grammatical, formal factors towards semantic, pragmatic and contextual factors. The two different types, however, of LC and LP models (i.e. "generative models" and "probability models") developed so far have not reached beyond the stage of hypotheses in spite of the amount of experimental data assembled and they demand further exact and systematic verification.

2. A more adequate explanation of the language behaviour requires a change in the basic approach in the production of psycholinguistic models: so far psycholinguistics has mostly developed by submitting linguistic (grammatical) models to an experimental psychologic examination[11] without making use of the inverse approach consisting in the production of psycholinguistic models on the basis of psychological

knowledge and criteria and their confrontation with linguistic theories. On the other hand it is true that no use has been made in psycholinguistics of the knowledge of some "non-grammatical" linguistic areas, in particular of the new stylistic and sociolinguistic theories.

3. Existing psycholinguistic models endeavour to account primarily for the process of decoding messages, i.e. they represent primarily models of the listener. It is advisable that the study of the processes of the actual production of messages, i.e. the study of the speaker, should be developed more intensely.[12]

Under the circumstances, we fully agree with A. A. Leontev (169, p. 4) that "it is high time that psycholinguistics should take up a serious theoretical re-assessment of the achieved results and proceed to a new, higher plane on which psycholinguistics would cease to be an isolated area related to the mathematical theory of generative grammars rather than to psychology, and to structural linguistics more than to sociology, and would find its place within the general synthesis of the studies of man."[13]

REFERENCES

AMMON, P. R.: The Perception of Grammatical Relations in Sentences: A Methodological Exploration. JVLVB, *7*, 1968, pp. 869—875.

BEVER, T. G.: Associations to Stimulus-Response Theories of Language. In: Verbal Behavior and General Behavior Theory, New York, Englewood Cliffs 1968, pp. 478—494.

BLUMENTHAL, A. L.: Prompted Recall of Sentences. JVLVB, *6*, 1967, pp. 203—206.

CHOMSKY, N.: Current Issues in Linguistic Theory. The Hague 1964.

CHOMSKY, N.: Aspects of the Theory of Syntax. Cambridge, Mass. 1965.

CHOMSKY, N. — MILLER, G. A.: Introduction to the Formal Analysis of Natural Languages. In: Handbook of Mathematical Psychology, vol. 2, New York, 1963, pp. 269—321.

CLARK, H. H.: Some Structural Properties of Simple Active and Passive Sentences. JVLVB, *4*, 1965, pp. 365—370.

CLARK, H. H. and BEGUN, J. S.: The Use of Syntax in Understanding Sentences. British Journal of Psychology, *59*, 1968, pp. 219—229.

[11] "But the assumption underlying much recent work in psycholinguistics, that the study of the form of grammars is the right place to begin the study of performance models, is surely misleading. As long as this assumption is made we are in danger of asking questions to which there are no answers" (Thorne, 1966, p. 10).

[12] Cf. Fromkin's (1968) suggestion on general requirements for a model of performance and specific requirements for a model of speech performance.

[13] When reading the present text at the end of 1973 I must state that still many "psycholinguists are finding that their colourless green ideas sleep furiously while they toss and turn in their self-made Procrustean grammatical beds" (K. Salzinger), while some others are moving to the study of "language in action and context" as it is manifested by the new discip ine — *psycho-sociolinguistics* [see Průcha 1972).

CLIFTON, Ch.—KURCZ I. — JENKINS, J. J.: Grammatical Relations as Determinants of Sentence Similarity. JVLVB, *4*, 1965, pp. 112—117.

COLEMAN, E. B.: The Comprehensibility of Several Grammatical Transformations. Journal of Applied Psychology, *43*, 1964, pp. 186—190.

DANKS, J. H.: Grammaticalness and Meaningfulness in the Comprehension of Sentences. JVLVB, *8*, 1969, pp. 687—696.

EPSTEIN, W.: Recall of Word Lists Following Learning Sentences and of Anomalous and Random Strings. JVLVB, *8*, 1969, pp. 20—25.

FODOR, J. — GARRET, M.: Some Reflections on Competence and Performance: In: Psycholinguistics Papers, Edinburgh 1966, pp. 135—154.

FORSTER, K. J.: Left-to-Right Processes in the Construction of Sentences. JVLVB, *5*, 1966, pp. 285—291.

FROMKIN, V.: Speculations on Performance Models. Journal of Linguistics, *4*, 1968, pp. 47—68.

HARMON, G. H.: Psychological Aspects of the Theory of Syntax. The Journal of Philosophy, *64*, 1967, pp. 75—87.

ILJASOV, I.: Eksperiment Dž. Millera po proverke psichologičeskoj realnosti transformacionnoj modeli. In: Psichologija grammatiki, Moskva 1968, pp. 50—66.

JONCKHEERE, A. R.: Discussion in: Psycholinguistics Papers, Edinburgh 1966, pp. 84—89.

JOHNSON, N. F.: Sequential Verbal Behavior. In: Verbal Behavior and General Behavior Theory, New York, Englewood Cliffs 1968, pp. 421—450.

KATZ, J. J.—POSTAL, P. M.: An Integrated Theory of Linguistic Descriptions. Cambridge, Mass. 1964.

LEONTEV, A. A.: Psicholingvistika. Leningrad 1967.

LEONTEV, A. A.: Psicholingvističeskaja značimost transformacionnoj porožd ajuščej modeli. In: Psichologija grammatiki, Moskva 1968, pp. 5—49.

LEONTEV, A. A.: Psicholingvističeskije jedinicy i porožd enije rečevogo vyskazyvanija. Moskva 1969.

LEVELT, W. J. M.: The Perception of Syntactic Structure. Mimeo, Heymans Bulletins Psychologische Instituten R. U. Groningen. NR: HB — 69 — 30 Ex., 1969.

LOEWENTHAL, K.: Discussion in: Psycholinguistics Papers, Edinburgh 1966, pp. 93—94.

MARKS, L. E.: Some Structural and Sequential Factors in the Processing of Sentences. JVLVB, *6*, 1967, pp. 707—713.

MARKS, L. E.: Judgments of Grammaticalness of Some English Sentences and Semi-Sentences. American Journal of Psychology, *80*, 1967, pp. 196—204.

MARTIN, J. E.: Semantic Determinants of Preferred Adjective Order. JVLVB, *8*, 1969, pp. 697—704.

MARTIN J. E. — ROBERTS, K. H.: Grammatical Factors in Sentence Retention. JVLVB, *5*, 1966, pp. 211—218.

MARTIN, J. E.—ROBERTS, K. H. — COLLINS, A. M.: Short-Term Memory for Sentences. JVLVB, *7*, 1968, pp. 560—566.

MEHLER, J.: Some Effects of Grammatical Transformations on the Recall of English Sentences. JVLVB, *2*, 1963, pp. 346—351.

MEHLER, J.—MILLER, G. A.: Retroactive Interference in the Recall of Simple Sentences. British Journal of Psychology, *54*, 1964, pp. 295—301.

MEHLER, J.—CAREY, P.: Role of Surface and Base Structure in the Perception of Sentences. JVLVB, *6*, 1967, pp. 335—338.

MILLER, G. A.: Some Psychological Studies of Grammar. American Psychologist, *17*, 1962, pp. 748—762.

MILLER, G. A.—CHOMSKY, N.: Finitary Models of Language Users. In: Handbook of Mathematical Psychology, Vol. 2, New York 1963, pp. 419—491.
MILLER, G. A.—GALANTER, E. — PRIBRAM, K.: Plans and the Structure of Behavior, New York 1960.
MILLER, G. A.—McKEAN, K.: A Chronometric Study of Some Relations Between Sentences. Quarterly Journal of Exp. Psych., *16*, 1964, pp. 297—308.
MILLER, G. A.—McKEAN, K.—SLOBIN, D.: The Exploration of Transformations by Sentence Matching. In: MILLER, G. A.: Some Psychological Studies of Grammar, 1962.
MORRIS, V. A.—RANKINE, F. C.—REBER, A. S.: Sentence Comprehension, Grammatical Transformations and Response Availability. JVLVB, *7*, 1968, pp. 1113—1115.
OSGOOD, Ch. E.: On Understanding and Creating Sentences. American Psychologist, *18*, 1963, pp. 735—751.
OSGOOD, Ch. E.: Toward a Wedding of Insufficiencies. In: Verbal Behavior and General Behavior Theory, Englewood Cliffs 1968, pp. 495—520.
PRŮCHA' J.: Psycholinguistics and Sociolinguistics — Separate or Integrated? International, Journal of Psycholingustics, *1*, 1972, pp. 9—23.
ROBERTS, K. H.: Grammatical and Associative Constraints in Sentence Retention. JVLVB, *7*, 1968, pp. 1072—1076.
ROMMETVEIT, R. et al.: Processing of Utterances in Context. Mimeo, Univ. of Louvain 1967.
ROMMETVEIT, R.: Words, Meanings and Messages, New York—London—Oslo 1968.
ROMMETVEIT, R.: Studies on Context Effects in Verbal Message Transmission. In: Experimental Social Psychology (Papers and Reports from the Intern. Conference on Social Psychology, Prague 1968), Prague 1969.
ROSENBERG, S.: Association and Phrase Structure in Sentence Recall. JVLVB, *7*, 1968, pp. 1077—1081.
SACHS, J. D. S.: Recognition Memory for Syntactic and Semantic Aspects of Connected Discourse. Perception and Psychophysics, *2*, 1967, pp. 437—442.
SALZINGER, K.—ECKERMAN, C.: Grammar and the Recall of Chains of Verbal Responses. JVLVB, *6*, 1967, pp. 232—239.
SAVIN, H. B.—PERCHONOCK, E.: Grammatical Structure and the Immediate Recall of English Sentences. JVLVB, *4*, 1965, pp. 348—353.
SCHLESINGER, J. M.: A Note on the Relationship between Psychological and Linguistic Theories. Foundations of Language, *3*, 1967, pp. 397—402.
SLAMA-CAZACU, T.: Introducere in psiholingvistică. București 1968.
SLOBIN, D.: Recall of Full and Truncated Passive Sentences in Connected Discourse. JVLVB, *7*, 1968, pp. 876—881.
SUTHERLAND, N. S.: Discussion in: Psycholinguistics Papers, Edinburgh 1966, pp. 154—163.
TANNENBAUM, P. H.—EVANS, R. R.—WILLIAMS, F.: An Experiment in the Generation of Simple Sentence Structure. Journal of Communication, *14*, 1964, pp. 113—117.
TAYLOR, I.: Content and Structure in Sentence Production. JVLVB, *8*, 1969, pp. 170—175.
THORNE, J. P.: On Hearing Sentences. In: Psycholinguistics Papers, Edinburgh 1966, pp. 3—10.
UHLENBECK, E. M.: An Appraisal of Transformation Theory. Lingua, *12*, 1963, pp. 1—18.
UHLENBECK, E. M.: Some Further Remarks on Transformational Grammar. Lingua, *17*, 1967, pp. 263—316.
WALES, R. J.—MARSHALL, J. C.: The Organization of Linguistic Performance. In: Psycholinguistics Papers, Edinburgh 1966, pp. 29—80.

On a Measure of Divergence of a Context-free Language from Finite State Languages

LÁSZLÓ KALMÁR, SZEGED

1. It is usual to consider a language as an entity given by two data: its (by Chomsky [1] so-called terminal) *vocabulary* V, i.e. the set of its words (more exactly, word forms), and the set S of its (grammatical) *sentences*. Here, V is a non-empty finite set whereas S is considered, in general, as infinite, each element of which (each sentence) being a finite word string. Or, in mathematicians' dialect, a language is defined as an ordered pair $L = \langle V, S \rangle$ formed of a non-empty finite set V and a subset S of the free semigroup $F(V)$ generated by V (which means the set of all finite strings formed of elements of V, equipped with some algebraic structure).

In this paper, I shall use a more general conception of language which leads to the following mathematical definition: a *language* is an ordered triple $L = \langle V, C, f \rangle$ formed of two disjoint, non-empty, finite sets V and C (called *vocabulary* and *set of* (linguistic) *category symbols* respectively) and an application f of the set C into the set of all subsets of $F(V)$. Philosophically, this conception means that I consider, besides the question, which word strings are sentences of L, for each linguistic category c which is applicable for the language L (and symbols for such categories are the elements of C), also the question, which word strings belong to the category of which c is a symbol (e.g. which word strings are noun groups, or possible predicates of a sentence), as a question referring to the language L itself, rather than to some of its grammars. (As it is well-known, the same language can be described by different grammars, hence, a grammar is always arbitrary to a certain extent, whereas a language, at least a natural one, belongs to the objective reality.) In the above mathematical model, f is the mapping which makes correspond to each category symbol c which is an element of C the set of strings formed of elements of V belonging to the category symbolized by c. Obviously, in the case if C is formed of a single element s (symbolizing the category of sentences), we get, setting $S = f(s)$, the usual conception as a particular case.

2. The notion of a finite state language can be easily fitted our more general conception by using, instead of Chomsky's finite state diagrams, more general ones which I shall call *finite state flag diagrams*. Informally speaking, a finite state flag diagram is an oriented graph, some edges of which are marked by some word v (belonging to V),

as "v-edges", whereas some vertices of which are marked, in one of two different ways, by some category symbol c (belonging to C), viz. either as "starting c-vertices" or "ending c-vertices". More exactly, a finite state flag diagram is defined as an ordered sextuple $D = \langle V, C, H, f_1, f_2, g\rangle$ formed of two disjoint, non-empty, finite sets V and C (called as above, vocabulary or set of words, and set of category symbols, respectively), a finite oriented graph H and three mappings f_1, f_2, g of two disjoint, non-empty subsets P_1 and P_2 of the set P of points (vertices) of H and a non-empty subset E_1 of the set E of edges of H onto C onto C and onto V, respectively. ($f_1 : P_1 \rightarrow \rightarrow C; f_2 : P_2 \rightarrow C; g : E_1 \rightarrow V$.) A vertex $p_1 \in P_1$ with $f_1(p_1) = c \in C$ is called a *starting c-vertex*; a vertex $p_2 \in P_2$ with $f_2(p_2) = c \in C$ is called an *ending c-vertex*; an edge $e \in E_1$ with $g(e) = v \in V$ is called a *v-edge*. An (oriented, possibly self-intersecting) path Q of H (i.e. going possibly several but a finite number of times through the same point or/and edge) is called a *c-path* ($c \in C$) if it leads from a starting c-vertex to an ending c-vertex but otherwise, it does not go through any starting or ending c-vertex.

Let be $e_1, e_2, \ldots, e_n$ the edges belonging to E_1 of a c-path Q of H, each written as many times as Q goes through it and written in the order in which Q goes through them. Let $v_i = g(e_i)$ for $i = 1, 2, \ldots, n$. Then the string $v_1 v_2, \ldots, v_n$ is called the *word string to be read along Q*. We call the language $L = \langle V, C, f\rangle$, where, for any $c \in C$, $f(c)$ is the set of all word strings to be read along some c-path of H, the *language represented by the finite state flag diagram* $D = \langle V, C, H, f_1, f_2, g\rangle$. A *finite state language*, in our conception, is a language which is represented by some finite state flag diagram. E.g. Fig. 1 indicates a finite state flag diagram representing a well-

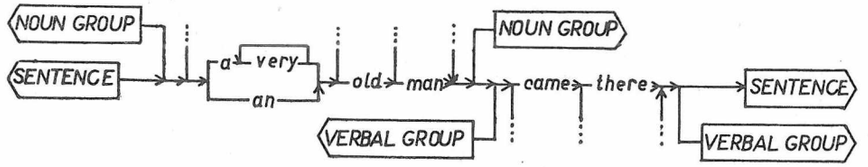


Fig. 1. Example of a flag diagram representing a finite state language.

known finite state sub-language of English with an infinity of sentences, whereas Fig. 2 indicates a finite state flag diagram representing French noun and adjective morphology as a finite state language (with the set of French noun and adjective stem and desinence morphemes as V, and $C = \{$FORME NOMINALE, FORME DE SUBSTANTIF, FORME D'ADJECTIF$\}$) used, according to an oral communication of Professor Vauquois, in programming the French morphological synthesis within the Grenoble system of Russian-French automatic translation. (In our figures, a v-edge is indicated by inscription of the symbol v in an edge; a starting and an ending c-vertex is indicated by marking a vertex either by a "flag head" or by the

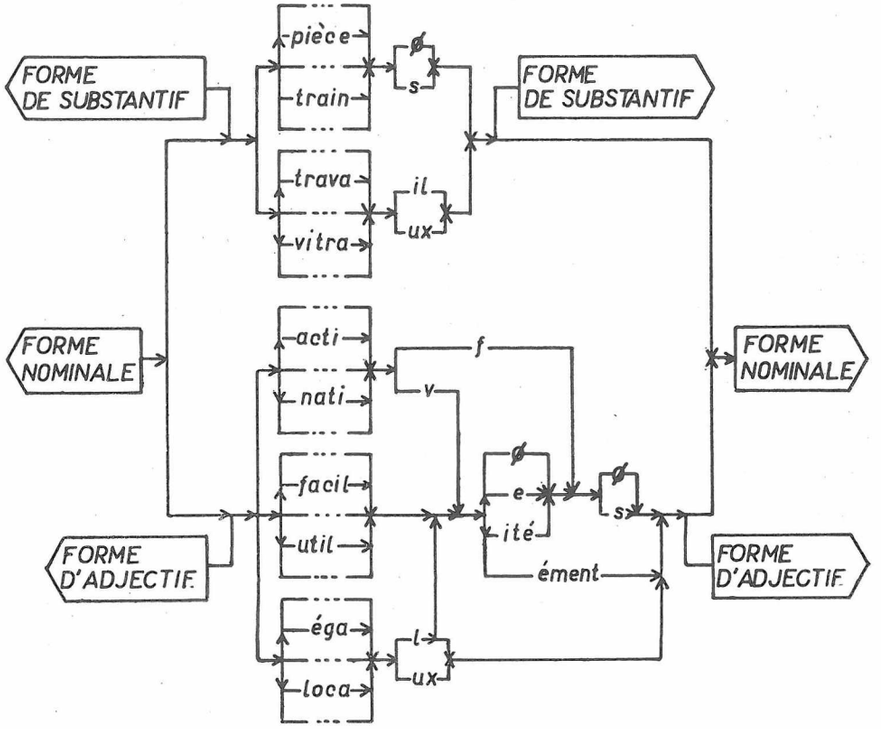


Fig. 2. Flag diagram representing French noun and adjective morphology.

"handle of the staff of a flag", bearing the symbol c, pointing to the left and to the right, respectively. The "flagstaff" is not to be considered as an edge of the graph.)

3. The notion of a language generated by a context-free grammar is fitted immediately our conception of language. For this purpose, one has only to omit the distinguished category (that of sentences) from the definition of a *context-free grammar*; i.e. to define the latter as an ordered triple $G = \langle V, C, R\rangle$ formed of two disjoint, non-empty, finite sets V and C (called again vocabulary and set of category symbols or, in this connection, according to Chomsky rather terminal and non-terminal vocabulary, respectively) and a non-empty, as a rule, finite, subset of the Cartesian product of C with the free semigroup $F(V \cup C)$ generated by the union of V and C. The elements of R, which are of the form $\langle c, \sigma\rangle$, written here as "$c : \sigma$" as in the presentation of the programming language ALGOL 68 [2] rather than "$c \rightarrow \sigma$" in Chomsky, where c is a category symbol and σ a "mixed string" formed of elements of the union of V and C, are called (rewriting, or production) *rules*. As usual, a mixed string σ is called a *direct production* of a category symbol c if $c : \sigma$ is a rule; and *productions* of a category symbol c are defined recursively as (i) its direct productions and (ii)

mixed strings formed of productions of c by replacing a category symbol c by any of its direct productions. The *language L generated by a context-free grammar $G = \langle V, C, R\rangle$* is defined as $L = \langle V, C, f\rangle$ where, for any category symbol $c \in C$, $f(c)$ is the set of all terminal productions of c, i.e. of all of its productions which are elements of $F(V)$ (i.e. strings containing no category). A language is a *context-free language* if it is generated by some context-free grammar.

4. As it is well-known, there are context-free languages which are not finite state languages. A simple example is the language whose sentences are the strings of the form $a^n bc^n$, a^n denoting a string of n words each of which is a, or, as I prefer to say, the language $L_0 = \langle V_0, C_0, f_0\rangle$ with $V_0 = \{(, O,)\}$, $C_0 = \{E\}$, $f_0(E)$ being the set o "expressions" O, (O), $((O))$, ... A context-free grammar generating L_0 is $G_0 = \langle V_0, C_0, R_0\rangle$ with $R_0 = \{r_1, r_2\}$, r_1 and r_2 being the rules $E : O$ and $E : (E)$, respectively.

Intuitively, L_0 is a non-finite state language whose "divergence" from finite state languages is minimal, in any case less than e.g. in the case of the language $L_1 = \langle V_1, C_1, f_1\rangle$ of the Church lambda conversion (see [3]; or rather that of the lambda-K conversion, see [4]) with $V_1 = \{x, |, \{, \}, (,), \lambda, [,]\}$, $C_1 = \{\text{VARIABLE, FORMULA}\}$, generated by the context-free grammar $G_1 = \langle V_1, C_1, R_1\rangle$ with $R_1 = \{r_1, r_2, r_3, r_4, r_5\}$, r_1, r_2, r_3, r_4, r_5 being the rewriting rules

VARIABLE : x
VARIABLE : VARIABLE
FORMULA : VARIABLE
FORMULA : {FORMULA} (FORMULA)
FORMULA : λ VARIABLE [FORMULA]

respectively.

For an exact definition of the notion of *divergence of a context-free language from finite state languages* underlying this intuition, we need a generalization of the notion of a flag diagram which allows representation of context-free languages in general. Informally speaking, the generalization consists in allowing to mark edges of the oriented graph H by category symbols. An edge marked by $c \in C$ stands for any subgraph of H connecting a starting c-vertex by an ending c-vertex; in the case this subgraph contains again an edge marked by a category symbol, this "standing for" is to be meant recursively. The idea of such a generalization, with statistical applications, can be found in Brodda [5]; on the other hand, generalized flag diagrams are used at the Department of Foundations of Mathematics and Computer Science of the Attila József University, Szeged, since several years, for tutorial purposes, viz. in teaching programming language syntaxes, e.g. that of ALGOL 60 and, in the last years, also of the metalanguage of ALGOL 68; see the figures appended to the present paper. (In the last figure, a v-edge is indicated by writing the symbol v *above* an edge.) See also Kalmár [6].

5. We define a *fiag diagram* in general, more exactly, as an ordered septuple $D = \langle V, C, H, f_1, f_2, g_1, g_2\rangle$ formed of two disjoint, non-empty, finite sets V and C (called again vocabulary and set of category symbols, respectively), a finite oriented graph H, and four mappings f_1, f_2, g_1, g_2 of two disjoint, non-empty subsets P_1 and P_2 of the set P of vertices of H and of two disjoint subsets E_1 and E_2 of the set E of edges of H, of which E_1 is non-empty, onto C, onto C, onto V and onto C, respectively. ($f_1 : P_1 \to C; f_2 : P_2 \to C; g_1 : E_1 \to V; g_2 :$ $E_2 \to$ into C.) *Starting* and *ending c-vertices* for $c \in C$ are defined literally as in the case of finite state flag diagrams. An edge $e_1 \in E_1$ with $g_1(e_1) = v \in V$ is called a *v-edge*; an edge $e_2 \in E_2$ is called a *recursive edge* and, in particular, a *c-edge* if $g_2(e_2) = c \in C$. For $c \in C$ a *c-path* is defined now as an (oriented, possibly self-intersecting) path Q of H leading from a starting c-vertex to an ending c-vertex but otherwise not going through any starting or ending c-vertex and not containing any recursive edge. The *word string to be read along a c-path Q* is defined analogously to the case of a finite state flag diagram, viz. as the string $v_1 v_2, \ldots, v_n$ with $v_i = g_1(e_i)$ for $i = 1, 2, \ldots, n$, $e_1, e_2, \ldots, e_n$ denoting the edges belonging to E_1 of V, each written as many times as Q goes through it and written in the order in which Q goes through them.

To define the language represented by a flag diagram, we need the notion of the derivatives of a flag diagram D. Given a flag diagram $D = \langle V, C, H, f_1, f_2, g_1, g_2\rangle$, we define first, for $c \in C$, a *c-graph of D* as a subgraph of H connecting any starting c-vertex p_1 with any ending c-vertex p_2, i.e. the graph having as vertices all vertices of H to which at least one (oriented) path of H leads from p_1 and from which at least one (oriented) path of H leads to p_2 and as edges all edges of H which connect such vertices. (If there is no such vertex of H, then the subgraph of H connecting p_1 with p_2 is the empty graph; otherwise, both p_1 and p_2 are vertices of this subgraph.) The *derivatives of the flag diagram D* are defined recursively as (i) D itself and (ii) any flag diagram which can be obtained from some derivative of D by replacing one of its c-edges ($c \in C$) by any c-graph of D. Here, replacement of a c-edge e of a derivative $D' = \langle V, C, H', f_1', f_2', g_1', g_2'\rangle$ of D by a c-graph $\overline{H}$ of D is to be understood as follows. Suppose, $\overline{H}$ is the subgraph of H connecting the starting c-point p_1 with the ending c-point p_2; and e is an edge of H leading from the vertex p_3 to the vertex p_4. First, modify $\overline{H}$ by replacing p_1 by p_3 and p_2 by p_4; then insert the graph obtained thus instead of e between the vertices p_3 and p_4 of H'. Thus, we obtain an oriented graph H'' each vertex of which is either a vertex of H' or a vertex of $\overline{H}$ and the same holds for its edges. Now, define, for vertices p and edges e of H'',

$$f_1''(p) = \begin{cases} f_1(p) \text{ if } p \text{ is a vertex of } \overline{H} \text{ and } f_1(p) \text{ is defined,} \\ f_1'(p) \text{ if } p \text{ is a vertex of } H' \text{ and } f_1'(p) \text{ is defined,} \\ \text{undefined otherwise;} \end{cases}$$

$$f_2''(p) = \begin{cases} f_2(p) \text{ if } p \text{ is a vertex of } \overline{H} \text{ and } f_2(p) \text{ is defined,} \\ f_2'(p) \text{ if } p \text{ is a vertex of } H' \text{ and } f_2'(p) \text{ is defined,} \\ \text{undefined otherwise;} \end{cases}$$

$$g_1''(e) = \begin{cases} g_1(e) \text{ if } e \text{ is an edge of } \overline{H} \text{ and } g_1(e) \text{ is defined,} \\ g_1'(e) \text{ if } e \text{ is an edge of } H' \text{ and } g_1'(e) \text{ is defined,} \\ \text{undefined otherwise;} \end{cases}$$

$$g_2''(e) = \begin{cases} g_2(e) \text{ if } e \text{ is an edge of } \overline{H} \text{ and } g_2(e) \text{ is defined,} \\ g_2'(e) \text{ if } e \text{ is an edge of } H' \text{ and } g_2'(e) \text{ is defined,} \\ \text{undefined otherwise,} \end{cases}$$

where, e.g., "$f_1(p)$ is defined" means $p \in P_1$. Then the flag diagram $D'' = \langle V, C, H'', f_1'', f_2'', g_1'', g_2'' \rangle$ is understood as the result of the replacement in question. (By induction, one verifies that for all derivatives of D, the vocabulary and the set of category symbols are the same; hence, the above definition of replacement is complete for our purposes, i.e. we do not need to consider the more general case $D' = \langle V', C', H', f_1', f_2', g_1', g_2' \rangle$.)

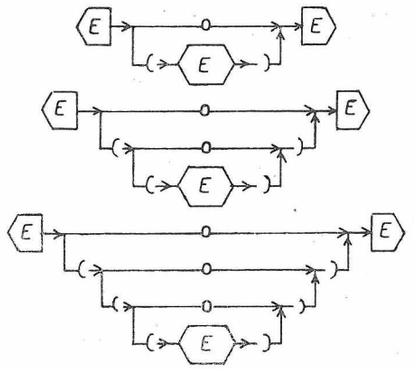


Fig. 3. A flag diagram representing a non-finite state language, and some of its derivatives.

Now, we define the language represented by a flag diagram $D = \langle V, C, H, f_1, f_2, g_1, g_2 \rangle$ as the language $L = \langle V, C, f \rangle$ where, for any $c \in C$, $f(c)$ is defined as the set of all word strings to be read along some c-path of the oriented graph H of some derivative $D' = \langle V, C, H', f_1', f_2', g_1', g_2' \rangle$ of D. E.g. Fig. 3 indicates a flag diagram representing the above context-free language L_0 which is not a finite state language as well as some of its derivatives, whereas Fig. 4 besides repeating the rewriting rules of the above context-free language L_1, indicates two different flag diagrams representing L_1. (In our figures, a c-edge is indicated by inserting a "double flag head", i.e. a hexagon with two horizontal and four slant sides which form two angles, pointing to the left and to the right, respectively, and bearing the symbol c, into an edge, reminding that, in forming the derivatives of the flag diagram, such an edge has to be replaced by a subgraph connecting some vertex marked by a flag pointing to the left with some vertex marked by a flag pointing to the right, both bearing the symbol c.)

6. One readily shows that *any context-free language can be represented by some flag diagram*. Indeed, consider any context-free language L, and a context-free grammar $G = \langle V, C, R \rangle$ which generates L. Each rewriting rule $r \in R$ has the form $c_r : s_{r1} s_{r2}, \ldots, s_{rn_r}$ with $c_r \in C$ and, for $i = 1, 2, \ldots, n_r$, $s_{ri} \in V \cup C$, i.e. either $s_{ri} \in V$, or $s_{ri} \in C$. To each such rule r, form an oriented graph H_r with $n_r + 1$ different vertices $p_{r0}, p_{r1}, \ldots, p_{rn_r}$ and n_r edges $e_{r1}, e_{r2}, \ldots, e_{rn}$, where, for $i = 1, 2, \ldots, n_r$, e_{ri} leads from $p_{r,i-1}$ to p_{ri}. Also, for $r \in R$, $r' \in R$, $r \neq r'$, any two vertices p_{ri} and $p_{r'j}$ have to be different. From the union of these oriented graphs H_r for all rules $r \in R$, mark, for

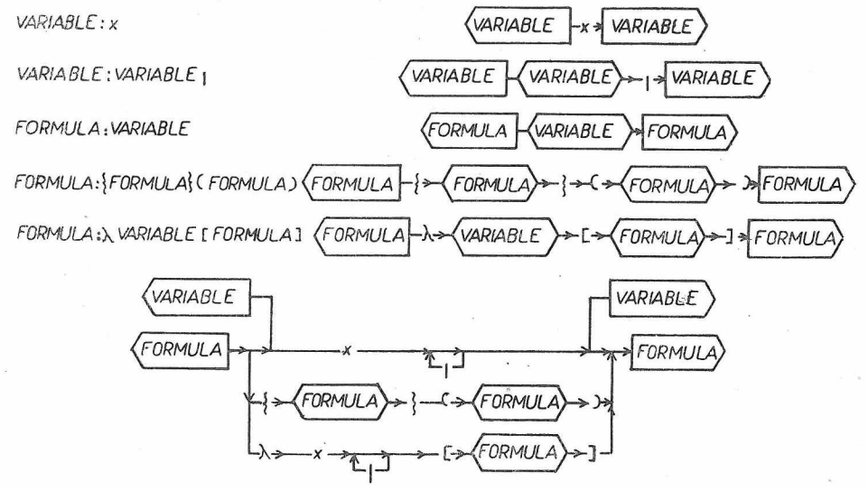


Fig. 4. Context-free rewriting rules translated into a flag diagram, and a simpler flag diagram representing the same language.

each $r \in R$, p_{r0} as a starting c_r-vertex, p_{rn_r} as an ending c_r-vertex and, for $i = 1, 2, \ldots, n_r$, the edge e_{ri} as an s_{ri}-edge (i.e., for $s_{ri} = v \in V$ as a v-edge and for $s_{ri} = c \in C$ as a c-edge). Thus, we obtain a flag diagram which, as readily shown, represents the language L. (On the other hand, one can show that any language represented by a flag diagram is a context-free language; however, we do not need this fact here.) E.g. the upper part of Fig. 4 indicates, how one can "translate" the rewriting rules for language L_1 into a flag diagram.

7. Now, the finite state flag diagrams can be considered as the particular case of flag diagrams $D = \langle V, C, H, f_1, f_2, g_1, g_2 \rangle$ in which E_2 is empty (hence, g_2 is the empty mapping, viz. that of the empty set E_2 into C and thus, as a trivial data, can be omitted, writing $D = \langle V, C, H, f_1, f_2, g_1 \rangle$ instead of $D = \langle V, C, H, f_1, f_2, g_1, g_2 \rangle$), or, with other words, in which there is no recursive edge. Hence, *the number of the recursive edges can be used as a measure of divergence of a flag diagram from finite state flag diagrams*. For a given context-free language L, *the minimal number of recursive edges in all flag diagrams representing L can be used as a measure of the divergence of L from finite state languages*, or of the "non-finite state character of L".

Using this measure, L_0 is indeed a non-finite state language for which the measure of divergence from finite state languages is the smallest, viz. 1, as shown by the flag diagram indicated by Fig. 2, having a single recursive edge (and the fact that L_0 is not a finite state language, hence, cannot be represented by a flag diagram without any recursive edge).

Unfortunately, no method is known to find, for any given context-free grammar G,

a flag diagram with minimal number of recursive edges, representing the language generated by G. In view of known algorithmically unsolvable problems of algebraic linguistics, one cannot expect a general algorithm for finding such a "minimal" flag diagram for any given context-free grammar G. However, this does not exclude the possibility of partial algorithms, covering important special cases (special classes of context-free grammars), or "approximative" algorithms, leading to "almost minimal" flag diagrams. So far, I cannot report even on partial results in this sense. We have only some empirical knacks for simplifying flag diagrams, among others for lowering the number of their recursive edges, which can be used in some cases. E.g. the flag diagram indicated by the lower part of Fig. 4 has been obtained from that on the upper part of the same figure by such knacks: "contraction" of disconnected components of a flag diagram which connect starting and ending vertices marked by the same category symbol, replacing them by a branching flag diagram, "inserting" a component of a flag diagram connecting a starting c-vertex with an ending c-vertex instead of a c-edge, replacing a flag diagram of a definite kind by another with a "loop" etc. The situation ressembles to the early history of optimisation of switching circuits when only such empirical rules where available. The fact that since that time theoretical results have been reached on that field which have also practical relevance, allows to hope an analogous development in our field as well.

To the appended figures indicating flag diagrams to represent ALGOL 60 (or a part of the meta-language used for definition of ALGOL 68), I remark, that technical and didactical considerations (to have not too big disconnected components, by which parts of the syntax can be visualized which are often used simultaneously) have restricted even usage of the empirical "minimization knacks" mentioned above.

REFERENCES

[1] CHOMSKY, N.: Three models for the description of language. IRE Transactions, *2*, 1956, pp. 113—124.

[2] MAILLOUX, B. J.—PECK, J. E. L.—KOSTER, C. H. A.: ALGOL 68 (Report on the Algorithmic Language ALGOL 68). Ed. A. van Wingaarden. Second printing by the Mathematic Centrum, MR 101, 1969.

[3] CHURCH, A.: A set of postulates for the foundation of logic. Annals of Math., *2*, 33, 1923, pp. 346—366.

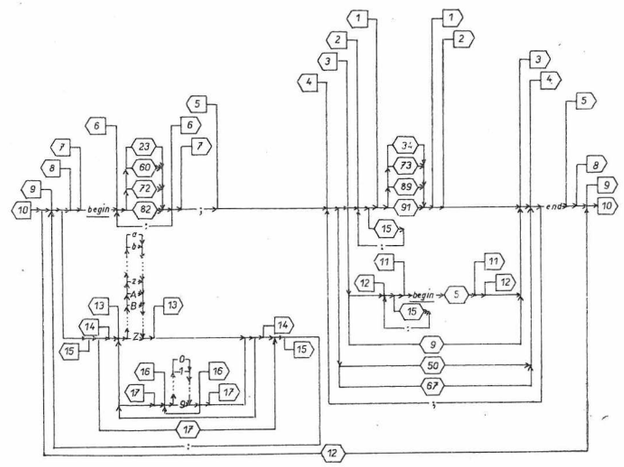
[4] CHURCH, A.: The calculi of lambda-conversion. Princeton 1941.

[5] BRODDA, B.: The entropy of recursive Markov Processes. 2ème Conference Internationale sur le traitement automatique des langues. Grenoble, 25—25 août 1967, No. 38, pp. 1—7, esp. p. 4.

[6] KALMÁR, L.: An intuitive representation of context-free languages. Reprint No. 66, International Conference on Computational Linguistics (COLING) in Sanga-Säby, Stockholm 1969, pp. 1—10.

ALGOL 60

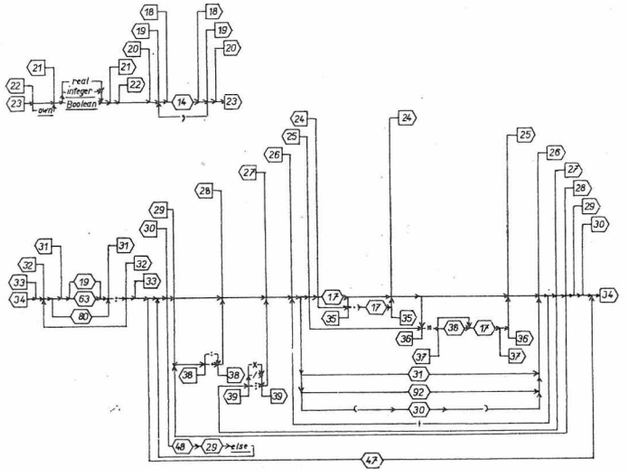
1. {UNLABELLED BASIC STATEMENT}
2. BASIC STATEMENT
3. {UNCONDITIONAL STATEMENT}
4. STATEMENT
5. COMPOUND TAIL
6. DECLARATION
7. BLOCK HEAD
8. {UNLABELLED BLOCK}
9. BLOCK
10. PROGRAM
11. {UNLABELLED COMPOUND}
12. {COMPOUND STATEMENT}
13. LETTER
14. IDENTIFIER
15. LABEL
16. DIGIT
17. {UNSIGNED INTEGER}



PROGRAM, BLOCK, STATEMENTS

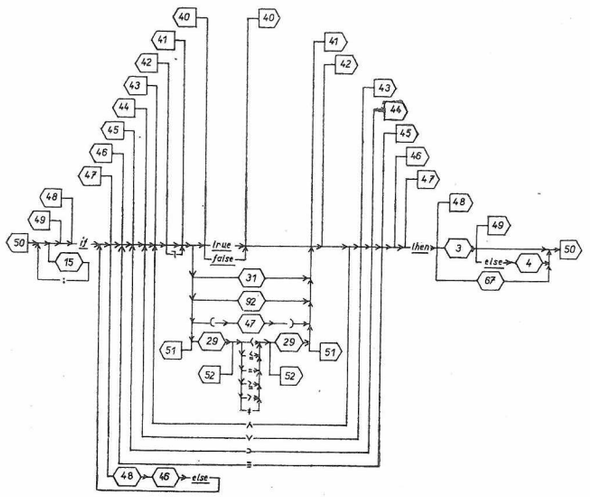
ALGOL 60

18. {VARIABLE IDENTIFIER}
19. SIMPLE VARIABLE
20. TYPE LIST
21. TYPE
22. LOCAL OR OWN TYPE
23. TYPE DECLARATION
24. DECIMAL NUMBER
25. UNSIGNED NUMBER
26. PRIMARY
27. FACTOR
28. TERM
29. {SIMPLE ARITHMETIC EXPRESSION}
30. {ARITHMETIC EXPRESSION}
31. VARIABLE
32. LEFT PART
33. LEFT PART LIST
34. {ASSIGNMENT STATEMENT}

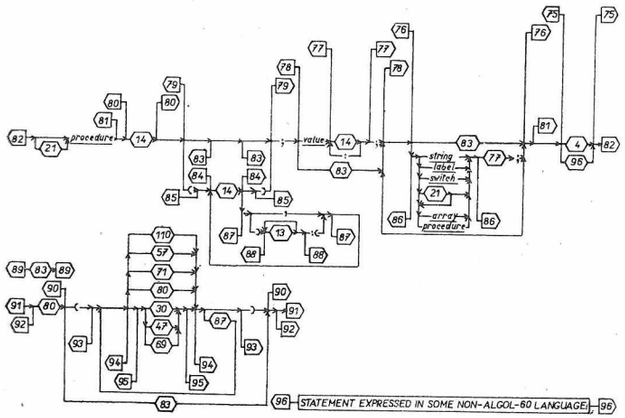


TYPE DECLARATION, ARITHMETIC EXPRESSION, ASSIGNMENT STATEMENT

35. DECIMAL FRACTION
36. EXPONENT PART
37. INTEGER
28. ADDING OPERATOR
39. {MULTIPLYING OPERATOR}



BOOLEAN EXPRESSION, CONDITIONAL STATEMENT



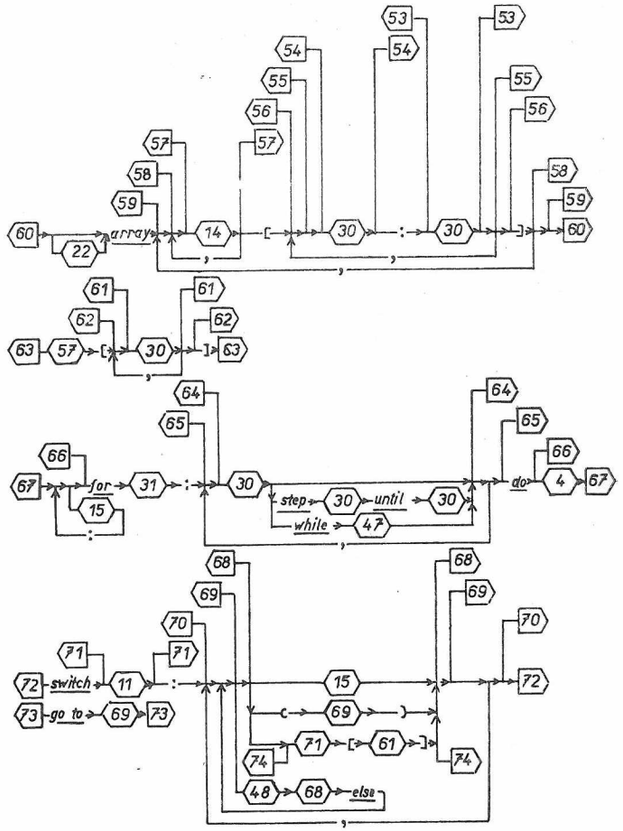
PROCEDURE DECLARATION, DUMY STATEMENT,
PROCEDURE STATEMENT, FUNCTION DESIGNATOR

90. ACTUAL PARAMETER PART
91. PROCEDURE STATEMENT
92. FUNCTION DESIGNATOR
93. ACTUAL PARAMETER LIST
94. ACTUAL PARAMETER
95. EXPRESSION
96. CODE

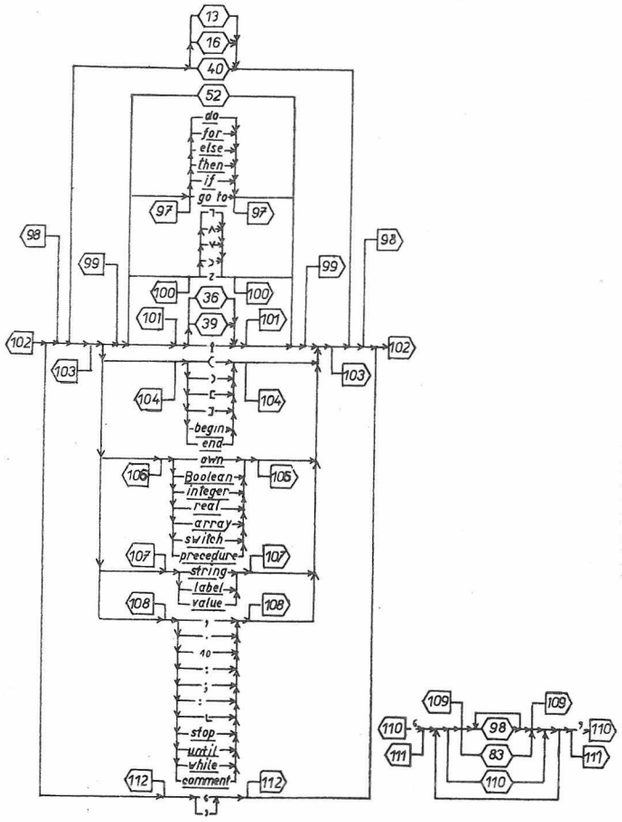
40. LOGICAL VALUE
41. BOOLEAN PRIMARY
42. BOOLEAN SECONDARY
43. BOOLEAN FACTOR
44. BOOLEAN TERM
45. IMPLICATION
46. SIMPLE BOOLEAN
47. BOOLEAN EXPRESSION
48. IF CLAUSE
49. IF STATEMENT
50. CONDITIONAL STATEMENT
51. RELATION
52. RELATIONAL OPERATOR

75. PROCEDURE BODY
76. SPECIFICATION PART
77. IDENTIFIER LIST
78. VALUE PART
79. FORMAL PARAMETER PART
80. PROCEDURE IDENTIFIER
81. PROCEDURE HEADING
82. PROCEDURE DECLARATION
83. EMPTY
84. FORMAL PARAMETER
85. FORMAL PARAMETER LIST
86. SPECIFIER
87. PARAMETER DELIMITER
88. LETTER STRING
89. DUMMY STATEMENT

53. UPPER BOUND
54. LOWER BOUND
55. BOUND PAIR
56. BOUND PAIR LIST
57. ARRAY IDENTIFIER
58. ARRAY SEGMENT
59. ARRAY LIST
60. ARRAY DECLARATION
61. SUBSCRIPT EXPRESSION
62. SUBSCRIPT LIST
63. SUBSCRIPTED VARIABLE
64. FOR LIST ELEMENT
65. FOR LIST
66. FOR CLAUSE
67. FOR STATEMENT
68. SIMPLE DESIGNATIONAL EXPRESSION
69. DESIGNATIONAL EXPRESSION
70. SWITCH LIST
71. SWITCH IDENTIFIER
72. SWITCH DECLARATION
73. GO TO STATEMENT
74. SWITCH DESIGNATOR



ARRAY DECLARATION, SUBSCRIPTED VARIABLE,
ARRAY DECLARATION, SUBSCRIPTED VARIABLE,
SWITCH DECLARATION, DESIGNATIONAL EXPRESSION,
GO TO STATEMENT

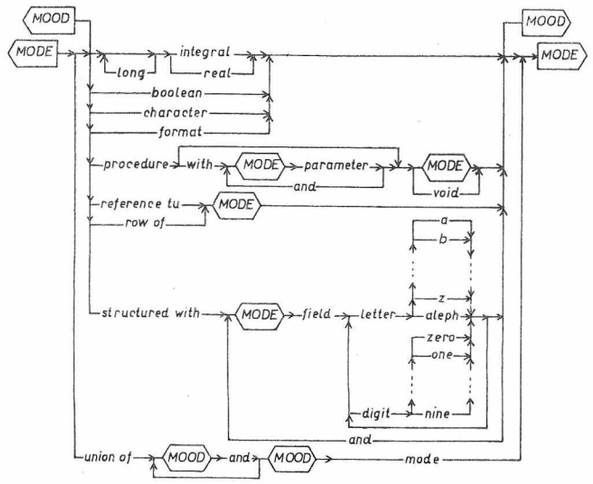


SURVEY OF BASIC SYMBOLS, STRING

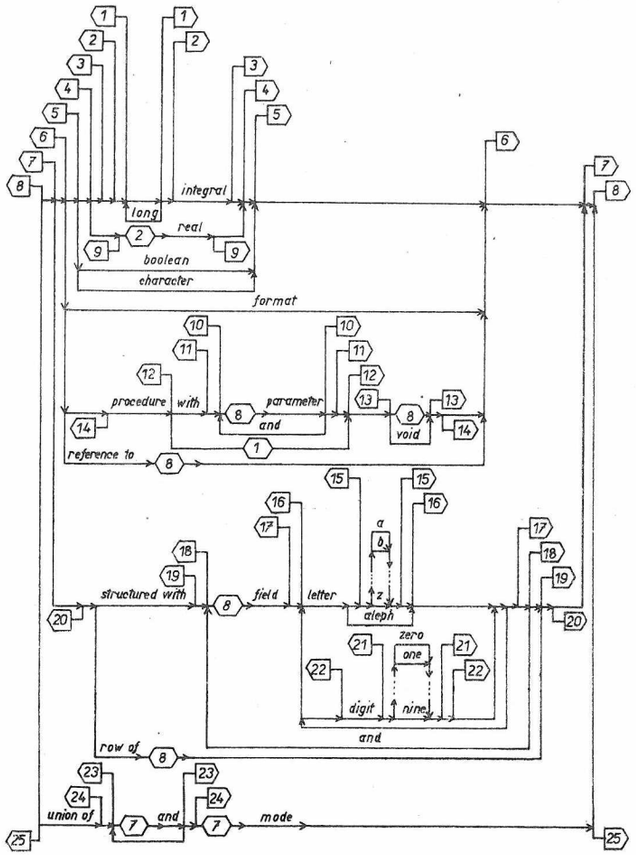
ALGOL 60

97. {SEQUENTIAL OPERATOR
98. {BASIC SYMBOL EXCEPT QUOTES
99. OPERATOR
100. {LOGICAL OPERATOR
101. {ARITHMETIC OPERATION
102. BASIC SYMBOL
103. DELIMITER
104. {BRACKET EXCEPT QUOTES
105. BRACKET
106. DECLARATION
107. SPECIFICATOR
108. SEPARATOR
109. PROPER STRING
110. STRING
111. OPEN STRING
112. QUOTATION MARK

ALGOL 68



SURVEY FLAG DIAGRAM OF MODES



ALGOL 68

1. EMPTY
2. LONGSETY
3. INTEGRAL
4. INTEGRAL
5. PLAIN
6. TYPE
7. MOOD
8. MODE
9. REAL
10. PARAMETER
11. PARAMETERS
12. PARAMETER
13. MOID
14. PROCEDURE
15. ALPHA
16. LETTER
17. TAG
18. FIELD
19. FIELDS
20. STOWED
21. FIGURE
22. DIGIT
23. LMOOD
24. LMOODS
25. UNITED

FLAG DIAGRAM REPRESENTING THE LANGUAGE GENERATED BY THE METAPRODUCTION RULES OF MODES IV ALGOL 68

On Some Connections between the Generative and Analytic Models of Languages

MIROSLAV NOVOTNÝ, BRNO

In this lecture, we shall deal with models of languages. Let V be a finite set, V^* the free monoid over V, i.e. the set of all finite sequences of elements of V (strings) including the empty sequence in which the operation of concatenation is defined. By $|\,x\,|$ we denote the *length* of $x \in V^*$, i.e. the number of symbols of x. Let $L \subseteq V^*$ be an arbitrary set. Then the pair (V, L) is an *analytic model of a language*: the elements of V can be interpreted as word-forms, the elements of V^* as finite sequences of word-forms, the elements of L as correct sentences. If a generative grammar is given which generates (V, L), then (V, L) is called a *generative model of a language*.

For the sake of brevity, we say "*language*" instead of "analytic model of language" and "*constructive language*" instead of "generative model of language". Constructive languages are called languages of the type 0 in the classification of Chomsky which I suppose everybody to be acquainted with. Thus, our problem is to study some connections between constructive languages and languages.

It is well-known that there are languages which are not constructive. We shall describe some constructions which can be applied to each language. The result of such construction is a pair of sets of strings; the finiteness of these sets is a sufficient condition for the constructibility of the given language.

The main concept of our theory is the concept of a configuration which is due to O. S. Kulagina (Probl. Kib. 1, 1958, 203—214) and A. V. Gladkij (Probl. Kib., *10*, 1963, 251—260). Configurations can be defined by means of certain relations on languages.

Let (V, L) be a language.

For $x \in V^*$ we put $xv(V, L)$ (x necessary in (V, L)) if there exist some strings $u, v \in V^*$ such that $uxv \in L$.

For $x, y \in V^*$ we put $x > y\ (V, L)$ (x can be substituted by y in (V, L)) if, for all $u, v \in V^*$, $uxv \in L$ implies $uyv \in L$.

For $x, y \in V^*$ we put $x \equiv y\ (V, L)$ (x equivalent to y in (V, L)) if $x > y\ (V, L)$ and $y > x\ (V, L)$.

The definition of a configuration is the following:

Let $x, y \in V^*$ be strings. Then x is called a *configuration of* (V, L) *with the result* y if the following conditions hold: $xv\ (V, L)$, $x \equiv y\ (V, L)$, $|\ x\ | > |\ y\ |$.

A configuration x of (V, L) is called *short* if the following condition is satisfied: There exists a string $z \in L$ containing a configuration such that the length of each configuration contained in z is $> |\ x\ |$.

A language (V, L) is called *finitely generated* if the set of all strings in L containing no configurations and the set of all short configurations are finite. It can be easily demonstrated that a language (V, L) is finitely generated if there exists such a natural number N that each $x \in L$ with the property $|\ x\ | > N$ contains a configuration of length $\leqq N$.

Example. Let $V = \{v, g, m\}$, let L be the set of all strings of the form xm, where x is an arbitrary string of elements v, g which is terminated by g if it is not empty.

Then, e.g. v^2 is a configuration of (V, L) with the result v, vg is a configuration with the result g, gv is a configuration with the result v, g^2 is a configuration with the result g, but gm is not a configuration with the result m. Then m, gm are the only strings in L containing no configurations, the set of all short configurations is empty. Thus, our language is finitely generated.

We describe some operations with languages: let (V, L) be a language, U a finite set. For each $x \in V^*$, we denote by $t^U_*(x)$ the string which is obtained from x by cancelling all symbols of $V - U$. The language $(U, t^U_*(L))$ is called *the trace of* (V, L) *in* U^*. If (V, L), (W, M) are languages, then the language $(V \cap W, L \cap M)$ is called their *intersection*.

Theorem 1. *Let (V, L) be a finitely generated language, U, W finite sets. Then the intersection of (W, W^*) with the trace of (V, L) in U^* is a language of the type 0 and each language of the type 0 can be constructed in the described way.*

Theorem 1 yields a complete characterization of languages of the type 0 in the terms of the configuration theory. It follows, especially, that each finitely generated language is constructive. Our special construction yields, for each language (V, L), the set of all strings in L containing no configurations and the set of all short configurations. The finiteness of both sets is a sufficient condition for the constructibility of (V, L).

The characterization of languages of the type 3 in the terms of our theory is the following:

Theorem 2. *Let (V, L) be a language. Then (V, L) is of the type 3 if and only there exists such a natural number N that each $x \in V^*$, $xv\ (V, L)$ with the property $|\ x\ | > N$ contains a configuration.*

Clearly, the language of our example is of the type 3.

The concept of a configuration can be modified by substituting the condition $|\ x\ | > |\ y\ |$ by the condition $|\ x\ | > |\ y\ | = 1$. Such configurations will be called *strong configurations*. A strong configuration of the language (V, L) is called *simple* if it contains no strong configuration as a proper substring. A language (V, L) is called a *language of depth 1* if the set of all strings in L containing no strong configurations

and the set of all simple configurations are finite. A finite set U is called *essential with respect to* (V, L) if the result of each simple strong configuration of (V, L) is in U.

Theorem 3. *Let (V, L) be a language of depth 1, U, W finite sets, U essential with respect to (V, L). Then the intersection of (W, W^*) with the trace of (V, L) in U^* is a language of the type 2 and each language of the type 2 can be constructed in the described way.*

This is a complete characterization of languages of the type 2 in the terms of the configuration theory. It follows, especially, that each language of the depth 1 is of the type 2. Our special construction yields, for each language (V, L), the set of all strings in L containing no strong configurations and the set of all simple strong configurations. The finiteness of both sets is a sufficient condition for (V, L) to be of the type 2 which implies the constructibility of (V, L).

The problem of finding a similar complete characterization of languages of the type 1 is open.

It follows from these results that the concept of a configuration belongs to the fundamental concepts of algebraic linguistics: It enables to characterize several important classes of constructive languages without the explicit use of grammars. But the possibilities of using configurations in the study of languages are not exhausted by the theorems I have presented here: It is possible to build up a theory of configurational grammars. Some of my results in this direction have been published in the Publications de la Faculté des Sciences de l'Université J. E. Purkyně Brno, the others are prepared for the print.

On Conditional Context-free Grammars for Programming and Natural Languages*

KAREL ČULÍK, PRAHA

It was shown, e.g. by R. W. Floyd [7], that the set of all programs of the Revized ALGOL 60 [14] which satisfy all further requirements included in sections entitled "Semantics", is no more a context-free language but a context-sensitive one. Therefore, it becomes necessary to start the study of context-sensitive languages which need not mean to start the study of context-sensitive or even more general grammars. This apparent paradox — context-free languages no, but context-free grammars yes — will be solved by the introduction of a new tool of syntactical description of a language consisting in conditions (of different types) under which only simple context-free rules may be applied during the generation process.

1. The necessity of preserving the context-free rules

There are at least three reasons showing the importance of the context-free rules in each grammar which should be used for an actual language, either a programming language or a natural one.

a) Only a *context-free rule* $a_0 \to a_1 a_2 \ldots a_n$ (and no other more general one) can be considered as a *usual* or *ostensive definition* according to whether at least one a_i, $1 \leqq i \leqq n$ is an *auxiliary symbol* or not, i.e. a_i is a *language*-symbol for each $i = 1, 2, \ldots, n$, respectively (of course, a_0 is always an auxiliary symbol). Each auxiliary symbol denotes a set of strings of the language-symbols, see e.g. [3].

Exactly for this reason the context-free rules or, in other words the "Backus normal forms" were used as *syntactical definitions* in the ALGOL 60 Report [14], where the above-mentioned auxiliary symbols were called *metalinguistic variables* and the language-symbols were called *basic-symbols*.

b) Only the context-free rules admit to assign one or more *structural descriptions* to each string of language-symbols generated by them in a very natural and simple form of a *tree-structure*.

* See also in Automatentheorie und formale Sprachen (editors J. Dörr and G. Hotz), Bibliographisches Institut, Mannheim 1970, pp. 209—220.

These structural descriptions are called *parsing trees* in programming languages and *phrase-markers* in natural ones (according to N. Chomsky [1]). According to N. Chomsky the language-symbols are called the *terminal symbols* and the auxiliary symbols are called the *non-terminal* ones. Further both types of rules mentioned in a) are called non-terminal and terminal resp., i.e. the right-hand side of terminal rules consists of terminal symbols only.

The form of *context-sensitive rule* $xa_0y \to xa_1a_2 \dots a_ny$, where x and y are arbitrary strings has been introduced by N. Chomsky probably exactly for its strong similarity to the context-free rule, because again just one single auxiliary symbol is rewritten by it. In fact each context-sensitive rule can be divided into two parts: first of all one context-free rule $a_0 \to a_1a_2 \dots a_n$ and then an ordered pair of strings $[x, :, y]$ called left-neighbouring and right-neighbouring contexts.

N. Chomsky had the possibility of choosing a more general type of rules, e.g. non-shortening ones having the form $v \to w$ where v and w are arbitrary strings such that the length of w is not shorter than the length of v and that v contains at least one auxiliary symbol, because it can easily be shown that the non-shortening grammars (i.e. those using the non-shortening rules only) have the same generative power as the context-sensitive ones. But it is not so important how powerful certain grammars are if they are not suitable for the main purpose to describe fully the given languages. And the non-shortening grammars, although they are convenient for proving some properties, do not allow any useful interpretation and therefore any application.

c) Only the context-free rules admit to introduce a very useful binary relation of *dependency* among the auxiliary symbols, i.e. one writes $a_0 > b_0$ if there are context-free rules $a_0 \to a_1a_2 \dots a_n$ and $b_0 \to b_1b_2 \dots b_m$ such that $b_0 = a_i$ for some i, $1 \leqq i \leqq n$, which leads to a natural *equivalency relation among the auxiliary symbols* (see [3], e.g. in [12] it was shown that ALGOL 60 has exactly two equivalence classes of this sort). In the same way a *binary relation of application of rules* among the rules themselves is introduced in [4] and used as a basis of a homomorphism among the context-free grammars which leads to new classifications of them.

2. Different syntactical tools in grammars

There is a series of different ways how to make the grammars, which use the context-free rules only, more powerful than the usual context-free grammars are. In Prague we have tried to go in the following three directions.

a) In [8] I. Friš has shown that by the *ordered context-free grammars*, i.e. the grammars together with a partial ordering of their rules, the context-sensitive languages may be generated also, when the partial ordering of rules determines which rule must or need not be applied sooner or later than another one.

Thus the partial ordering of rules is one of possible tools but it seems to be rather

complicated to use this tool everywhere because sometimes it is an unproper and unnatural one. Therefore we decided to search for some further tools yet.

In fact, this tool was studied for a formal motivation consisting in carrying over the (total) orderings of rules in Markoff algorithms into the grammars.

The next two tools have another informal motivation as follows. If in a context-free grammar we want to derive from c_i and from c_j occurring in a string $c_1c_2 \dots c_i \dots c_j \dots c_p$, i.e. $1 \leqq i < j \leqq p$, then there is no possibility to control or to coordinate both derivations, although it would very often be desirable. Both these derivations, one starting with c_i and the second with c_j, are completely independent on each other. If $j - i$ is always under a fixed bound, a certain control can be reached by context-sensitive rules, but if $j - i$ can be arbitrarily large, there is no possibility to control this case at all.

b) An idea of a *multiple rule* (being a sequence of finite number of usual simple rules) appeared in [2] and it was applied to finite automata in [5]. In [13] J. Král developed a theory of multiple grammars and he has shown that the *multiple context-free grammars* can generate the context-languages too. Here the required control is determined by a simultaneous application of all simple rules belonging to the same multiple rule. E.g. by the double rule $[c_i \to d, c_j \to d]$ applied to the mentioned string $c_1c_2 \dots c_p$ one gets $c_1c_2 \dots c_{i-1}dc_{i+1} \dots c_{j-1}dc_{j+1} \dots c_p$ etc.

In [6] K. Čulík developed similar ideas for *relational grammars* too.

Thus the multiplicity of context-free rules is the second of possible tools, but although sometimes this tool is very natural it is again suitable only for certain cases.

c) An idea of joining a condition to a context-free rule and to allow to apply this rule just if that condition is satisfied arised naturally as mentioned in b) by cutting the context-sensitive rules into two parts. Of course there is a great freedom as to types of conditions to be introduced and investigated instead of those corresponding to the context-sensitive rules.

In [10] I. Havel studied the most general case when he considered *certain terminating Markoff algorithms* as conditions which were applicable to the strings x and y from the considered string xa_0y and the context-free rule $a_0 \to a_1a_2 \dots a_n$ was applicable to the occurrence a_0 in xa_0y if and only if the result of both algorithms was a void word. It was shown that these conditional grammars are as powerful as possible.

In [15] E. Navrátil and similarly earlier in [8] I. Friš have considered a little less general conditions. There are prescribed two *regular events* L and R to the rule $a_0 \to \to a_1a_2 \dots a_n$ which is applicable to a string xa_0y if and only if $x \in L$ and $y \in R$. It was shown again that these conditional grammars are more powerful than the context-free ones.

In the following much more special conditions are introduced.

3. Different types of conditions

The following types of conditions should be judged from that point of view whether or not they are a proper or even a sufficient tool for a complete syntactical description of natural and programming languages.

a) All the conditions of the first type are called *occurrence conditions* because they consist in several decisions whether or not certain strings are substrings of x and y and where they are located if the considered string is xa_0y and the considered rule $a_0 \rightarrow$ $\rightarrow a_1a_2 \dots a_n$. These conditions are clarified in the following example where just one colon occurs which separates the left and the right part of the condition but several semicolons and commas can occur, which separate certain strings provided with signs + or — (it is assumed that no of the symbols "$:\ ;\ ,\ +\ -$" is language-symbol of the considered language):

$$[+a, -bc; :, +b; -ac, +ab].$$

This condition is satisfied for the string xa_0y if there are strings x_0, x_1 and y_1, y_0 such that $x_0ax_1 = x$ and $by_1aby_0 = y$ and such that $x_1 \neq bct$ for each string t and y_1 does not contain ac as a substring. E.g. $x = baa$ and $y = bcaba$ satisfy our condition but $x = abc$ or $y = cab$ do not. It should be mentioned that a separation by a comma means that both strings must be concatenated but a separation by a semicolon means that between both separated strings may be an arbitrarily long string. Of course + indicates a required occurrence an — a forbidden one.

In [11] I. Havel investigated one-side context-sensitive grammars, i.e. with conditions $[+x, :]$ (or on the contrary $[:, +y]$) only, and he showed that already these grammars are more powerful than the context-free ones.

It is clear that $[+x, :, +y]$ is the occurrence condition joined to the context-free rule $a_0 \rightarrow a_1a_2 \dots a_n$ corresponding to the context-sensitive rule $xa_0y \rightarrow xa_1a_2 \dots a_ny$. Already the condition $[+x; :; +y]$ is essentially different from the previous one, because now it is required that anywhere to the left from a_0 the string x should appear. This difference can be proved as follows.

It is clear what is a negation of a condition and what is a conjunction and a disjunction of two conditions. Therefore a Boolean algebra of conditions may be established. What should be mentioned here is that the negation of $[+a, :]$ if there are e.g. just three different symbols a, b, c, can be expressed as $[-b, :]$ & $[-c, :]$ but it is no more possible in the case $[+a; :]$, the expression of which requires an infinite number of conditions.

b) All the conditions of the second type are called *counting conditions* because they require to count the particular occurrences of strings. Let us restrict ourselves here to a simple special case of them, which will be clarified by the following example where all the previous symbols have the same meaning as in a) and the new symbols $n_1, n_2, \dots$ are either certain fixed positive integers or integer-variables which should

satisfy certain further arithmetical condition. The following four conditions will further be used:

$$[n_1(+a); :; n_2(+b)] \text{ where } n_1 \geqq n_2 + 2, \tag{1}$$
$$[n_1(+a); :; n_2(+b)] \text{ where } n_1 = n_2 + 1, \tag{2}$$
$$[n_2(+b); :; n_3(+c)] \text{ where } n_2 \geqq n_3 + 2, \tag{3}$$
$$[n_2(+b); :; n_3(+c)] \text{ where } n_2 = n_3 + 1. \tag{4}$$

If the condition (1) is joined to a rule $a_0 \rightarrow a_1a_2 \dots a_n$ which should be applied to a string xa_0y, then (1) is satisfied if and only if the number of occurrences n_1 of the symbol a in x and the number of occurrences n_2 of the symbol b in y satisfy the mentioned inequality.

Example. The following conditional context-free grammar $G = \langle V_N, V_T, S, R\rangle$ using some simple occurrence and counting conditions generates a very well-known context-sensitive language $L_G = \{a^nb^nc^n;\ n = 1, 2, \dots\}$:

$$V_N = \{S, A, B, C\},\ V_T = \{a, b, c\},$$
$$R = \{S \rightarrow ABC;\ A \rightarrow Aa;\ A \rightarrow a;\ B \rightarrow Bb \text{ (Cond. 1)};$$
$$B \rightarrow b \text{ (Cond. 2)};\ C \rightarrow Cc \text{ (Cond. 3)};\ C \rightarrow c \text{ (Cond. 4)}\},$$

where

$$\text{(Cond. 1)} = [-A; :] \ \&\ (1),$$
$$\text{(Cond. 2)} = [-A; :] \ \&\ (2),$$
$$\text{(Cond. 3)} = [-A; :] \ \&\ [-B; :] \ \&\ (3),$$
$$\text{(Cond. 4)} = [-A; :] \ \&\ [-B; :] \ \&\ (4).$$

c) Finally all the conditions of the third type are called *length conditions* because they concern the lengths of strings without taking any care of the basic symbols composing these strings. Again let us restrict ourselves here to simple special cases and let us illustrate them in the following example, where the previous symbols are used again as in a) and b) and the new symbols $l_1, l_2, \dots$ are either certain fixed positive integers or integer-variables which should satisfy certain further arithmetical condition:

$$[+a; l_1, +b, : l_2; +a; l_3],$$

where $l_1 \geqq l_2 + l_3$.

If this condition is joined to the rule $a_0 \rightarrow a_1a_2 \dots a_n$ which should be applied to the string xa_0y, then this condition is satisfied for $x = acb$ and $y = ac$ but is not satisfied for $x = acb$ and $y = aac$ etc., because here $l_1 = l(c) = 1$, $l_2 = l(a) = 1$ and $l_3 = l(c) = 1$.

4. Applications in programming languages

In [9] I. M. Glassover, K. V. Hanford and C. B. Jones give the following three types of constraints in generation of phrases in a programming language:

a) The *equality constraints* should give the strings of the form *declare A use A*

where instead of A an arbitrary identifier but the same in both occurrences of A can be written.

If one assumes that all these phrases are obtained from the single phrase *declare* ⟨Ident⟩ *use* ⟨Ident⟩, then it is sufficient to use the scheme of a double rule [⟨Ident⟩ → → A, ⟨Ident⟩ → A] from which all the particular rules are obtained by substituting any string from ⟨Ident⟩ for both occurrences of A.

On the other hand it is possible to use the occurrence conditions too, e.g. it is sufficient to take the following scheme of the rule ⟨Ident⟩ → A ([+ *declare*, : , + *use* A] ∨ [+ *declare* A, *use*, :]).

If one considers a little more realistic situation by accepting the block structure of the programs, the starting phrase may be modified as follows: *begin* X *declare* ⟨Ident⟩ Y *use* ⟨Ident⟩ Z *end* where of course the first *begin* and the last *end* are the corresponding brackets, i.e. between them the same number of occurrences of *begin's* and of *end's* can occur and X, Y and Z are certain strings.

In this case the following condition can help: ⟨Ident⟩ A ([+ *declare*, : , n_1 (+ *begin*) & n_2 (+ *end*); + *use* A] ∨ [+ *declare* A; n_1 (+ *begin*) & n_2 (+ *end*); *use*, :]) where $n_1 = n_2$.

In fact even more realistic situation should be considered when several *use's* are admitted etc.

b) The *inequality constraints* should give the strings of the form *declare* A *declare* B where instead of A and B arbitrary but always mutually different identifiers are written.

If one assumes that all these phrases are obtained from the single phrase *declare* ⟨Ident⟩ *declare* ⟨Ident⟩, then the following scheme of a conditional rule may be used: ⟨Ident⟩ → A ([+ *declare*, : , + *declare*, —A] ∨ [+ *declare*, —A, + *declare*, :]).

Similarly as in the previous case a more realistic situation can be described.

c) The *length constraints* should give the strings of the form *declare* A (x, x, ..., x) *use* A (?, ?, ..., ?) and there is the same number of asterisks as of queries. It is clear that also this case can be described by suitable counting or eventually length conditions.

Finally let us conclude that there is a hope to describe the complete syntax of ALGOL 60 [14] and of ALGOL 68 [16] too using the mentioned new syntactical tools and then to omit all types of "context conditions". In fact in this point the syntax of the new ALGOL 68 is unsufficient in the same measure as it was in ALGOL 60 and even more because a very strong and powerful grammar in two levels is used for ALGOL 68.

REFERENCES

[1] CHOMSKY, N.: Chapters 11—13 in the Handbook of Mathematical Psychology. II. New York, Wiley 1963.

[2] ČULÍK, K.: On some axiomatic systems for formal grammars and languages. Proceedings of IFIP Congress 62, Amsterdam, North Holland 1963, pp. 313—317.

[3] ČULÍK, K.: Formal structure of ALGOL and simplification of its description. Symbolic languages in data processing. New York—London, Gordon-Breach 1962, pp. 75—82.

[4] ČULÍK, K.: On some transformations in context-free grammars and languages. Czech. Math. Journal, *17*, 1967, pp. 278—311.

[5] ČULÍK, K.—HAVEL, I.: On multiple finite automata. 3. Colloquium über Automatentheorie. Basel—Stuttgart, Birkhäuser 1967, pp. 158—169.

[6] ČULÍK, K.: The n-ary grammar and the description of mapping of languages. Kybernetika, *6*, 1970, pp. 99—116.

[7] FLOYD, R. W.: On the nonexistence of a phrase-structure grammar for ALGOL 60. Comm. ACM, 1962, pp. 483—484.

[8] FRIŠ, I.: Grammars with partial ordering of rules. Information and Control, *12*, 1968, pp. 415—425.

[9] GLASSOVER, S. M.—HANFORD, K. V.—JONES, C. B.: The syntax machine. Technical Report 12.077, IBM United Kingdom 1968.

[10] HAVEL, I.: On conditional grammars and regular approximations (in Czech), thesis 1967.

[11] HAVEL, I.: A note on one-sided context-sensitive grammars. Kybernetika, *5*, 1969, pp. 186—189.

[12] KOPŘIVA, J.: Some notes on the formal structure of ALGOL 60. Publ. Fac. Sci. Univ. J. E. Purkyně, Brno, No. 456, 1964, pp. 409—418.

[13] KRÁL, J.: On multiple grammars. Kybernetika, *5*, 1969.

[14] NAUER, P.: Revised report on the algorithmic language ALGOL 60. Comm. ACM, 1963, pp. 1—23.

[15] NAVRÁTIL, E.: Context-free grammars with regular conditions. Kybernetika, *6*, 1970, pp. 118—125.

[16] WIJNGAARDEN, A. van: Report on the algorithmic language ALGOL 68. Mathem. Centrum, Amsterdam, AS MR, 101, 1968.

Karlgren's Decision Grammars

ADRIAN BIRBĂNESCU, BUCUREŞTI

Introduction

The *categorial grammars* defined independently by Y. Bar-Hillel [1, 2] and by J. Lambek [3] provide means of analysing a given string of words and deciding whether it is or not grammatical according to the particular grammar involved. J. Cohen [4] has demonstrated the equivalence between the categorial grammars of Bar-Hillel and those of Lambek. Bar-Hillel, Gaifman and Shamir [5], H. Gaifman [6] and H. Karlgren [7] have shown that the above-mentioned categorial grammars are equivalent with phrase-structure systems. H. Karlgren has indicated in [7] a way of extending the categorial grammar method to the analysis of strings belonging to context-sensitive languages. Later on, in a paper presented at the 1969 International Conference on Computational Linguistics [8], H. Karlgren has given a general definition for grammars which allow to decide whether or not a given string is grammatical. He has exemplified also ways of applying such a *decision grammar*, with rules called by him *context-free*, to the analysis of non context-free languages. The "computing" procedure of a decision grammar is the converse of that of a decision generative. In this paper, at a suggestion made by Solomon Marcus, it is shown that the context-free decision grammars of Karlgren are equivalent to the *context-free generative grammars* as defined by N. Chomsky and G. Miller [9]. From [5], [6] and [7] it follows then immediately that Karlgren's context-free decision grammars are also equivalent to the categorial grammars defined in [1—4].

Although a new proof would be thus unnecessary, it is shown here directly that the traditional categorial grammars are included in Karlgren's definition of context-free decision grammars. This independent proof was thought useful for showing that Karlgren's definitions are more general than previous ones.

Definitions

Karlgren defines a *reduction grammar* by $G = \langle S, R, I, T\rangle$ where S is an *alphabet*, $I \subset S$ the *input alphabet*, $T \subset S^*$ a set of *target symbols* and R a set of *reduction rules* such that

$$R = \{\alpha \dot{\to} \beta \mid \alpha = a_1 a_2 \dots a_m, \qquad \beta = b_1 b_2 \dots b_n, a_i \in S, b_j \in S, m \geqq n\}.$$

In the definition of R, α and β are thus *strings* over the alphabet S. As Karlgren admits also rules of the form $\alpha \dot{\to} \alpha$, $\dot{\to}$ is a reflexive, asymmetric and transitive relation. We have deliberately denoted this relation by $\dot{\to}$ with the intention of making a difference between it and the analogue relation $\to$ usually met in generative grammars.

In traditional notations S—$I = V_N$ is called a *non-terminal* alphabet.

If T contains only one element, let it be T, Karlgren says that the reduction grammar is a *decision grammar* "which specifies for each input string whether or not it is grammatical".

Further on, Karlgren defines as a *context-free rule* a reduction rule with the right hand side containing exactly one symbol, and as a *context-free grammar* a grammar the rules of which are all context-free.

Evidently the language $L(G)$ is the set of strings over the alphabet I such that they can be reduced by R to T, i.e.

$$L = L(\langle S, R, I, T\rangle) = \{\sigma \mid \sigma \in I^*, \sigma \dot{\Rightarrow} T\},$$

where $\sigma \dot{\Rightarrow} T$ means that there is a sequence of strings $\sigma_1, \sigma_2, \dots, \sigma_n$, such that

$$\sigma \dot{\to} \sigma_1,$$
$$\sigma_1 \dot{\to} \sigma_2,$$
$$\sigma_{n-1} \dot{\to} \sigma_n,$$
$$\sigma_n \dot{\to} T.$$

Lemma 1. *For every context-free generative grammar G_C there is an equivalent context-free decision grammar G_D.*

Proof. To make this proof more pregnant we shall start from the definition of a context-free generative grammar as given by Chomsky and Miller in [2]. According to this definition

$$G_C = \langle V, \to, V_T, S\rangle,$$

where V is a finite set of symbols called *vocabulary*, $V_T \subset V$ is the *terminal vocabulary*, $\to$ a finite, reflexive and asymmetric relation and S the *initial symbol* which can be

read *sentence*. (In their original definition, Chomsky and Miller consider $\to$ an irreflexive relation. The assumption of reflexivity does not reduce the generality of the proof.) The rules of G_C are of the form

$$A \to \omega,$$

where $A \in V_N = V - V_T$ and ω is a string of elements of V.

We say that a sentence σ is generated by G_C, and we write $s \Rightarrow \sigma$, if, by applying the rules of G_C, there is a sequence

$$s \to \sigma_n,$$
$$\sigma_n \to \sigma_{n-1},$$
$$\sigma_1 \to \sigma$$

and $\sigma \in V_T^*$.

Let us construct now a context-free decision grammar $G_D = \langle S, R, I, T\rangle$ by putting

$$S = V,$$
$$R = \{\omega \dot{\to} A \mid A \to \omega, \qquad A \in V_N = S - I, \omega \subset S^*\},$$
$$I = V_T,$$
$$T = s.$$

a) Every sentence generated by G_C is specified by G_D as grammatical.

Indeed, let σ be the sentence generated by G_C. By its very definition $\to$ is the inverse of $\dot{\to}$. This means that from $\sigma_1 \to \sigma$ it follows, by applying a certain rule of R, that $\sigma \dot{\to} \sigma_1$, and so on, until we find $\sigma_n \dot{\to} s$. But $s = T$, and so we can write

$$\sigma \dot{\Rightarrow} T.$$

This proves our statement which implies

$$L(G_C) \subset L(G_D).$$

b) Every string specified by G_D as grammatical is a sentence generated by G_C.

The proof is again based on the fact that $\to$ is the inverse relation of $\dot{\to}$, which permits us to reverse the order of the sequence implied by $\sigma \dot{\Rightarrow} T$ and find the sequence which implies $s \Rightarrow \sigma$. From this it follows that

$$L(G_D) \subset L(G_C).$$

Evidently, a) and b) imply

$$L(G_D) = L(G_C)$$

which proves our lemma.

Lemma 2. *If a context-free decision grammar* G_D *contains rules whose right hand side is an input symbol, i.e.* $a_1a_2 \dots a_n \to b$ *and* $b \in I$, *we can construct an equivalent grammar* G'_D *with all rules containing in the right hand side only non-terminal symbols.*

Proof. Let us examine a rule of the form

$$a_1a_2 \dots a_n \to b, b \in I. \tag{1}$$

Since b is an input symbol, no application of this rule alone can lead to a decision, i.e. to the symbol T. Then, the analysis of the string under consideration is determined by further rules containing b, if such rules exist. Since the rules are in fact couples of a binary relation, b can occur or not at two places, which means $2^2 = 4$ possibilities. Since the rules are context-free, b can occur on the right-hand alone only, and on the left hand either alone or together with other symbols. This adds two further possibilities, rising the total to the six contained in the following table.

right hand		left hand	
	b absent	b occurs alone	b occurs accompanied
b absent	a	b	c
b occurs	d	e	f

a) This means that b occurs in no other rule than (1). Then we can write directly

$$a_1a_2 \dots a_n \to B$$

with $B \in V_N$ and B not appearing in any other rule. Incidentally we may remark that in this case no string containing b is grammatical, so that b can be omitted from I.

b) This means that b occurs in a rule of one of the forms

$$b \to B, B \in V_N,$$

$$b \to c, \quad c \in I.$$

In the first case we may replace (1) by

$$a_1a_2 \dots a_n \to B.$$

The second case may be examined according to the kind of other rules containing c, these being evidently of one of the types listed under a) to f).

c) The symbol b appears in a rule of one of the forms

$$b_1b_2 \dots b \dots b_n \to B,$$

$$b_1b_2 \dots b \dots b_n \to c.$$

In the first case we can replace (1) by the rules

$$b \to C,$$

$$a_1a_2 \dots a_n \to C,$$

$$b_1b_2 \dots C \dots b_n \to B.$$

In the second case the examination must proceed on until it is possible to apply one or more of the solutions listed under a) to f).

d) The symbol b occurs also in a rule of the form

$$b_1b_2 \dots b_n \to b. \tag{2}$$

We shall now replace (1) and (2) by

$$b \to B,$$

$$a_1a_2 \dots a_n \to B,$$

$$b_1b_2 \dots b_n \to B.$$

e) The symbol b occurs also in a rule of the form

$$b \to b.$$

This rule is trivial and can be erased from the set R as giving birth to no interesting derivation. Then case e) reduces to case a).

f) The symbol b occurs also in a rule of the form

$$b_1b_2 \dots b \dots b_n \to b.$$

This is the second case of c) and it should be treated according to whether or not b occurs also in other rules of R. If not, for example, we should replace the rules

$$a_1a_2 \dots\dots\dots a_n \to b,$$

$$b_1b_2 \dots b \dots b_n \to b,$$

by

$$b \to B,$$

$$a_1a_2 \dots\dots\dots a_n \to B,$$

$$b_1b_2 \dots B \dots b_n \to B.$$

All other cases can be solved by recursively combining the solutions listed under a) to f).

Lemma 3. *For every context-free decision grammar G_D there is an equivalent context-free generative grammar G_C.*

Let us consider the general case of a context-free decision grammar $G_D = \langle S, R, I, T\rangle$ containing also rules with input symbols on the right-hand side. By Lemma 2 we shall construct an equivalent grammar $G'_D = \langle S, R', I, T\rangle$ with rules containing only non-terminal symbols on the right-hand side. Further on we shall construct a generative grammar $G_C = \langle V, \rightarrow, V_T, s\rangle$, where

$$V = S,$$

$$\rightarrow = (\dot{\rightarrow})^{-1}, \qquad \text{for all } \dot{\rightarrow} \subset R',$$

$$V_T = I,$$

$$s = T.$$

By virtue of $\rightarrow = (\dot{\rightarrow})^{-1}$ the set of rules of G_C will contain only rules with one non-terminal symbol on the left-hand, that is G_C will be a context-free grammar. The proof is to be continued on the same lines as that of Lemma 1.

Theorem 1. *The class $T(G_D)$ of languages specified by context-free decision grammars is equal to the class $T(G_C)$ of languages generated by context-free generative grammars.*

Proof. From Lemma 1 it follows $T(G_C) \subset T(G_D)$, while from Lemma 3 results $T(G_D) \subset T(G_C)$. This means $T(G_D) = T(G_C)$.

Theorem 2. *Every categorial grammar G_K is a context-free decision grammar.*

Proof. Let us consider a categorial grammar G_K defined by:

— a vocabulary V grouped into categories C_i;

— a set of symbols Z_i, one for each category, Z_i being of one of the forms

$$X, Y \ldots$$

$$X/Y,\ Y\backslash X, \ldots$$

or more complex forms derived recursively from the above forms:

— a set of rules of one of the forms:

$$(X/Y)\ Y = X, \tag{3}$$

$$Y(Y|X) = X,$$

— a category s meaning sentence.

It can be seen immediately that this grammar falls directly under the definition of a context-free decision grammar when putting

$$V = I,$$

$$\{Z_i, i = 1, 2, \ldots, n\} = S - I,$$

$$s = T,$$

while in the rules (3) there are two symbols on the left-hand side and only one on the right-hand side. It must be stressed here that X/Y, for instance, is only one symbol.

Acknowledgements

The author thanks Prof. Solomon Marcus of the Bucharest University and Alexandru Cărăușu of the Jassy University for their comments on the first draft of this paper.

REFERENCES

[1] BAR-HILLEL, Y.: A quasi-arithmetical notation for syntactic description. Language, *29*, 1953, pp. 47—53.

[2] BAR-HILLEL, Y.: Language and Information. Addison-Wesley and the Jerusalem Academic Press, 1964.

[3] LAMBEK, J.: The mathematics of sentence structure. American Mathematical Monthly, *65*, 1958, pp. 154—170.

[4] COHEN, J. M.: The equivalence of two concepts of categorial grammars. Information and Control, *11*, No. 5, May 1967, pp. 475—484.

[5] BAR-HILLEL, Y.—GAIFMAN, H.—SHAMIR, E.: On categorial and phrase structure grammars. In: Bulletin of the Research Council of Israel, 9F, 1960, pp. 1—16; republished in [2].

[6] GAIFMAN, H.: Dependency systems and phrase-structure systems. Information and Control, *8*, No. 3, June 1965, pp. 304—337.

[7] KARLGREN, H.: Categorial grammar analysis of context-sensitive languages. KVAL Rep. No. 441, 1968.

[8] KARLGREN, H.: Multi-index syntactical calculus. International conference on computational linguistics, Stockholm 1969. Reprint No. 68.

[9] CHOMSKY, N.—MILLER, A.: Introduction to the formal analysis of natural languages. Handbook of Mathematical Psychology, Vol. II. Ed. R. D. Luce et al., New York, Wiley 1963.

Meaning of Tense and Its Recursive Properties (Summary)*

EVA HAJIČOVÁ, JARMILA PANEVOVÁ, AND PETR SGALL, PRAHA

In the present paper, the linguistic meaning of Czech verbal tenses is examined, in comparison with the English ones; the study of the nature of semantic units the presence of which in the system of the given languages is granted by linguistic (grammatical) criteria rather than a systematic investigation of the cognitive content of tenses and temporal relations is the objective here.

Reichenbach's notions of the point of speech (S), the point of event (E) and the point of reference (R) served as a point of departure for our discussion. The placement of a predication (clause, embedded sentence) in time is its relation to some point of reference, determined in accordance with the characterization of a predication as "content" clauses (reported speech in a broader sense) and "adjunct" clauses and in accordance with its syntactic position (its position in the dependency tree). Every sentence has its S (a point of speech which is fixed for the whole complex sentence), every predication has its E (the point of event) and its R (point of reference), which coincides either with S (with all main predications and with some dependent ones) or with E of some other predication. According to statements common in Czech grammars, R of a content clause would coincide with E of its governing clause, while R of adjunct clause coincides with S. However, our examination of complex sentences with deeper embeddings has shown that only the first half of this statement holds; examples as

(i) Slíbil (E_1), že nám celý postup objasní (E_2), protože se na přednášce seznámí (E_3) s jeho podstatou.
He promised (E_1) he would explain (E_2) the whole procedure to us because he would get acquainted (E_3) with its principles,

when compared with

(ii) Slíbil (E_1), že nám celý postup objasní (E_2), protože se na přednášce seznámil (E_1) s jeho podstatou.

---

* The full text of the paper was published in Philologica Pragensia, *14*, 1971, 1—15, 57—64. for a preliminary version, see The Prague Bulletin of Mathematical Linguistics, *13*, 1970, 9—42.

He promised (E_1) he would explain (E_2) to us the whole procedure because he got acquainted (E_3) with its principles show that it is the relation of E_3 to E_1 (in (i) E_3 is after E_1, in (ii) E_3 is before E_1) which is expressed by the form of the verb in the deepest embedded clause (adjunct clause in both cases) and not that of E_3 to S. That means that R of the adjunct clause coincides with R of the content clause to which it is subordinated; and it does not equal S. To determine the point of reference of every predication it is necessary to apply a recursive principle (the content clause in question can be itself subordinated to another content clause, etc.), which can be in a shortened form written as a sequence of three rules:*

(x) $R_{\text{main}} = S$,
(xx) $R_{\text{content}} = E_{\text{govern}}$,
(xxx) $R_{\text{depend}} = R_{\text{govern}}$,

where the subscripts **main**, **content**, **depend**, and **govern** stand for the main clause, the content clause the dependent clause of another type, and governing clause, respectively.

The discussion of some apparent counterexamples and of the properties of rules that could describe these phenomena in Czech has shown that the above recursive principle stated by Panevová is adequate for the description of meaning of tenses in Czech; we may assume that it has a more general validity and that it accounts in the main also for the English verb system. However, as our tentative investigation has shown, for a description of the rich repertoir of English verbal forms some additional rules are needed; some of them are sketched in our paper.

* See J. Panevová, Některé otázky závislé predikace v generativním popisu češtiny. Dissertation, Charles University, Prague 1969.

Un modèle mathématique intégral de l'oeuvre dramatique

SOLOMON MARCUS, BUCUREȘTI

Construction du modèle

Une oeuvre dramatique est une collection $\Omega = \langle P, L, R, \mathscr{I}, f, g, h, \varphi, \psi, T, \Sigma, \mu \rangle$ de 12 objets. Par P, L, R et $\mathscr{I}$ on désigne des ensembles finis non vides, disjoints deux à deux. Les éléments de P sont des *personnages*. Les éléments de L sont des *lieux*. Les éléments de R sont des *répliques*. Les éléments de $\mathscr{I}$ sont des *indications de régie*. Les fonctions f, g et h ont les valeurs dans $\mathscr{I}$ et sont définies respectivement dans P, L et R, de sorte que $f(P) \cup g(L) \cup h(R) = \mathscr{I}$. Les fonctions φ et ψ sont définies sur R et on a $\varphi(R) = P$, $\psi(R) \subseteq L$. On désigne par T une certaine phrase $x_1 x_2 \dots x_n$ sur le vocabulaire $R \cup \mathscr{I}$, telle que pour chaque $x \in R \cup \mathscr{I}$ il existe un entier positif i ($1 \leqq i \leqq n$) avec la propriété $x_i = x$.

Une *situation* est une sous-phrase de T de la forme $x_i x_{i+1} \dots x_{j-1} x_j$ avec $1 \leqq i < j \leqq n$. Deux situations sont dites *consécutives* si leur concaténation est aussi une situation, c'est-à-dire si elles sont de la forme $x_i x_{i+1} \dots x_j$ et $x_{j+1} x_{j+2} \dots x_k$. Désignons par Σ une certaine décomposition de T en situations consécutives: $T = S_1 S_2 \dots S_m$. Ici donc les situations S_i et S_{i+1} sont consécutives pour chaque i tel que $1 \leqq i \leqq m - 1$. Les situations $S_i (1 \leqq i \leqq m)$ sont des *situations marquées*. Posons $\mathscr{C} = \{S_1, S_2, \dots, S_m\}$. L'application μ sera définie sur l'ensemble $\mathscr{C}$ et aura les valeurs dans le produit cartésien $2^P \times L$, où par 2^P on a noté l'ensemble des parties de P. Toute collection de personnages est une *configuration de personnages*. Une configuration γ pour laquelle il existe une situation marquée S_i et un lieu $d \in L$ tels que $\mu(S_i) = \langle \psi, d \rangle$ est une *configuration marquée*. Désignons par $\mathscr{C}$ l'ensemble des configurations marquées (parmi ces configurations il peut se trouver aussi la configuration vide). L'ensemble $\mathscr{C}$ (donc, en dernière instance, l'application μ) sera soumis aux deux conditions suivantes: 1° Pour chaque personnage p il existe une configuration marquée γ telle que $p \in \gamma$; 2° pour chaque lieu d il existe une situation marquée S_i et une configuration marquée γ, telles que $\mu(S_i) = \langle \gamma, d \rangle$.

Commentaire du modèle

Les ensembles P et L constituent l'axe paradigmatique, tandis que les ensembles R et $\mathscr{I}$ constituent l'axe syntagmatique. Un lieu représente la totalité des éléments de décore existant à un certain moment dans la scène. Les applications f, g et h ayant les valeurs dans $\mathscr{I}$, expriment la primordialité des indications de régie dans toute oeuvre dramatique, le fait que tout élément de l'oeuvre dramatique exige une indication de régie. La surjection φ est suggérée par l'idée de St. Jansen; selon cet auteur, la réplique est la condition nécessaire et suffisante de l'existence d'un personnage. Une personne présente dans la scène, mais qui ne prononce jamais une réplique, n'est pas un personnage, mais un élément du décore. Un objet qui parle au moins une fois, par exemple un appareil de radio, est un personnage, donc un élément de P. La surjection φ exprime le fait qu'une réplique ne peut appartenir qu'à un seul personnage. Cela signifie que, dans l'analyse d'une oeuvre déterminée, un groupe de personnes qui prononcent, à un certain moment, la même réplique sera envisagé comme un seul personnage. Cette convention ajoute, à la liste des personnages donnée par l'auteur, quelques personnages supplémentaires.

Une situation marquée est une séquence de répliques et d'indications de régie, dans l'ordre où elles sont données par l'auteur (ici intervient d'une manière essentielle le fait que le modèle ci-dessus concerne des oeuvres dramatiques écrites et non pas des représentations des oeuvres; il concerne des pièces et non pas des spectacles); une situation marquée correspond à un intervalle maximum de temps où aucune modification ne se produit en ce qui concerne la configuration des personnages et le décore. Ce fait est visible sur la définition de l'application μ, qui associe à une situation marquée un ensemble déterminé de personnages (l'ensemble des personnages présentes dans la scène pendant la situation envisagée) et un décore bien déterminé (c'est-à-dire, le décore où se déroule la situation marquée envisagée). On s'apperçoit donc que la segmentation Σ du texte T d'une oeuvre dramatique en situations marquées correspond approximativement (mais pas exactement!) à la division en scènes. Évidemment, toute réplique et toute indication de régie doivent figurer dans T; autrement, elles seraient parasites. L'implication $\ll x \in R \cup \mathscr{I}$ — il existe un i $(1 \leqq i \leqq n)$ tel que $x = x_i \gg$ exprime donc le fait que les ensembles R et $\mathscr{I}$ ne contiennent pas des éléments parasites. La condition 1° imposée à μ exprime le fait qu'il n'y a dans P aucun personnage parasite, tandis que la condition 2° imposée à μ exprime le fait que L ne contient aucun élément parasite.

Lieux et personnages en corrélation

Un ensemble L_1 de lieux et un ensemble P_1 de personnages sont en corrélation si pour chaque lieu $d \in L_1$ il existe une situation marquée S telle que $\mu(S) = \langle P_1, d \rangle$, mais pour aucun lieu $d \notin L_1$ il n'existe aucune situation marquée S telle que $\mu(S) = \langle P_1, d \rangle$. Les ensembles P_2 et L_2 sont en semicorrélation s'il existe deux ensembles P_1 et L_1 en corrélation, telles que les inclusions $P_2 \subset P_1$ et $L_2 \subset L_1$ soient strictes.

Nous dirons que le personnage β est présent dans la situation S, si $p \in P'$, où $\mu(S) = \langle P', d \rangle$. Les personnages p_1 et p_2 sont *équivalents* s'ils sont présents exactement dans les mêmes situations. Le personnage p_1 *domine* le personnage p_2 si, dans toute situation où p_2 est présent, le personnage p_1 est aussi présent. Les personnages p_1 et p_2 sont *interférents* s'il existe au moins une situation où tous les deux sont présents. Les personnages interférants p_1 et p_2 sont *indépendants* si aucun d'entre eux ne domine pas l'autre.

Proposition 1. *Si le personnage p_1 domine le personnage p_2, alors il existe un ensemble L_2 de lieux qui est en semi-corrélation avec l'ensemble $\{p_1, p_2\}$.*

Proposition 2. *Si γ est une configuration marquée de personnages, il existe un ensemble λ non-vide de lieux, en corrélation avec γ.*

Proposition 3. *S'il existe un ensemble λ non-vide de lieux qui est en corrélation avec la configuration γ, alors γ est marquée.*

Théorème 1. *Afin qu'il existe un ensemble λ de lieux en corrélation avec la configuration de personnages γ il faut et il suffit que γ soit une configuration marquée.*

Domination entre configurations

On dira que la configuration γ_1 *domine* la configuration γ_2 (et on va écrire $\gamma_1 \to \gamma_2$) si pour toute configuration marquée γ telle que $\gamma_2 \subseteq \gamma$ on a $\gamma_1 \subseteq \gamma$.

Proposition 4. *Si les configurations marquées forment une suite monotone ascendante $\gamma_1 \subset \gamma_2 \subset \ldots \subset \gamma_n$, alors $\gamma_1 - \gamma_2 - \ldots - \gamma_{n-1} - \gamma_n$.*

On dira que la configuration marquée γ de personnages est *irréductible* s'il n'existe aucune configuration marquée γ_1 strictement contenue dans γ. On dira que γ' est une *sous-configuration marquée* de γ si $\gamma' \subset \gamma$ et si γ et γ' sont des configurations marquées. Si γ n'est une sous-configuration marquée d'aucune autre configuration marquée, alors on dira que γ est une *configuration marquée maximale*.

Proposition 6. *Si chaque configuration marquée est irréductible, alors chaque configuration marquée est maximale et il n'y a pas de configurations marquées distinctes en relation de domination.*

Une configuration sera dite *semimarquée* si elle est contenue dans une configuration marquée. Autrement, elle est *parasite*. On dira que l'ensemble $\mathscr{C}$ des configurations marquées est *stable par rapport à l'opération d'intersection* si pour $\gamma_1, \gamma_2 \in \mathscr{C}$ on a $\gamma_1 \cap \gamma_2 \in \mathscr{C}$.

Théorème 2. *Soit $\mathscr{C}$ stable par rapport à l'intersection et soit B une configuration non-marquée. Alors il existe une configuration A disjointe de B, qui domine B.*

Remarque. Dans le cas particulier où $A = \{p_1\}$, $B = \{p_2\}$, $p_1 p_2 \in P$, $p_1 \neq p_2$ et $\mathscr{C}$ est une topologie dans l'ensemble des configurations, on retrouve un résultat de Mihai Dinu (théorème 1 de [6]).

Theorème 3. *Soit $\mathscr{C}$ stable par rapport à l'intersection. Soit n un entier positif. Si aucune configuration formée de n personnages au plus n'est marquée, alors il existe au moins deux configurations distinctes, formées chacune de n personnages au plus, qui se dominent réciproquement.*

Remarque. Dans le cas particulier $n = 1$ et $\mathscr{C}$ formant une topologie, on retrouve un résultat de Mihai Dinu (théorème 2 de [6]).

La signification du théorème 3 est éclaircie par la

Proposition 6. *Si deux configurations A et B se dominent réciproquement, alors toute configuration marquée contenant $A(B)$ contient aussi $B(A)$.*

Noyaux

Nous rappelons que le personnage p est présent dans la situation S si $p \in P'$, où $\mu(S) = \langle P', d \rangle$. Nous associons à chaque oeuvre dramatique Ω une matrice booléenne $M(\Omega)$ ayant le nombre de lignes égal au nombre des personnages et le nombre des colonnes égal au nombre des situations marquées. A l'intersection de la ligne de rang i avec la colonne de rang j on a le chiffre 1 si le personnage de rang i est présent dans la situation S_j et on a le chiffre O s'il en est absent. Envisageons le graphe G dont les sommets sont les personnages et où deux sommets p_1 et p_2 sont adjacents s'il existe une situation où p_1 et p_2 sont tous les deux présents. Un noyau de ce graphe définit une conception régisorale de la pièce. On a, en général, plusieurs noyaux, donc plusieurs conceptions régisorales. (Rappelons qu'un noyau est un ensemble de sommets qui ne sont pas adjacents deux à deux, mais tel que tout autre sommet est adjacent avec un sommet du noyau.) Un noyau peut contenir des personnages envisagés comme secondaires dans la pièce. La recherche des noyaux est facilitée par les propositions suivantes.

Proposition 7. *Si A et B sont des noyaux tels que $A \subseteq B$, alors $A = B$.*

Corollaire. *S'il existe un noyau formé d'un seul personnage, ce personnage n'appartient à aucun autre noyau.*

Proposition 8. *Deux personnages interférants n'appartiennent jamais à un même noyau.*

Proposition 9. *Soit p_1 et p_2 deux personnages équivalents. Dans ces conditions: 1° il n'existe aucun noyau contenant p_1 et p_2 à la fois; 2° Si N est un noyau contenant p_1, alors l'ensemble $(N \cup \{p_2\}) - \{p_1\}$ est aussi un noyau.*

Théorème 4. *Si A est un noyau entre les situations de rangs i et j $(i < j)$ et si B est un noyau entre les situations de rang $j + 1$ et k $(j < k)$ tels que les personnages de $B - A$ ne sont pas présents dans les situations dont les rangs sont compris entre i et j, tandis que les personnages de $A - B$ ne sont pas présents dans les situations dont les rangs sont compris entre $j + 1$ et k, alors l'ensemble $A \cup B$ est un noyau entre la situation de rang i et la situation de rang k* (c'est à dire, lorsque on fait abstraction des situations dont le rang est inférieur à i ou supérieur à k).

Langages associés à une oeuvre dramatique

Nous dirons qu'une suite de personnages $p_1 p_2 \dots p_s$ est *marquée* s'il existe une suite de situations consécutives $S_{i_1} S_{i_2} \dots S_{i_s}$ telle que p_{i_j} est présent dans la situation S_{i_j} $(j = 1, 2, \dots, s)$. L'ensemble des suites marquées est un langage λ sur le vocabulaire P. Ce langage est toujours fini; la longueur de ses phrases ne peut pas dépasser le nombre total des situations marquées. Mais nous pouvons „prolonger" ce langage, de sorte qu'il devient éventuellement infini et reflète certaines tendances récursives de l'oeuvre dramatique. Nous allons donc définir un nouveau langage sur le vocabulaire P. On dira que la phrase $p_1 p_2 \dots p_s$ est *quasimarquée* si pour chaque entier positif i $(1 \leqq i \leqq s - 1)$ la phrase $p_i p_{i+1}$ est marquée, c'est-à-dire elle est dans λ. Désignons par λ^* l'ensemble des phrases quasi-marquées sur P. Le langage λ^* peut être infini; c'est justement la situation habituelle, car on a le

Théorème 5. *Afin que le langage λ^* associé à une oeuvre dramatique Ω soit infini il suffit qu'il existe dans Ω au moins deux situations consécutives qui correspondent à des configurations non-disjointes de personnages.*

Proposition 10. *Les langages λ et λ^* sont héréditaires* (c'est-à-dire toute sous-phrase d'une phrase marquée est marquée et toute sous-phrase d'une phrase quasi-marquée est quasi-marquée) *et on a toujours $\lambda \subseteq \lambda^*$.*

Les langages λ et λ^* nous permettent d'étudier la syntagmatique des personnages, leurs relations de dépendance et de subordination, leurs configurations syntaxiques (au sens de Kulagina, Gladkiĭ ou Novotný; voir [16], chapitre V).

La segmentation morphémique de Harris (voir [10] ou [17], ch. III) pourrait être appliquée à la phrase T, afin d'obtenir des unités dramatiques plus fines que les actes. Mihai Dinu a utilisé une méthode de Ferdinand de Saussure pour obtenir la segmentation en syllabes dramatiques [5], [6].

En ce qui concerne le langage λ^*, il pourrait montrer son efficacité dans l'étude fonctionnelle du mécanisme de génération des phrases marquées de personnages. Une telle étude ne pourrait être accomplie à l'aide du langage λ, car celui-ci est toujours fini.

Indications bibliographiques

Certains aspects combinatoirs de l'oeuvre dramatique ont été investigué par G. Politi [23] et E. Souriau [30]. La géometrie dramatique a été étudiée par P. Ginestier [8], qui utilise d'une manière implicite les graphes, mais pas leur théorie mathématique. De ce point de vue, le travail de Ginestier rappelle le livre de L. Tesnière [31]. Nous avons donné dans [15], [18] et [19] quelques aspects des application dramatiques de la matrice booléenne $M(\Omega)$. Un résumé de ces applications a été donné par L. Kalmár [14]. L'étude informationnelle du théâtre est dûe à F. von Cube

Theorème 3. *Soit $\mathscr{C}$ stable par rapport à l'intersection. Soit n un entier positif. Si aucune configuration formée de n personnages au plus n'est marquée, alors il existe au moins deux configurations distinctes, formées chacune de n personnages au plus, qui se dominent réciproquement.*

Remarque. Dans le cas particulier $n = 1$ et $\mathscr{C}$ formant une topologie, on retrouve un résultat de Mihai Dinu (théorème 2 de [6]).

La signification du théorème 3 est éclaircie par la

Proposition 6. *Si deux configurations A et B se dominent réciproquement, alors toute configuration marquée contenant $A(B)$ contient aussi $B(A)$.*

Noyaux

Nous rappelons que le personnage p est présent dans la situation S si $p \in P'$, où $\mu(S) = \langle P', d \rangle$. Nous associons à chaque oeuvre dramatique Ω une matrice booléenne $M(\Omega)$ ayant le nombre de lignes égal au nombre des personnages et le nombre des colonnes égal au nombre des situations marquées. A l'intersection de la ligne de rang i avec la colonne de rang j on a le chiffre 1 si le personnage de rang i est présent dans la situation S_j et on a le chiffre O s'il en est absent. Envisageons le graphe G dont les sommets sont les personnages et où deux sommets p_1 et p_2 sont adjacents s'il existe une situation où p_1 et p_2 sont tous les deux présents. Un noyau de ce graphe définit une conception régisorale de la pièce. On a, en général, plusieurs noyaux, donc plusieurs conceptions régisorales. (Rappelons qu'un noyau est un ensemble de sommets qui ne sont pas adjacents deux à deux, mais tel que tout autre sommet est adjacent avec un sommet du noyau.) Un noyau peut contenir des personnages envisagés comme secondaires dans la pièce. La recherche des noyaux est facilitée par les propositions suivantes.

Proposition 7. *Si A et B sont des noyaux tels que $A \subseteq B$, alors $A = B$.*

Corollaire. *S'il existe un noyau formé d'un seul personnage, ce personnage n'appartient à aucun autre noyau.*

Proposition 8. *Deux personnages interférants n'appartiennent jamais à un même noyau.*

Proposition 9. *Soit p_1 et p_2 deux personnages équivalents. Dans ces conditions: 1° il n'existe aucun noyau contenant p_1 et p_2 à la fois; 2° Si N est un noyau contenant p_1, alors l'ensemble $(N \cup \{p_2\}) - \{p_1\}$ est aussi un noyau.*

Théorème 4. *Si A est un noyau entre les situations de rangs i et j $(i < j)$ et si B est un noyau entre les situations de rang $j + 1$ et k $(j < k)$ tels que les personnages de $B - A$ ne sont pas présents dans les situations dont les rangs sont compris entre i et j, tandis que les personnages de $A - B$ ne sont pas présents dans les situations dont les rangs sont compris entre $j + 1$ et k, alors l'ensemble $A \cup B$ est un noyau entre la situation de rang i et la situation de rang k* (c'est à dire, lorsque on fait abstraction des situations dont le rang est inférieur à i ou supérieur à k).

Langages associés à une oeuvre dramatique

Nous dirons qu'une suite de personnages $p_1 p_2 \dots p_s$ est *marquée* s'il existe une suite de situations consécutives $S_{i_1} S_{i_2} \dots S_{i_s}$ telle que p_{i_j} est présent dans la situation S_{i_j} $(j = 1, 2, \dots, s)$. L'ensemble des suites marquées est un langage λ sur le vocabulaire P. Ce langage est toujours fini; la longueur de ses phrases ne peut pas dépasser le nombre total des situations marquées. Mais nous pouvons „prolonger" ce langage, de sorte qu'il devient éventuellement infini et reflète certaines tendances récursives de l'oeuvre dramatique. Nous allons donc définir un nouveau langage sur le vocabulaire P. On dira que la phrase $p_1 p_2 \dots p_s$ est *quasimarquée* si pour chaque entier positif i $(1 \leqq i \leqq s - 1)$ la phrase $p_i p_{i+1}$ est marquée, c'est-à-dire elle est dans λ. Désignons par λ^* l'ensemble des phrases quasi-marquées sur P. Le langage λ^* peut être infini; c'est justement la situation habituelle, car on a le

Théorème 5. *Afin que le langage λ^* associé à une oeuvre dramatique Ω soit infini il suffit qu'il existe dans Ω au moins deux situations consécutives qui correspondent à des configurations non-disjointes de personnages.*

Proposition 10. *Les langages λ et λ^* sont héréditaires* (c'est-à-dire toute sous-phrase d'une phrase marquée est marquée et toute sous-phrase d'une phrase quasi-marquée est quasi-marquée) *et on a toujours $\lambda \subseteq \lambda^*$.*

Les langages λ et λ^* nous permettent d'étudier la syntagmatique des personnages, leurs relations de dépendance et de subordination, leurs configurations syntaxiques (au sens de Kulagina, Gladkiĭ ou Novotný; voir [16], chapitre V).

La segmentation morphémique de Harris (voir [10] ou [17], ch. III) pourrait être appliquée à la phrase T, afin d'obtenir des unités dramatiques plus fines que les actes. Mihai Dinu a utilisé une méthode de Ferdinand de Saussure pour obtenir la segmentation en syllabes dramatiques [5], [6].

En ce qui concerne le langage λ^*, il pourrait montrer son efficacité dans l'étude fonctionnelle du mécanisme de génération des phrases marquées de personnages. Une telle étude ne pourrait être accomplie à l'aide du langage λ, car celui-ci est toujours fini.

Indications bibliographiques

Certains aspects combinatoirs de l'oeuvre dramatique ont été investigué par G. Politi [23] et E. Souriau [30]. La géometrie dramatique a été étudiée par P. Ginestier [8], qui utilise d'une manière implicite les graphes, mais pas leur théorie mathématique. De ce point de vue, le travail de Ginestier rappelle le livre de L. Tesnière [31]. Nous avons donné dans [15], [18] et [19] quelques aspects des application dramatiques de la matrice booléenne $M(\Omega)$. Un résumé de ces applications a été donné par L. Kalmár [14]. L'étude informationnelle du théâtre est dûe à F. von Cube

[4], qui a utilisé la notion d'entropie elective dûe à J. C. Moreno [22]. Un exposé plus détaillé de notre modèle, avec toutes les démonstrations des propositions et des théorèmes, peut être trouvé dans le chapitre VIII de notre livre [20]. Certains objets de notre modèle remontent à St. Jansen [11], [12], [13]. L'utilisation de la matrice $M(\Omega)$ exige l'utilisation de la théorie des graphes, pour laquelle nous renvoyons au livre de Cl. Berge [1]. Pour l'étude des noyaux on peut utiliser avec profit l'article [26] de S. Rudeanu. D'autres idées de notre modèle remontent à Th. Brijitte [2], L. Egri [7], A. J. Greimas [9], A. G. Matache [21], L. J. Prieto [24], A. I. I. Radcliffe [25], W. Sacksteder [27], J. Schérer [28], E. Souriau [29] et P. von Tieghem [32].

BIBLIOGRAPHIE

[1] BERGE, C.: Théorie des graphes et ses applications. Paris, Dunod 1958.

[2] BRIJITTE, Th.: Szenenschluss, Szenenanfang und Szenennaht in Shakespeare Historien und Tragödien. Dissertation, München 1965.

[3] CHOMSKY, N.: Formal properties of grammars. Handbook of Mathematical Psychology, vol. 2 (Editors R. D. Luce, R. R. Bush, E. Galanter). New York, John Wiley 1963.

[4] CUBE, F. van: Das Drama als Forschungsobjekt der Kybernetik. Mathematik und Dichtung, 1965.

[5] DINU, M.: Structures linguistiques probabilistes dans l'étude du théâtre. Cahiers de linguistique théorique et appliquée, *5*, 1968, p. 29—46.

[6] DINU, M.: Contributions à l'étude mathématique du théâtre. Revue roumaine de mathématiques théorique pures et appliquées, *15*, 1970, No. 4.

[7] EGRI, L.: The art of dramatic writing. Its basis in the creative interpretation of human motives. Boston, The Writer, Inc. Publishers 1960.

[8] GINESTIER, P.: Le théâtre contemporain dans le monde. Paris, Presses Universitaires de France 1961.

[9] GREIMAS, A. J.: La structure des actans du récit. Essai d'approche générative. Word, *23*, 1967, p. 221—238.

[10] HARRIS, Z. S.: From phoneme to morpheme. Language, *31*, 1955, No. 2, p. 190—222.

[11] JANSEN, St.: Sur les rôles des personnages dans Andromaque. Orbis Litterarum, *22*, 1967, No. 1—4, p. 77—87.

[12] JANSEN, St.: Analyse d'Andromaque. Revue Romane, *3*, 1968, p. 16—29.

[13] JANSEN, St.: Esquisse d'une théorie de la forme dramatique. Langages, *12*, décembre 1968, p. 71—93.

[14] KALMÁR, L.: Matematika és színház. S. Marcus bukaresti professor kutatásai. Dél-Magyar ország, *56*, 1966, No. 263, p. 8.

[15] MARCUS, S.: Modèles mathématiques dans l'étude du drame. T. A. Informations, 1967, No. 2, Paris, p. 86—87.

[16] MARCUS, S.: Algebraic Linguistics. Analytical Models. New York—London, Academic Press 1967.

[17] MARCUS, S.: Introduction mathématique à la linguistique structurale. Paris, Dunod 1967.

[18] MARCUS, S.: Metode matematice în studiul dramei. Strategia personajelor, I. Metodologia istoriei și criticii literare. București, Editura Academiei R. S. R. 1969, p. 163—170.

[19] MARCUS, S.: Metode matematice în studiul dramei. Strategia personajelor, II. Revista de istorie și teorie literară, *18*, 1969, No. 4, p. 649—657.

[20] MARCUS S.: Poetica matematică. Bucuresti, Editura Academiei R. S. R. 1970.

[21] MATACHE, G. A.: Proiecția tragică—element compozițional în Romeo și Julieta. Studii de literatură universală, *12*, 1968, p. 15—25.

[22] MORENO, J. C.: Die Grundlagen der Soziometrie. 1954.

[23] POLTI, G.: Les 36 situations dramatiques. Mercure de France, 1934.

[24] PRIETO, L.: Langue et style. La linguistique, 1969, No. 1, p. 5—24.

[25] RADCLIFFE, A. I. I.: Romeo and Juliet. London 1946.

[26] RUDEANU, S.: Notes sur l'existence et l'unicité du noyau d'un graphe. Revue française de recherche opérationelle, 1964, No. 33, p. 345—352.

[27] SACKSTEDER, W.: Éléments du modèle dramatique. Diogène, vol. 52, 1965, p. 29—60.

[28] SCHÉRER, J.: Tartuffe. Histoire et structure. Paris, Les course de Sorbonne, C. D. U. 1965, p. 50.

[29] SOURIAU, É.: Les différents modes d'existence. Paris 1943.

[30] SOURIAU, É.: Les deux cent mille situations dramatiques. Bibliothèque d'Esthétique. Paris, Flammarion 1950.

[31] TESNIÈRE, L.: Éléments de syntaxe structurale. Paris, C. Klincksieck 1959.

Étude algébrique comparative de la structure syntaxique et sémantique des variantes d'un texte poétique

LIANA SCHWARTZ, BUCURESTI

1. Une poésie est d'habitude le point terminus d'une succesion de variantes qui, pour des raisons diverses, ont provoqué le mécontentement de l'auteur. Mais il y a des cas où l'auteur envisage comme valables plusieurs variantes, des cas où il ne peut pas se décider pour une variante unique et alors il propose au lecteur plusieurs expressions approchées de sa pensée poétique. C'est le cas de la poésie *Mai am un singur dor* (*Je n'ai qu'un seul désir*), du grand poète roumain Mihai Eminescu. Le poète a proposé encore trois variantes de cette poésie: „*De-oi adormi curînd*" („*Si je m'endors bientôt*"); „*Nu voi mormînt bogat*" („*Je ne veux pas une tombe riche*") et „*Iar cînd voi fi pămînt*" („*Et lorsque je serai poussière*). Désignons ces variantes, dans l'ordre où elles ont été spécifiées, par *A*, *B*, *C*, et *D*. Chacune de ces variantes a un grand nombre de sous-variantes, de sorte que l'ensemble de toutes ces variantes et sous-variantes se lève à 40. Nous sommes donc en présence d'une véritable arborescence de variantes et sous-variantes, qui nous permettent de pénétrer dans le laboratoire de création de ce grand poète, de suivre ses pensées et ses hésitations.

Mais une telle situation nous confronte avec quelques questions difficiles, qui exigent une réponse. Qu'est-ce-que c'est une variante? Combien et de quelle manière doivent se rapprocher deux textes poétiques afin que l'un d'eux soit considéré une variante de l'autre? En quelle situation peut-on affirmer qu'une poésie est subordonnée à une autre poésie? Peut-on fixer dans cet ordre d'idées une frontière tranchante, ou s'agit-il ici d'une question de gradation? Mais, dans ce dernier cas, comment doit-on améliorer les critères d'analyse, afin qu'ils soient assez sensibles et, en présence de quelques dizaines de textes „soupçonnés" d'être variantes d'un texte donné, ces critères donnent des paramètres assez fins pour permettre une hiérarchie rigoureuse, du point de vue de leur degré de dépendence?

Sans doute, on arrive ainsi à d'autres questions, qui constituent l'essence même d'une recherche scientifique du langage poétique: Quelles sont les structures fondamentales d'une oeuvre poétique? Qulles sont les coordonnées qui définissent une telle oeuvre, qui délimitent le poétique du non poétique? Ce sont justement ces structures et ces coordonnées qui devraient fournir les critères de la hiérarchie cherchée. Il ne s'agit pas de résoudre ici ces problèmes et peut-être une solution générale de ces pro-

blèmes n'existe pas. Nous nons contentons d'adopter une solution ad-hoc, de detecter quelques structures qui, dans les poésies envisagées, remplissent une fonction poétique importante et de dégager, de l'examen de ces structures, quelques paramètres dont l'utilisation fournira la hiérarchie cherchée. Mais il faut décourager le lecteur au moins d'un certain point de vue. Il est aisé de comprendre que la recherche comparée de 40 variantes d'un texte ne peut être conçue sans l'utilisation d'un ordinateur. En absence de cet aide précieux, nous avons limité notre recherche à une partie seulement de ces 40 variantes, notamment les variantes A, B, C, D et les 13 sous-variantes de la variante B.

Nous prenons comme point de départ cinq types de structures: la structure d'ordre (concernant le rang du vers dans le texte), la structure prosodique (concernant le type de versification), la structure syntaxique (concernant la nature des catégories syntaxiques utilisées), la structure lexicale (concernant l'identité ou la non-identité des mots pleins — aux sens de Pierre Guiraud [4] — utilisés dans les textes comparés) et la structure du contenu (concernant l'appartenance ou la non-appartenance des mots au même champ sémantique, utilisés dans les textes comparés, conçus au niveau dénotatif). Étant donné deux textes poétiques de la même longueur (mesurée en nombre de vers) et entre lesquels on a une certaine correspondance vers par vers, on introduit une numérotation de sorte que deux vers qui sont en correspondance aient le même numéro. Si la longueur commune des deux textes envisagés est égale à n, on associe à chaque numéro p, $1 \leqq p \leqq n$, une suite de cinq chiffres 0 et 1, de la manière suivante. Le chiffre de rang i ($1 \leqq i \in 5$) dans cette suite est égale à 0 si les vers numérotés avec p sont identiques du point de vue de la structure de rang i; ce chiffre est 1 dans le cas contraire ([11], ch.VII).

Prenons, par exemple, le vers *Doar toamna glas să dea* (traduction mot à mot: *Seulement l'automne donne la voix*) de rang 15 dans la variante A et son correspondant dans la variante B *Doar moartea glas să dea* (traduction mot à mot: *Seulement la mort donne la voix*), dont le rang dans B est égal à 23. Les rangs de ces vers dans les deux variantes sont différents, donc le premier chiffre, correspondant à la structure d'ordre, est égal à 1. Les structures prosodique et syntaxique sont les mêmes, donc les deux chiffres suivants sont 0. La structure lexicale est différente, donc le quatrième chiffre est 1. Les mots *automne* et *mort* appartient au même champ sémantique (disons, le champ sémantique de la decrépitude, de la décadence, de l'épanouissement), donc le cinquième chiffre est 0.

De cette sorte, les deux textes donnent naissance à ce qu'on appelle un code binaire. Le nombre des mots de ce code est égal à n, c'est-à-dire à la longueur commune des deux textes, tandis que la longueur des mots du code est égale à 5, c'est-à-dire au nombre de structures envisagées. La différence de structure entre les vers numérotés avec p est donnée par le nombre de chiffre 1 dans le mot d'ordre p du code envisagé. La somme de ces n différences sera, par définition, la distance structurale

entre les deux textes soient-ils X et Y, et sera désignée par $\alpha(X, Y)$. Dans le cas où les deux textes ne sont pas exactement de la même longueur, on peut utiliser un procédé proposé par Dom. J. Froger [3], afin de corriger le résultat. (Voir les détails dans [13].) On a trouvé dans [13] les résultats suivants: 1) $\alpha(B\ .\ C) = 82$, 2) $\alpha(B, D) = 109$, 3) $\alpha(A, B) = 117$, 4) $\alpha(A, D) = 119$, 5) $\alpha(A, C) = 137$, 6) $\alpha(C, D) = 144$.

Désignons pour chaque variante X, par $\alpha(X)$ la somme entre les numéros d'ordre des distances structurales qui engagent, dans leur expression, la variante X. On obtient $\alpha(A) = 3 + 4 + 5 = 12$, $\alpha(B) = 1 + 2 + 3 = 6$, $\alpha(C) = 1 + 5 + 6 = 12$, $\alpha(D) = 2 + 4 + 6 = 12$. On constate donc la priorité de B, mais l'impossibilité de hiérarchiser les variantes A, C et D [9]. Afin d'obtenir une hiérarchie plus fine des variantes A, B, C et D, nous allons définir, pour chaque variante X, un nouveau paramètre, $\beta(X)$, qui sera, par définition, la somme des distances structurales de X par rapport aux autres trois variantes. On a $\beta(A) = \alpha(A, B) + \alpha(A, C) + \alpha(A, D) = 117 + 137 + 119 = 373$, $\beta(B) = 308$, $\beta(C) = 363$, $\beta(D) = 372$. On obtient donc la hiérarchie suivante: 1. $B(\beta[B] = 308)$, 2. $C(\beta[C] = 363)$, 3. $D(\beta[D] = 372)$, 4. $A(\beta[A] = 373)$. Cette hiérarchie, tout en respectant la hiérarchie précédante, donne la possibilité de départager les variantes A, C et D. On constate toutefois que les variantes A et D restent sensiblement égales, la variante C étant la seule qui est effectivement départagée. Mais C reste bien loin de B, ce qui confirme d'une manière plus convaincante le caractère plus représentatif de la variante B par rapport aux variantes A, C et D. En tout cas, la dernière place occupée par A est en contraste avec le fait que la variante A est présentée, d'habitude, comme la plus importante. Ce fait même montre qu'il faut chercher d'autres critères de hiérarchiser les variantes et voir en quelle mesure ces nouvelles hiérarchies coïncident avec celle déjà trouvée.

2. Les traits sémantiques des mots et des syntagmes utilisés dans les variantes A, B, C et D correspondent à certaines catégories sémantiques. D'après la remarque de F. Kiefer [7], un ensemble K, de catégories sémantiques, pour être utile dans la recherche, doit remplir les deux conditions suivantes: a) il doit être fini; b) les catégories de K doivent être pertinentes du point de vue linguistique. Une catégorie est pertinente du point de vue linguistique s'il existe au moins deux morphèmes dont la distinction est donnée justement par la présence (respectivement l'absence) de la catégorie sémantique envisagée. Selon Kiefer on introduit dans un ensemble de catégories sémantiques une relation de domination, de la manière suivante: Si tout morphème qui correspond à la catégorie sémantique C_j correspond aussi à la catégorie sémantique C_i, alors on dit que C_i domine C_j. Il est aisé de voir que la catégorie dominante est plus générale que la catégorie dominée. L'analyse sémantique des variantes A, B, C et D nous a conduit à envisager l'ensemble de catégories sémantiques représenté dans la fig. 1, ensemble qui remplit les deux conditions posées par Kiefer. Les catégories sémantiques sont ici représentées à l'aide d'une arborescence dont le

centre correspond à la catégorie très générale *OBJET*, tandis que l'ordre de haut en bas est celui du général vers le particulier. Les catégories situées sur la même ligne ont le même degré de généralité. La catégorie *OBJET* est la seule qui n'est dominé par aucune autre catégorie. Toute autre catégorie est dominée par une seule catégorie, située à un niveau immédiatement supérieur. L'*ABSTRAIT* et le *CONCRET* sont dominés par l'*OBJET*; le *TERRESTRE* et le *EXTRATERRESTRE* sont dominés par le *CONCRET* et ainsi de suite.

L'arborescence sémantique adoptée ci-dessus nous permet d'utiliser certaines notions et méthodes qui ont prouvé déjà leur efficacité dans l'étude du langage poétique. Nous avons spécialement en vue le fameux principe de Roman Jakobson, qui envisage la poésie comme le résultat d'une projection spécifique de l'axe paradigmatique sur l'axe syntagmatique du langage [6]. On trouve d'ailleurs une conception semblable chez R. Barthes [1] (voir, dans ce sens, la discussion de [9]) et, récemment, chez Walter A. Koch [8]). La manière concrète d'appliquer le principe de Jakobson-Barthes dans l'analyse des variantes A, B, C et D sera la suivante: Nous acceptons que le degré de poéticité d'un texte est donné par le degré de nonconcordance entre la distance paradigmatique et la distance syntagmatique des termes dont la juxtaposition forme les figures poétiques du texte envisagé. Mais parceque la distance syntagmatique est toujours très petite dans ces cas, c'est la valeur de la distance paradigmatique qui mesure le degré de figuration, de poéticité de l'expression envisagée. Il faut donc trouver une méthode de mesurer la distance paradigmatique entre deux termes x et y. Ce sont les suggestions de N. Chomsky concernant les degrés de grammaticalité qui nous donnent cette méthode [2]. Associons à chaque terme la catégorie sémantique la plus particulière figurant dans l'arborescence sémantique et concernant le terme envisagé. La distance, dans le sens de la théorie des graphes, entre le sommet qui correspond à la catégorie associée à x et le sommet qui correspond à la catégorie asssociée à y sera, par définition, la distance paradigmatique entre x et y. (Cette définition nous a été proposée par S. Marcus.) En effet, cette distance est très complexe, car elle tient compte à la fois du degré de généralité et du degré d'hétérogéneité des traits sémantiques envisagés (voir, pour une discussion plus détaillée et plus profonde, le livre [11]). Par exemple, dans l'expression (rencontrée dans *Mai am un singur dor*) *glas să dea frunzişului* ((que l'automne) *donne la voi au feuillage*) *glas* (*voix*) correspond à la catégorie *HUMAIN*, tandis que *frunziş* (*feuillage*) correspond à la catégorie VÉGÉTAL, donc la distance paradigmatique est égale à 5 (voir la figure 1). En effet, en partant de *HUMAIN*, pour arriver à VÉGÉTAL, il faut parcourir les sommets *ANIMÉ*, *TERRESTRE*, *INANIMÉ*, *UNIREGNE*, donc la chaîne ainsi obtenue est de longueur égale à 5. Il faut remarquer que la distance paradigmatique ainsi définie tient compte à la fois du degré d'hétérogénéité des catégories envisagées et de leur degré de généralité. Par exemple, *HUMAIN* et *INANIMÉ* ont le même degré d'hétérogénéité que *ANIMÉ* et *INANIMÉ*, mais le degré de généralité de *HUMAIN* est inférieur au degré de généralité de *INANIMÉ*,

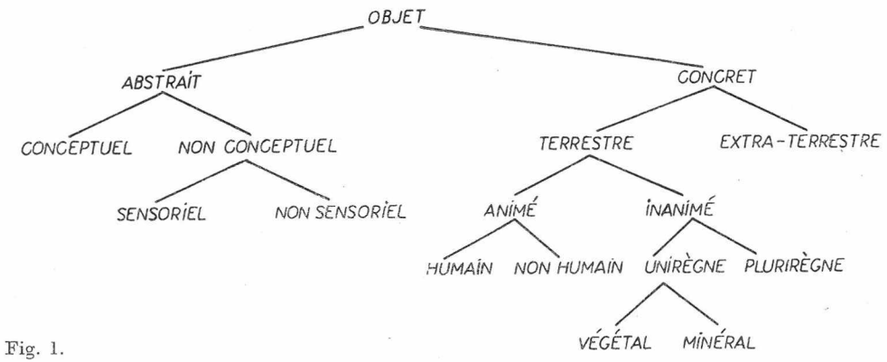


Fig. 1.

tandis que *ANIMÉ* et *INANIMÉ* ont le même degré de généralité. C'est pour cela que la distance paradigmatique entre *HUMAIN* et *INANIMÉ* est égale à 3, tandis que la distance paradigmatique entre *ANIMÉ* et *INANIMÉ* est égale à 2.

Nous avons envisagé dans [15], pour les figures rencontrées dans les variantes A, B, C et D, les catégories sémantiques associées et leurs distances paradigmatiques. La somme de ces distances calculée pour chaque variante prise à part, va donner son degré de figuration, de poéticité. La hiérarchie des variantes sera donc donnée par l'ordre décroissant de ces sommes.

On obtient la hiérarchie suivante: B — 75, C — 53, A — 48, D — 39. Pour les détails voir [15]. On constate donc de nouveau que la variante B est, loin, sur la première place, tandis que C est sur la deuxième place. Le seul changement de places concerne les variantes A et D.

Il y a dix années, T. Vianu remarquait qu'il y a deux périodes distinctes dans l'évolution de la poésie de Eminescu. La première période est caractérisée par une terminologie du domaine de l'art, par le caractère monumental des constructions (dont un prototype est le poème *Mélancolie*), tandis que la deuxième période est caractérisée par une terminologie du domaine de la nature, dont une illustration est le poème *Mai am un singur dor*. L'analyse faite ci-dessus montre bien la justesse de la remarque de Tudor Vianu, en ce qui concerne *Mai am un singur dor* (et ajoutons-nous, ses variantes aussi), où le végétal, le minéral et l'extraterrestre dominent les figures poétiques.

3. Une autre méthode de hiérarchiser les variantes A, B, C et D repose sur le degré de connexité des diverses catégories grammaticales. Nous allons diriger notre attention vers deux catégories qui semblent être primordiales dans ces quatre variantes: la catégorie du temps et la catégorie de la personne. Étant donnée une catégorie K susceptible de prendre les valeurs V_1, $V_2 \ldots V_n$, nous dirons (selon une proposition de S. Marcus [11]) qu'une partie P d'un texte T est connexe par rapport à K (ou K-connexe) s'il existe une valeur $i (1 \leqq i \leqq n)$ telle que chaque occurence de la catégorie K

dans P possède la valeur V_i. Une partie connexe maximale (c'est-à-dire, qui n'est pas contenue dans une autre partie connexe) est dite une composante K-connexe du texte T, ou une composante connexe de T, par rapport à la catégorie K, ou simplement une composante.

Le nombre des composantes des divers temps, pour les variantes A, B, C, et D est donné dans le tableau 1.

Tableau 1

	Indic. prés.	Subj. prés.	Futur
A	4	4	2
B	4	3	2
C	4	3	1
D	1	3	3

On peut même préciser la structure des composantes connexes. Par exemple, pour la variante A la première composante connexe du subjonctif présent va du vers de rang 2 jusqu'au vers de rang 8, la deuxième va du vers de rang 11 jusqu'au vers de rang 15, la troisième va du vers de rang 19 jusqu'au vers de rang 24, tandis que le vers de rang 32 donne la quatrième composante. La première composante de l'indicatif présent est donné par le vers de rang 1, la deuxième composante est donnée par les vers de rang 9 et 10, la troisième par le vers de rang 17, tandis que la quatrième est donné par le vers de rang 29. La première composante connexe du futur est formée par les vers de rang 25 à 27, tandis que sa deuxième composante va du vers de rang 33 jusqu'au vers de rang 35. Pour les variantes B, C et D la structure détailée des composantes connexes est donnée dans [15].

Ce qu'il y a de commun à toutes les variantes c'est l'itération de la boucle formée par le présent de l'indicatif et le présent du subjonctif. Ce qu'il y a de différent c'est la manière dont cette itération est interrompue par certaines excursions vers le futur. (Voir les figures 2, 3, 4 et 5 donnant la succession des temps dans les variantes A, B, C et, respectivement D.) La succession des temps est, dans le déroulement de la variante A, la suivante: présent de l'indicatif, présent du subjonctif, présent de l'indicatif, présent du subjonctif, présent de l'indicatif, présent du subjonctif, futur, présent de l'indicatif, présent du subjonctif, futur. C'est un chemin dans le graphe défini par les flèches de la figure 2, un chemin où la boucle présent de l'indicatif — présent du subjonctif — présent de l'indicatif est parcourue plusieurs fois, avec deux excursions vers le futur. La succession des temps est, dans le déroulement de la variante B, la suivante: futur, subjonctif présent, indicatif présent, subjonctif présent, indicatif présent, subjonctif présent, futur, indicatif présent ce qui représente un chemin dans le graphe

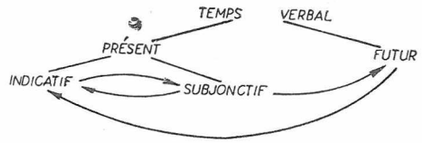


Fig. 2.

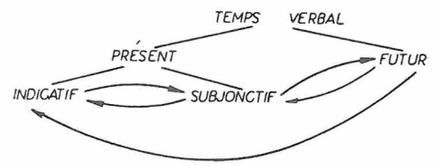


Fig. 3.

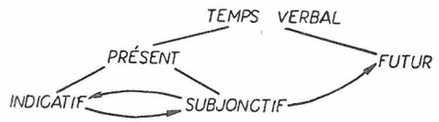


Fig. 4.

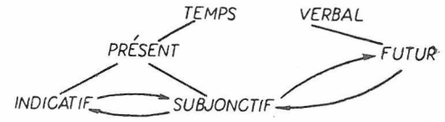


Fig. 5.

defini par les flèches de la figure 3. La succession des temps dans la variante C est la suivante: présent de l'indicatif, présent du subjonctif, présent de l'indicatif, présent du subjonctif, présent de l'indicatif, présent du subjonctif, futur, ce qui représente un chemin dans le graphe des flèches de la figure 4. La succession des temps dans la variante D est la suivante: futur, présent du subjonctif, présent de l'indicatif, présent du subjonctif, futur, présent du subjonctif, futur. C'est un circuit dans le graphe des flèches de la figure 5. C'est dans la variante C que le futur a la moindre importance, sa seule occurence se produisant à la fin du poème. C'est dans la variante D que le futur a la meilleure importance, c'est lui qui couvre ici, à la fois, le commencement et la fin du poème, c'est lui aussi qui forme, avec le présent du subjonctif, une boucle qui ne peut être rencontrée dans aucune des variantes A, B et C.

Il est aisé de voir (en regardant le tableau 1) que la variante A est la plus nonconnexe, car elle comporte 10 composantes de connexion. On a ensuite la variante B, avec 9 composantes de connexion, la variante C avec 8 composantes et, enfin, la variante D avec 7 composantes. C'est donc la variante A, qui se présente, de ce point de vue, la plus figurée. Mais si nous voulons établir quelle est la variante la plus représentative du point de vue de la structure des temps, il faut recourir à une autre méthode, dont la suggestion se trouve dans la théorie des codes. D'après le tableau 1, chaque variante est une suite ordonnée de trois chiffres. Définissons une distance entre deux variantes, égale au nombre de positions où ces variantes comportent des chiffres différents. En désignant par d cette distance, on a donc $d(A, B) = 1$, $d(A, C) = 2$, $d(A, D) = 3$, $d(B, C) = 1$, $d(B, D) = 2$, $d(C, D) = 2$. Posons $d(A) = d(A, B) + d(A, C) + d(A, D)$, $d(B) = d(B, A) + d(B, D)$, $d(C) = d(C, A) + d(C, B) + d(C, D)$, $d(D) = d(D, A) + d(D, B) + d(D, C)$. On a $d(A) = 6$, $d(B) = 4$, $d(C) = 5$, $d(D) = 7$. On peut admettre que la variante la plus représentative, du point de vue de la structure des catégories temporelles, est justement la variante la plus proche, par rapport aux autres variantes, du point de vue de la distance d. C'est donc l'ordre croissant des valeurs de d qui donne la hiérarchie cherchée des variantes A, B, C et D. Cet ordre est: B, C, A, D, ce qui prouve de nouveau la supériorité de la variante B. Il faut remarquer aussi que la hiérarchie ainsi obtenue est justement la hiérarchie

obtenue dans le paragraphe précédent, à l'aide de la distance paradigmatique. Le fait que deux critères tout à fait différents conduisent à la même hiérarchie des variantes montre la raison profonde de cette hiérarchie.

Mais quel est le sens poétique des structures des temps que nous venons de mettre en évidence? L'oscillation entre le présent de l'indicatif et le présent du subjonctif est en fait l'oscillation entre les faits accomplis et les états exprimant un désir et qui restent sous le signe de l'incertitude. Traduire cette oscillation par la boucle rencontrée dans tous les quatre graphes envisagés c'est préfigurer une récursivité, une virtuelle répétition à l'infini de cette hésitation entre les deux temps. C'est une méthode souvent utilisée dans les modèles mathématiques: de remplacer le fini par l'infini, afin de mieux mettre en évidence la tendance d'un certain procéssus. (Un tel procédé, dans l'étude des personnages dramatiques, a été proposé par S. Marcus dans [10], p. 245 et [11], cap. VIII.)

4. Nous allons soumettre la catégorie de la personne à un traitement analogue à celui appliqué à la catégorie du temps. Le mode spécifique de succession des diverses valeurs de cette catégorie, dans les variantes A, B, C et D, est très pertinent du point de vue artistique et nous allons discuter cet aspect à la fin de ce paragraphe.

Désignons par P_1, P_2 et P_3 la première, la deuxième et, respectivement, la troisième personne du singulier et par R_1, R_2 et R_3 la première, la deuxième et, respectivement, la troisième personne du pluriel. Le nombre des composantes connexes de ces six membres de la catégorie de la personne, pour chacune des variantes A, B, C et D, est donné dans le tableau 2.

Tableau 2

	P_1	P_2	P_3	R_1	R_2	R_3
A	5	0	5	0	2	2
B	3	0	2	0	2	2
C	2	0	3	0	2	4
D	5	0	5	0	2	4

Un trait commun de toutes les variantes c'est l'absence de la deuxième personne du singulier et de la première personne du pluriel. La première personne est reservée à la confession du poète, ce qui exige toujours le singulier, pour mieux marquer le caractère très personnel de cette confession. L'interlocuteur du poète est toujours l'humanité toute entière, il n'a jamais en vue une personne déterminée, il s'adresse à tous ceux qui l'entourent, à tous ceux qui le connaissent, ce qui exige que la deuxième personne soit toujours au pluriel. Il faut d'ailleurs remarquer le caractère stable, invariant, de cette catégorie; car en effet, la deuxième personne du pluriel

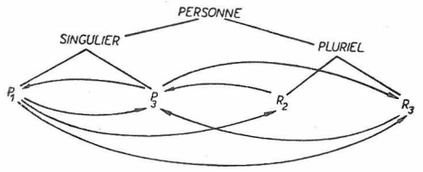


Fig. 6.

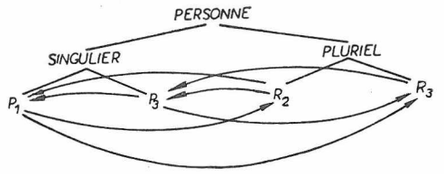


Fig. 7.

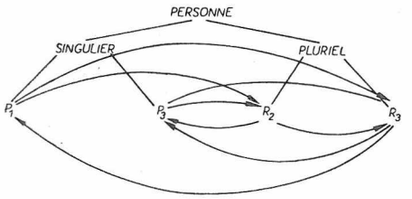


Fig. 8.

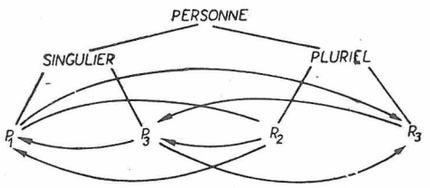


Fig. 9.

est la seule catégorie de personne dont le nombre des composantes connexes reste toujours le même, ce nombre étant égal à 2 pour chaque variante. (Voir aussi le tableau 2.) D'autre part, on pourrait dire que les variantes A et D sont isomorphes en ce qui concerne les catégories de personnes du singulier (le nombre des composantes de P_1 et de P_3 étant le même dans A et D), tandis que les variantes A et B sont isomorphes du point de vue des catégories de personnes du pluriel (le nombre des composantes de R_2 et R_3 étant le même dans A et B). Le même isomorphisme du point de vue du pluriel existe entre les variantes C et D.

La succesion des divers membres de la catégorie de la personne, dans les variantes A, B, C et D est donnée dans les figures 6, 7, 8 et 9.

Pour faire la hiérarchie des variantes du point de vue de leur caractère représentatif en ce qui concerne la structure des personnes, nous allons prendre comme point de départ les données du tableau 2, où chaque variante est un mot de longueur égale à 6 dans un code numérique. En imitant la distance de Hamming, utilisée dans la théorie des codes binaires, nous définissons la distance entre deux variantes comme égale au nombre de positions (dans le tableau 2) où les chiffres correspondants sont différents [11]. En désignant par d cette distance, nous avons $d(A, B) = 2, d(A, C) = 3, d(A, D) = 1, d(B, C) = 3, d(B, D) = 3, d(C, D) = 2$. Posons $d(A) = d(A, B) + d(A, C) + d(A, D), d(B) = d(B, A) + d(B, C) + d(B, D), d(C) = d(C, A) + d(C, B) + d(C, D), d(D) = d(D, A) + d(D, B) + d(D, C)$. On a $d(A) = 6, d(B) = 8, d(C) = 8, d(D) = 6$. Afin de pouvoir départager les variantes A et D, d'une part, et les variantes B et C d'autre part, nous allons recourir au nombre de composantes connexes de chaque variante. Ce nombre est égal à 14 pour A, à 9 pour B, à 11 pour C et à 16 pour D, la moyenne étant donc 12,50. En prenant comme point de départ l'écart par rapport à cette valeur moyenne on a la hiérarchie C, A et, sur la dernière place, B et D à la fois. Mais B et D ont été déjà départagées à l'aide de la distance précédente.

On a donc la hiérarchie finale (où l'ordre des valeurs de d est celui nondécroissant): A, D, C, B. Le fait le plus intéressant ici c'est la situation de B. Cette variante, qui du point de vue de la distance paradigmatique et des structures des temps était la plus représentative, s'avère, du point de vue de la structure des personnes grammaticales, la moins représentative. Ce fait est dû au nombre trop réduit de composantes connexes, donc aux variations trop faibles de la catégorie des personnes, dans la variante B. Cette faiblesse peut être rencontrée dans la variante C aussi, mais seulement en ce qui concerne des catégories de personnes du singulier; elle peut être rencontrée aussi dans la variante A, mais seulement en ce qui concerne les catégories de personnes du pluriel. C'est B la seule variante où cette faiblesse concerne le singulier et le pluriel à la fois.

Conclusions

Nous avons obtenu quatre hiérarchies des variantes A, B, C *et* D, hiérarchies qui correspondent à quatre points de vue, à savoir: la distance structurale, la distance paradigmatique, la structure des temps et la structure des personnes. Les résultats obtenus sont synthétisés dans le tableau 3.

Tableau 3

	1	2	3	4
Dist. struct.	B	C	D	A
Dist. parad.	B	C	A	D
Struct. des temps	B	C	A	D
Struct. des pers.	A	D	C	B

Il s'agit maintenant de déduire, de ces quatre hiérarchies, une hiérarchie unique, qui tient compte simultanément de tous les paramètres envisagés. On pourrait procéder de la manière suivante. Associons, à chaque variante, un nombre égal à la somme des numéros d'ordre de cette variante dans les quatre hiérarchies envisagées. On a alors: $B - 7$, $C - 9$, $A - 11$, $D - 13$. En effet, B occupe la première place dans les trois premières hiérarchies et la quatrième place dans la quatrième hiérarchie, donc le nombre associé à B est $1 + 1 + 1 + 4$. D'une manière analogue on trouve les nombres associés à C, A et D. On obtient donc la hiérarchie finale et de synthèse: B, C, A, D, ce qui prouve de nouveau et définitivement la priorité de la variante B par rapport à toutes les autres variantes.

On pourrait procéder d'une manière plus savante, en appliquant, par exemple (selon une proposition de S. Marcus) la méthode de Condorcet, utilisée dans l'analyse

algébrique d'un scrutin [3]. Ecrivons $X > Y$ pour dire que X est antérieur à Y, dans une certaine hiérarchie. On constate alors que $B > A$, $B > C$, $B > D$, $C > A$, $C > D$ d'après les trois premières hiérarchies et $A > D$ d'après les trois dernières hiérarchies, c'est-à-dire que toutes ces relations d'antériorité ont l'avis de la majorité des critères adoptés. En tenant compte de la transitivité de la relation $>$, on obtient les relations $B > C > A > D$, qui ont l'avis de la majorité (mais aucune d'elles de l'unanimité!) des critères et qui retrouvent la hiérarchie finale, de synthèse, déjà établie ci-dessus par une méthode directe. Mais, sans doute, l'application de la méthode de Condorcet est ici assez triviale, à la suite du nombre réduit de critères et de variantes.

LITERATURE

[1] BARTHES, R.: Le degré zéro de l'écriture, suivi de éléments de sémiologie. Paris, Editions Gouthier 1965.

[2] CHOMSKY, N.: Degrees of grammaticalness. The structure of language. In: Philosophy of Language, Eds. J. A. Fodor and J. J. Katz. New Jersey, Prentice-Hall, Inc. Englewood Cliffs, p. 384—389.

[3] FROGER, D. J.: La critique des textes et son automatisation. Paris, Dunod 1968.

[4] GUILBAUD, G. Th.—ROSENSTIEHL, P.: Analyse algébrique d'un scrutin. Mathématiques et sciences humaines, 1963, No. 4, p. 9—13.

[5] GUIRAUD, P.: Les caractères statistiques du vocabulaire. Paris, Presses Universitaire de France 1954.

[6] JAKOBSON, R.: Linguistics and poetics. Style in languages. Ed. Thomas A. Sebeok. New York, 1960, p. 350—377.

[7] KIEFER, F.: Some semantic relations in natural languages. Foundations of Language, *2*, 1966, No. 3, p. 228—240.

[8] KOCH, W. A.: Recurrence and a three-models approach to poetry. The Hague—Paris, Mouton & Co. 1966.

[9] MARCUS, S.: Les écarts dans le langage poétique. Revue roumaine de linguistique, *13*, 1968, No. 5.

[10] MARCUS, S.: Lingvistica, stiinţă pilot. Studii şi cercetări lingvistice, *20*, 1969, No. 3, p. 235.

[11] MARCUS, S.: Poetica matematică. Bucureşti, Editura Academiei R. S. România 1970.

[12] RUWET, N.: Limites de l'analyse linguistique en poétique. Langages, 1968, No. 12, p. 56—70.

[13] SCHWARTZ, L.: Studiul matematic al variantelor poeziei "Mai am un singur dor" de Mihai Eminescu. Studii şi cercetări matematice, *22*, 1970.

[14] SCHWARTZ, L.: Théorie des graphes et analyse littéraire (Étude des variantes de la poésie, "Mai am un singur dor" de Mihai Eminescu). Bullet in Mathématique de la Société des Sciences mathématiques de la R. S. R., *12*, 1969.

[15] SCHWARTZ, L.: Écart sémantique, structure des temps et structure des personnes dans le langage poétique. Revue roumaine de linguistique, 1970, No. 1.

[16] VIANU, T.: Statistica lexicală si o problemă a vocabularului eminescian. Limba română, *8*, 1959, No. 3, p. 25—33.

Un modèle markovien de l'influence à distance dans les langues naturelles

MIHAI DINU, BUCUREŞTI

1. Introduction

L'étude de la propagation le long d'un texte de l'influence d'une unité linguistique connue (on se rapportera en ce qui suit aux lettres, mais la problème se pose de la même sorte pour les phonèmes, les morphèmes, les syllabes et les mots) a été abordée jusqu'à présent par deux voies, illustrées par des méthodes différentes.

La première méthode part d'une statistique exhaustive des combinaisons de 2, 3, ..., n lettres de la langue écrite respective. En connaissant les fréquences des occurrences de toutes les séquences possibles de lettres dans un grand nombre de textes représentatifs, on calcule les entropies correspondantes d'ordre 1, 2, 3, ..., n ($H_1, H_2 \ldots H_n$). La suite $H_1, H_2 \ldots H_n, \ldots$ est décroissante et convergente. Autrement dit pour une valeur $\varepsilon > 0$, quelque petite qu'elle soit, il existe un H_i tel que $H_i - H_{i+1} < \varepsilon$. La signification linguistique de cette convergence est qu'après i positions l'influence de la lettre initiale s'éteint et la $(i + 1)$-ème lettre ne dépend plus d'aucune façon de celle-ci.

L'obstacle principal, actuellement insurmontable, qui apparaît quand il s'agit d'appliquer cette méthode est constitué par l'énorme volume de travail nécessaire pour déterminer les fréquences de toutes les combinaisons de lettres, puisque le nombre de celles-ci augmente exponentiellement avec l'ordre de l'entropie qu'on veut calculer. Pour obtenir des valeurs des fréquences suffisamment rapprochées des probabilités dans la langue des unités considérées, il est nécessaire d'inventorier des textes extrêmement longs, de sorte qu'on arrive à dépasser à un moment donné même les possibilités actuelles du calcul automatique.

La seconde méthode, celle des expériments de prédiction, initiée par Shannon en 1951 [7], essaie d'escamoter la difficulté mentionnée plus haut, en faisant appel à la compétence linguistique d'une personne quelconque, à laquelle on propose de deviner dans certaines conditions des lettres d'un texte partiellement connu. On obtient ainsi, avec une certaine précision (l'erreur peut être calculée), la distance moyenne jusqu'à laquelle se propage l'information fournie par une lettre.

Mais ce que cette méthode ne peut pas nous offrir c'est une description de la manière dont s'accomplit le transport d'information le long du texte. A l'origine de ce

transport se trouvent les contraintes imposées par la langue aux combinaisons de plusieurs lettres. Il est connu que dans la constitution des digrammes, trigrammes, tétragrammes etc., dans les langues naturelles interviennent certaines restrictions, qui facilitent la prédiction des lettres suivantes. Chaque combinaison de lettres apporte une contribution qui lui est propre à la propagation à distance de l'information. Par exemple, il est évident que si deux lettres quelconques x_i et x_j apparaissent avec les probabilités $p(x_i)$ et $p(x_j)$ et si la probabilité de l'apparition du digramme x_ix_j est dans la langue respective $p(x_ix_j) = p(x_i) \, . \, p(x_j)$, ce digramme ne contribue nullement à la transmission à distance de l'information fournie par la lettre x_j, parce qu'il n'introduit aucune contrainte qui facilite la prédiction de la lettre suivante. Dans un pareil cas, si l'influence de la lettre x_i se manifeste pourtant plus loin dans le texte, le vehicule de cette information n'est pas le digramme considéré, mais une autre combinaison de lettres grevée par une certaine restriction linguistique. Mais dans la clarification de l'apport différencié propre à divers types de séquences de lettres de longueur différente, dans la propagation de l'information, les expériments de prédiction, qui ne permettent que des estimations globales, ne peuvent aucunement nous aider.

Pour suppléer cette lacune, nous proposons ci-dessous un modèle mathématique, capable de saisir et d'expliciter la contribution séparée des digrammes, trigrammes.... n-grammes, dans la transmission à distance de l'influence des lettres dans un texte. Ainsi que l'on verra, ce modèle permet de mettre en lumière quelques particularités de la communication par l'intermède de la langue, qui ont échappé jusqu'à présent aux recherches effectuées à l'aide des moyens traditionnels.

2. Déscription du modèle

Soit un système susceptible de prendre un nombre n d'états: $S_1, S_2, S_3 \ldots . S_n$.

A des moments déterminés le système passe brusquement d'un état à l'autre.

En principe, il est possible de passer de n'importe quel état à n'importe quel autre état, sans respecter l'ordre de l'énumération ci-dessus. Mais les probabilités que l'état S_i soit suivi par l'état S_j (nous les notons par p_{ij}) sont, dans le cas général différentes pour des valeurs différentes de i et j.

Evidemment, si entre deux états le passage direct n'est pas possible, la probabilité p_{ij} est nulle.

L'existence de notre système se déroule donc par des changements brusques, la situation à un moment donné étant dépendante de la situation du moment antérieur par l'intermède d'un tableau (matrice) de probabilités de passage:

$$\begin{vmatrix} p_{11} & p_{12} & p_{13} & \ldots\ldots\ldots & p_{1n} \\ p_{21} & p_{22} & p_{23} & \ldots\ldots\ldots & p_{2n} \\ \cdot & & & & \\ \cdot & & & & \\ \cdot & & & & \\ \cdot & & & & \\ p_{n1} & p_{n2} & p_{n3} & \ldots\ldots\ldots & p_{nn} \end{vmatrix} \tag{1}$$

On remarque que la somme de tous les éléments d'une ligne de la matrice ci-dessus doit être égale à 1. Au cas contraire, il faudrait admettre qu'on pût sortir à un certain moment en dehors du système, en passant à un état qui n'appartient pas à l'ensemble des états initiaux.

La succession des états parcourus par le système dans une période de temps déterminée constitue une *chaîne Markov*.

Si la matrice de passage est la même pour n'importe quels deux moments succesifs, c'est-à-dire si elle ne dépend pas du temps, la chaîne Markov respective s'appelle *stationnaire*.

Supposons qu'à un moment déterminé i il existe la probabilité p_1 que le système se trouve dans l'état S_1, la probabilité p_2 qu'il se trouve dans l'état S_2, ..., la probabilité p_n qu'il se trouve dans l'état S_n. On dit que cette phase du processus est caractérisée par le vecteur stochastique:

$$\left| \, p_1 \; p_2 \; p_3 \; \ldots\ldots\ldots \; p_n \, \right| . \tag{2}$$

Il va de soi que la somme des éléments de ce vecteur doit aussi être strictement égale à 1, pour que le système se trouve avec certitude dans l'un des états $S_1, S_2 \ldots . S_n$ (en mathématiques on dit que le système d'événements doit être complet).

Au moment suivant $i + 1$, la situation du système sera décrite par un autre, vecteur stochastique, différent en général du premier:

$$\left| \, p_1 \; p_2 \; p_3 \; \ldots\ldots\ldots \; p_n \, \right| . \tag{3}$$

Pour obtenir les éléments de ce nouveau vecteur, il sera suffisant d'effectuer, suivant les règles de multiplication de l'analyse matricielle [1], le produit entre le vecteur (2) et la matrice (1). Le vecteur qui caracterise le moment $i + 2$ sera obtenu par la multiplication du vecteur (3) par la même matrice de passage (nous avons supposé le processus stationnaire) et ainsi de suite.

Certains processus Markov jouissent de la propriété qu'ils tendent à atteindre, après un temps suffisamment long, une situation probabilistique stable, appelée *régime permanent* ou *régime limite*, pour laquelle les vecteurs stochastiques qui correspondent à deux moments succesifs deviennent identiques. Par conséquent en commençant par un certain moment, le déroulement du processus ne dépend plus de l'état initial du système. Cette propriété, qui est en effet une propriété de la

matrice de passage, s'appelle *ergodicité*. (Pour des détails concernant les notions définies ci-dessus voir [3].)

En revenant au problème qui nous préoccupe, imaginons un processus Markov à temps discret, dont l'ensemble des états soit l'ensemble des lettres de l'alphabet d'une certaine langue. (Le blanc, c'est-à-dire l'espace vide entre deux mots succesifs, est traité, lui aussi, comme une lettre.)

Tout texte écrit dans la langue respective pourra être considéré, en ce cas, comme une chaîne Markov. Il est alors évident que la matrice de passage sera donnée par la fréquence dans la langue écrite de tous les digrammes.

L'intuition nous dit qu'à une distance suffisamment grande, par rapport au commencement d'un texte, l'apparition d'une lettre ou d'une autre n'est plus influencée par la lettre initiale, mais bien gouvernée par les lois générales de fréquence de la langue.

Nous nous attendons donc à ce que la chaîne Markov ainsi définie soit ergodique. En nous basant justement sur cette propriété de la chaîne, nous essayerons de résoudre le problème de l'influence à distance des lettres dans un texte.

Supposons qu'un texte commence par la lettre a. Le vecteur stochastique correspondant à ce moment initial du processus sera:

$$| 1 \quad 0 \quad 0 \quad 0 \ldots\ldots\ldots 0 |, \tag{4}$$

parce que l'occurrence dans la position considérée de toute autre lettre, a excepté, est exclue.

Que pouvons-nous attendre pour le moment suivant? (Plus précisément dans la position suivante; nous considérons pourtant, ainsi qu'on l'a montré, le processus déroulé non pas spatialement, mais bien sur l'axe du temps, comme d'ailleurs ont lieu la rédaction et la lecture du texte.)

Si nous disposons d'une statistique des digrammes de la langue respective, il ne nous reste qu'à multiplier le vecteur (4) par la matrice de passage offerte par la statistique et nous obtiendrons la caractérisation de la nouvelle phase du processus, c'est à-dire les probabilités d'apparition de chaque lettre dans la position faisant immédiatement suite à un a. En appliquant successivement le même procédé nous pourrons déterminer l'évolution du processus en nous éloignant indéfiniment du point de départ. A partir du moment où le régime permanent s'établit, nous saurons que l'influence de la lettre initiale s'est éteinte.

De cette manière on déterminera le rayon d'action informationnelle de la lettre considérée. La même méthode nous permettra d'analyser l'influence à distance de chaque lettre prise à part, chaque fois le vecteur initial ayant le chiffre 1 placé dans la position qui correspond à la lettre respective.

Il serait sans doute intéressant non seulement d'établir le moment où le régime limite est atteint, mais aussi de connaître la manière dont on tend vers ce régime.

Nous obtiendrons une caractérisation suggestive des phases du processus, si nous faisons appel à la notion d'entropie informationnelle.

Ainsi qu'il est connu, étant donné un système complet d'événements dont les probabilités sont $p_1, p_2 \ldots\ldots p_n$, l'entropie du système est définie comme la somme

$$H = -p_1 \log_2 p_1 - p_2 \log_2 p_2 \ldots\ldots\ldots -p_n \log_2 p_n \tag{5}$$

où par $\log_2 p_i$ on a noté le logarithme dans la base 2 de la probabilité p_i.

Comme chaque vecteur stochastique qui caractérise le processus à un moment donné se rapporte à un système complet d'événements, on pourra calculer à l'aide de la formule (5) l'entropie correspondante au moment respectif. Les valeurs obtenues nous donneront une idée quant à la mesure dans laquelle la connaissance d'une lettre facilite la prévision des lettres suivantes. En effet, l'entropie exprime l'indétermination du processus et plus sa valeur sera petite, d'autant il sera plus facile de deviner l'état dans lequel se trouvera le système au moment respectif [4].

Nous nous attendrons donc à ce que l'entropie augmente de 0, à l'endroit de la lettre connue où l'état du système est parfaitement déterminé, vers la valeur moyenne de l'entropie de premier ordre de la langue.

Si on tient compte du fait qu'il est plus que probable que la transmission de l'information procurée par une lettre ait lieu aussi vers l'amont de manière semblable à sa propagation en aval par rapport à la lettre connue (nous avons désigné par amont et aval l'ensemble des lettres situées à gauche et respectivement à droite par rapport à la position considérée) on peut affirmer qu'en précisant une lettre on produit dans le niveau général de l'entropie moyenne du texte une sorte d'entonnoir, une dépression de longueur plus ou moins grande, selon que la lettre est plus ou moins influente.

Il n'est pas difficile de remarquer que pour calculer le rayon d'action des lettres vers l'amont, il est suffisant d'utiliser une matrice de passage résultée de la première par la transformation des lignes en colonnes.

En ayant soin d'amener la nouvelle matrice à une forme stochastique (la somme des éléments par lignes doit être égale à 1) nous disposerons ainsi du moyen nécessaire pour donner une description complète de l'influence d'une lettre dans un texte.

3. Résultats et perspectives

L'application à la langue roumaine écrite du modèle préconisé s'est heurtée à une difficulté: dans les statistiques existantes, résultées de l'analyse de certains textes plus riches, le dénombrement des digrammes ou des paires de phonèmes n'est pas complet. Ainsi Octavian Tocaciu [8], qui a analysé des textes comprenant 442 730 lettres, a élaboré une matrice des fréquences des digrammes à l'intérieur des mots, qui ne tient pas compte de l'occurrence du blanc et n'est donc pas utilisable de notre point de vue. Alexandra Roceric-Alexandrescu dans son consistant travail de

phonostatistique de la langue roumaine [6] prend en considération parmi les paires de phonèmes seulement les groupes vocaliques et consonantiques, sans investiguer aussi les combinaisons mixtes voyelle-consonne et consonne-voyelle.

C'est pourquoi nous avons été obligés de faire appel à une statistique à dimensions plus modestes, visant certains textes poétiques empruntés à l'oeuvre de M. Eminescu et T. Arghezi, travail dû à E. Nicolau, C. Sala et Al. Roceric [5]. Ces auteurs ont dressé trois tableaux comprenant les probabilités d'apparition des digrammes dans:

1. — les premières poésies de M. Eminescu (celles apparues en 1866);
2. — le poème *Hypérion* du même auteur;
3. — quelques poésies (les titres ne sont pas précisés dans l'étude) de T. Arghezi.

Estimant que la langue des poésies de T. Arghezi, chronologiquement plus rapprochées de nous, reflète mieux la situation actuelle de la langue roumaine, nous avons opté pour l'utilisation de la troisième matrice. Mais en vérifiant le caractère stochastique de cette matrice on a constaté que dans la forme publiée se sont glissées certaines erreurs. Les lignes qui correspondent aux lettres *b*, *î*, *n*, *ş*, *u*, *v* et au *blanc* donnent des sommes différentes par rapport à 1 et elles sont donc incorrectes.

Nous avons préféré, au lieu de modifier arbitrairement certains éléments pour satisfaire la condition imposée, de remplacer entièrement les lignes en cause par les lignes homonymes de la matrice dressée pour le poème *Hypérion*, étant donné que, dans l'absence de quelque chose de meilleur, les probabilités de passage présentes là-bas respectent en quelque mesure les lois de fréquence de la langue. Mais pour la lettre *b* et pour le blanc, la matrice de *Hypérion* s'est avérée elle-même incorrecte et il a fallu recourir à la première statistique, celle des poésies publiées par M. Eminescu en 1866.

Toutes ces modifications ont altéré bien sûr le résultat de l'étude. Pourtant nous montrerons plus loin que si les conclusions qui découlent de l'analyse entreprise ne sont plus totalement valables pour aucun des textes qui ont servi à l'élaboration des matrices de passage considérées, en échange elle ne contredisent pas l'esprit de la langue et concordent en bonne partie avec ce que l'on savait par des investigations antérieures concernant les propriétés statistiques de la langue roumaine.

Le point de départ a été donc une matrice à 26 lignes et tout autant de colonnes, qui correspondent aux 26 lettres rencontrées dans les textes, à savoir: *a*, *ă*, *b*, *c*, *d*, *e*, *f*, *g*, *h*, *i*, *î*, *j*, *l*, *m*, *n*, *o*, *p*, *r*, *s*, *ş*, *t*, *ţ*, *u*, *v*, *z* et le blanc. (On remarque que les textes analysés ne comprennent pas les lettres *k*, *w*, *x* et *y* présentes pourtant, il est vrai, avec des fréquences très réduites, dans l'inventaire de la langue roumaine écrite.)

Pour effectuer les calculs, dont le volume est considérable, on a utilisé un ordinateur, ayant comme vitesse de travail environ 2000 opérations élémentaires par seconde.

L'influence de chaque lettre prise à part a été étudiée selon le schéma exposé. On prenait comme point de départ un vecteur stochastique avec tous les éléments nuls, exception faite pour celui afférent à la lettre choisie. Au moyen des multiplications

successives par la matrice de passage, on déterminait les phases suivantes du processus, en calculant aussi l'entropie qui leur correspond. Pour controler le bon fonctionnement de l'algorithme, on vérifiait pour chaque pas de calcul le caractère stochastique du vecteur résulté.

Le tableau 1 synthétise les résultats des calculs effectués. Dans la première colonne se trouvent inscrites les lettres dont nous nous sommes proposé d'étudier l'influence. Les colonnes 1—10 comprennent les entropies qui caractérisent les positions suivantes.

Par exemple, l'entropie des lettres qui peuvent apparaître en première position après un *j* est de 1,923 bits. En seconde position de 3,816 bits, en troisième de 4,165 bits et ainsi de suite.

Les conclusions plus importantes, du point de vue linguistique, résultées de l'examen du tableau 1, pourraient être formulées comme suit:

1° — L'ergodicité de la chaîne Markov est évidente. On remarque qu'indifféremment de l'état de départ, le régime limite s'établit après environ 10 pas. Autrement dit, une lettre influence les probabilités d'apparition des autres 9 lettres, mais à partir de la dixième lettre cette influence s'éteint.

On pourrait remarquer ici que la frontière à laquelle nous nous rapportons est tout à fait relative. Le régime limite étant une situation vers laquelle on tend asymptotiquement, le moment où nous pouvons considérer qu'il a été atteint dépend de la précision que nous imposons à la coïncidence des entropies. Ainsi du tableau 1, où l'on a pris en considération seulement trois chiffres décimales, il résulterait que pour la lettre *a* on obtient le régime permanent après huit positions. Mais si nous nous intéressons aux premières six décimales, les calculs montrent qu'on rencontre l'entropie limite du texte (4, 120 444 bits) seulement dans la 26ème position. Considérée à ce niveau de précision, l'influence d'une lettre est donc considérablement plus étendue.

Il résulte qu'on ne peut parler du rayon d'action d'une lettre qu'en fonction d'un certain seuil de proximité par rapport à l'entropie du texte, seuil que nous devrons nous imposer. Si nous considérons, par exemple, que l'influence d'une lettre devient négligeable à partir de la position pour laquelle l'entropie diffère avec moins de 1 % en plus ou en moins par rapport à l'entropie de premier ordre du texte (ces limites seraient dans notre cas $4{,}1204 \pm 0{,}0412 = 4{,}0792 + 4{,}1616$) les rayons d'action des lettres seront les suivants:

a	*ă*	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*î*	*j*	*l*
2	3	3	4	4	4	3	3	4	4	2	4	4

m	*n*	*o*	*p*	*r*	*s*	*ş*	*t*	*ţ*	*u*	*v*	*z*	blanc
4	1	3	3	3	3	5	4	4	3	3	4	4

Les différences de comportement informationnel sont clairement marquées: par rapport à la lettre *n* qui n'influence sensiblement que la lettre suivante, le rayon d'action d'un *ş* est 5 fois plus grand.

2° — L'entropie de premier ordre du texte est d'environ 4,12 bits (voir la colonne 10 du tableau 1). Ce résultat a dans notre cas une importance particulière. Il indique que les substitutions de lignes quelque peu arbitraires que nous avons été obligés d'effectuer dans la matrice de passage choisie ne nous ont pas trop éloignés des répartitions de probabilité qui caractérisent la langue roumaine écrite, pour laquelle des recherches antérieures ont montré que l'entropie de premier ordre est d'environ 4,11 bits.

D'ailleurs les probabilités en régime limite diffèrent assez peu des fréquences moyennes établies par des chercheurs différents [4], [5], [8] et en tout cas ne contredisent pas de manière flagrante l'esprit de la langue.

Elle sont les suivantes:

						t ...	4,70 %
a ...	7,17 %	*e* ...	8,44 %	*î* ...	2,11 %	*o* ... 3,17 %	*ţ* ... 0,84 %
ă ...	3,26 %	*f* ...	1,01 %	*j* ...	0,24 %	*p* ... 2,70 %	*u* ... 5,89 %
b ...	0,82 %	*g* ...	0,72 %	*l* ...	3,56 %	*r* ... 5,13 %	*v* ... 1,00 %
c ...	4,50 %	*h* ...	0,19 %	*m* ...	2,80 %	*s* ... 3,09 %	*z* ... 0,52 %
d ...	3,57 %	*i* ...	8,71 %	*n* ...	5,60 %	*ş* ... 1,46 %	blanc 18,81 %

3° — Un résultat plus difficile à prévoir à l'avance, et justement pour cela d'autant plus intéressant, se rapporte à l'allure de la courbe de variation de l'entropie en „aval" d'une lettre donnée. (Pour s'exprimer plus rigoureusement il ne s'agit pas d'une courbe, mais d'un ensemble de points isolés, l'interpolation entre ces points étant dépourvue de sens.)

Nous nous serions attendus qu'à partir de l'endroit où se trouve la lettre connue et où l'indétermination du choix entre les variantes est évidemment nulle, l'entropie augmentât continuellement en se rapprochant indéfiniment de la valeur de l'entropie moyenne du texte. L'intuition nous dit que la connaissance précise d'une lettre dans un texte inconnu devrait faciliter la prévision des lettres suivantes jusqu'à une certaine distance, ce qui se traduit en termes informationnels par une baisse de l'entropie, baisse d'autant plus petite que la lettre cherchée est plus éloignée de celle connue.

Mais les résultats des calculs contredisent cette supposition. Ainsi qu'il ressort du tableau 1, après un petit nombre d'intervalles (1 pour le blanc; 2 pour *a*, *ă*, *e*, *i*, *n*; 3 pour *b*, *c*, *d*, *f*, *g*, *h*, *î*, *j*, *l*, *m*, *p*, *r*, *s*, *ş*, *t*, *ţ*, *v*, *z*; 4 pour *o*) on enregistre des valeurs de l'entropie plus élevées que l'entropie moyenne du texte, après quoi on tend vers celle-ci par des oscillations amortisées.

Ici nous est révélé un paradoxe de la langue qui n'avait pas été mis jusqu'à présent en évidence, à savoir que lorsque l'on demande de deviner une lettre dans un texte au sujet duquel nous ne disposons d'aucune information, la connaissance d'une lettre

rapprochée de celle recherchée ne facilite pas toujours sa découverte. Au contraire même, l'information reçue peut nous dérouter en rendant l'identification encore plus difficile.

Un exemple éclairera mieux cet effet: Si on nous demande de deviner quelle est la lettre qui occupe une certaine position dans un texte que nous ne connaissons pas, la chance de donner une réponse correcte correspond environ à une entropie de 4,12 bits, ce qui équivaut à un choix entre 17,5 variantes également probables. Mais si nous sommes „aidés" en recevant l'information supplémentaire suivante: „la lettre située à deux intervalles à gauche par rapport à la lettre recherchée est un *e*", nos chances de deviner la reponse exacte baissent. L'entropie d'une lettre, placée dans la seconde position après un *e*, est de 4,358 bits (v. le tableau 1) et elle correspond au choix entre un peu plus de 20 variantes également probables.

Tableau 1

	1	2	3	4	5	6	7	8	9	10
a	3,518	4,162	4,027	4,151	4,119	4,114	4,124	4,120	4,120	4,120
ă	2,729	4,355	3,973	4,093	4,152	4,106	4,122	4,123	4,119	4,120
b	3,142	3,873	4,193	4,116	4,106	4,129	4,119	4,120	4,121	4,120
c	3,132	3,724	4,198	4,075	4,122	4,127	4,116	4,121	4,121	4,120
d	2,353	3,548	4,292	4,047	4,107	4,140	4,112	4,121	4,122	4,120
e	3,008	4,358	4,016	4,078	4,154	4,108	4,120	4,123	4,119	4,121
f	3,078	3,677	4,247	4,092	4,101	4,135	4,116	4,120	4,122	4,120
g	3,164	3,872	4,176	4,118	4,106	4,128	4,119	4,120	4,121	4,120
h	2,441	3,509	4,269	4,076	4,099	4,138	4,114	4,120	4,122	4,120
i	3,000	4,285	4,087	4,068	4,148	4,114	4,118	4,123	4,120	4,120
î	1,205	3,420	4,149	4,124	4,103	4,129	4,119	4,120	4,121	4,120
j	1,923	3,816	4,165	4,059	4,138	4,122	4,116	4,123	4,120	4,120
l	3,177	4,089	4,169	4,053	4,140	4,121	4,116	4,123	4,120	4,120
m	3,288	3,836	4,181	4,072	4,126	4,126	4,116	4,122	4,121	4,120
n	3,232	4,149	4,108	4,112	4,128	4,118	4,120	4,121	4,120	4,120
o	3,709	4,070	4,029	4,160	4,114	4,115	4,124	4,119	4,120	4,121
p	3,339	3,772	4,224	4,106	4,102	4,132	4,117	4,120	4,121	4,120
r	3,557	3,747	4,243	4,080	4,109	4,133	4,115	4,121	4,121	4,120
s	3,319	3,957	4,212	4,095	4,114	4,129	4,117	4,121	4,121	4,120
ş	1,564	3,470	4,252	4,111	4,077	4,142	4,117	4,118	4,122	4,120
t	2,880	4,066	4,181	4,049	4,139	4,123	4,115	4,123	4,120	4,120
ţ	1,441	3,162	4,299	4,071	4,079	4,146	4,113	4,119	4,123	4,120
u	3,332	4,094	4,058	4,144	4,114	4,119	4,122	4,120	4,120	4,121
v	2,563	3,638	4,250	4,082	4,102	4,136	4,115	4,120	4,122	4,120
z	2,692	3,531	4,279	4,074	4,098	4,139	4,114	4,120	4,122	4,120
blanc	4,142	3,811	3,935	4,204	4,094	4,115	4,128	4,117	4,121	4,121

La difficulté de la réponse a donc augmenté sensiblement.

En nous précisant la lettre auxiliaire, on nous a fourni en fait une „information négative". La mesure de cette information négative est la différence entre l'entropie moyenne à priori du texte et l'entropie de la position, après avoir reçu l'information supplémentaire (4,120—4,358 = —0,238 bits). Le tableau 2 comprend les informations „positives" et „negatives" que nous fournit la connaissance d'une lettre pour les dix autres lettres suivantes.

En vue d'une illustration spectaculaire du paradoxe ci-dessus, nous pourrons, en partant du tableau 2, combiner des textes possédant des propriétés étranges.

Tableau 2

	1	2	3	4	5	6	7	8	9	10
a	0,602	—0,042	0,093	—0,031	0,001	0,006	—0,004	0	0	0
ă	1,391	—0,235	0,147	0,027	—0,032	0,014	—0,002	—0,003	0,001	0
b	0,978	0,247	—0,073	0,004	0,014	—0,009	0,001	0	—0,001	0
c	0,988	0,396	—0,078	0,045	—0,002	—0,007	0,004	—0,001	—0,001	0
d	1,767	0,572	—0,172	0,073	0,013	—0,020	0,008	—0,001	—0,002	0
e	1,112	—0,238	0,104	0,042	—0,034	0,012	0	—0,003	0,001	—0,001
f	1,042	0,443	—0,127	0,028	0,019	—0,015	0,004	0	—0,002	0
g	0,956	0,248	—0,056	0,002	0,014	—0,008	0,001	0	—0,001	0
h	1,679	0,611	—0,149	0,044	0,021	—0,018	0,006	0	—0,002	0
i	1,120	—0,165	0,033	0,052	—0,028	0,006	0,002	—0,003	0	0
î	2,915	0,700	—0,029	—0,004	0,017	—0,009	0,001	0	—0,001	0
j	2,197	0,204	—0,045	0,061	—0,018	—0,002	0,004	—0,003	0	0
l	0,943	0,031	—0,049	0,067	—0,020	—0,001	0,004	—0,003	0	0
m	0,832	0,284	—0,061	0,048	—0,006	—0,006	0,004	—0,002	—0,001	0
n	0,888	—0,029	0,012	0,008	—0,008	0,002	0	—0,001	0	0
o	0,411	0,050	0,091	—0,040	0,006	0,005	—0,004	0,001	0	—0,001
p	0,781	0,228	—0,104	0,014	0,018	—0,012	0,003	0	—0,001	0
r	0,563	0,373	—0,123	0,040	0,011	—0,013	0,005	—0,001	—0,001	0
s	0,801	0,163	—0,092	0,025	0,006	—0,009	0,003	—0,001	—0,001	0
ş	2,556	0,650	—0,132	0,009	0,043	—0,022	0,003	0,002	—0,002	0
t	1,240	0,054	—0,061	0,071	—0,019	—0,003	0,005	—0,003	0	0
ţ	2,679	0,958	0,179	0,049	0,041	—0,026	0,007	0,001	—0,003	0
u	0,788	0,026	0,062	—0,024	0,006	0,001	—0,002	0	0	—0,001
v	1,557	0,482	—0,130	0,038	0,018	—0,016	0,005	0	—0,002	0
z	1,428	0,589	—0,159	0,046	0,022	—0,019	0,006	0	—0,002	0
blanc	—0,022	0,209	0,185	—0,084	0,026	0,005	—0,008	0,003	—0,001	—0,001

Soit par exemple le texte roumain suivant composé de dix lettres (texte que nous pourrions très bien rencontrer dans un article d'électronique):

— şi — diode — ?

où par — on a noté le blanc.

Quelles sont nos chances pour indiquer correctement la lettre qui suit (la onzième)?

On sait que la lettre cherchée est située immédiatement après un blanc. La connaissance de cette chose nous procure, comme il résulte du tableau 2, une information de —0,022 bits. Elle se trouve aussi à deux intervalles après un *e*, ce qui, ainsi que nous l'avons vu antérieurement, nous donne une nouvelle information de —0,238 bits. En allant plus loin à gauche on remarque qu'absolument toutes les lettres antérieures à celle cherchée ne font que rendre plus difficile l'indentification de celle-ci.

La quantité totale d'information fournie par les lettres précédentes est de —0,534 bits, donc l'entropie de la lettre cherchée sera de 4,120 + 0,534 = 4,654 bits, ce qui correspond à l'option entre 25 variantes également probables.

Voilà combien trompeuse peut être la connaissance partielle d'un texte! (A cet aspect devraient méditer les amateurs de mots croisés, mais aussi pas moins les archéologues et les épigraphistes qui cherchent à réconstituer les fragments indéchiffrables ou perdus des inscriptions anciennes!)

Considérons maintenant un autre texte, cette fois-ci de 8 lettres:

— deşi — îţ ?

où se pose le problème d'identifier la neuvième lettre.

Nous reprenons le raisonnement ci-dessus.

La dernière lettre qui précède celle cherchée est un *ţ* (on a donc une information de 2,679 bits), l'avant-dernière — un *î* (0,700 bits), l'antépénultième le blanc (0,185 bits) et ainsi de suite.

On constate que dans ce cas toutes les lettres connues concourent à faciliter notre tâche et elle nous mettent à la disposition une information totale de 3,685 bits. L'entropie de la lettre cherchée sera 4,120 — 3,685 = 0,435 bits, c'est-à-dire plus petite même que l'entropie de l'option entre deux variantes également probables. Ce fait ne surprend aucun Roumain, qui remplacera facilement le signe d'interrogation par un *i*, la seule lettre admisible dans le contexte donné.

Il serait encore à noter que, bien que d'après notre connaissance jusqu'à présent le paradoxe décrit plus haut n'ait pas été mis en évidence dans une forme explicite, il existait pourtant un indice potentiel pour l'apercevoir. Il y a déjà quelques années on a constaté [6] qu'en roumain l'entropie des phonèmes au commencement des mots est plus grande que l'entropie moyenne de la langue, ce qui revient à dire que la présence du blanc immédiatement à gauche de la lettre cherchée augmente le désordre du système et rend plus difficile l'identification de celle-ci, c'est-à-dire exactement

ce qui est exprimé par la valeur 4,142 de la première cassette de la dernière ligne du tableau 1.

4° — On devrait encore expliquer comment peuvent se mettre d'accord les conclusions de la présente contribution avec les résultats de Burton et Licklieder [2].

Au premier regard il semblerait que les premières contredisent les derniers. Le rayon d'action des lettres dans le modèle présenté ici a résulté quelques fois plus petit que celui trouvé par les chercheurs américains. Mais n'oublions pas qu'on n'a tenu compte ici que de la fréquence des lettres contiguës.

Il est pourtant évident pour le linguiste qu'un rôle important dans la transmission à distance de l'information revient aux restrictions imposées par la langue aux trigrammes, aux tétragrammes et aussi aux séquences plus longues. Si on ne tient compte que de la fréquence des digrammes, étant donné que des combinaisons comme *br* (*brad*) et *rt* (*cort*) sont possibles en roumain, on arrive par exemple à la conclusion erronée qu'avec une certaine probabilité, on peut aussi enregistrer l'occurrence du trigramme *brt*, que la langue rejette pourtant.

On devra donc analyser aussi l'apport des séquences plus longues. Comment procéder?

Pour étudier la contribution des trigrammes, il est suffisant d'inventarier les paires de lettres séparées l'une de l'autre par une troisième lettre. Une telle statistique n'est aucunement plus difficile que celle des digrammes et incomparablement plus simple que la statistique des trigrammes de le langue écrite, dont le nombre est évidement beaucoup plus grand.

Alors, on dresse une matrice de passage dans laquelle la valeur inscrite, par exemple, à l'intersection de la ligne i avec la colonne j représente la probabilité que si dans la position 1 d'un texte nous rencontrons la lettre i, dans la position 3 on retrouve la lettre j.

Le processus Markov, caractérisé par cette matrice, sera un processus à pas doubles par rapport à celui étudié plus haut, parce qu'il parcourra les positions du texte deux à deux. Le rayon d'influence ainsi déterminé, que nous supposons plus grand que dans le cas précédent, exprimera l'apport des trigrammes dans la transmission à distance de l'information fournie par une lettre.

Tenir compte aussi des tétragrammes, pentagrammes etc. que nous ne pourrions pas inventorier, même avec les moyens automatiques les plus modernes, équivaudra encore à établir les fréquences de certains digrammes dont les lettres sont emplacées à de plus grandes distances les unes des autres. Les processus Markov respectifs parcourront des pas triples, quadruples etc. comparativement à ceux du processus initial.

Pour chaque chaîne Markov ainsi conçue on trouvera un autre rayon d'action des lettres que nous supposons de plus en plus grand. Du moment où le rayon demeure constant, on peut affirmer qu'on a déterminé la distance maximum où l'influence

de la lettre considérée s'éteint. De cette manière on trouve pas seulement quel est le rayon global d'action de la lettre mais aussi quel est le rôle des digrammes, trigrammes et des autres séquences de lettres prises séparément dans la propagation de l'information obtenue. Nous avons en préparation une telle description complète accomplie sur un matériel quelque peu restreint, pour le moment. Mais rien n'empêche qu'à l'avenir, par l'examen d'un nombre suffisamment grand de textes, on fasse, fondés sur le modèle présenté ici, une analyse exhaustive d'une ou de plusieures langues du point de vue de l'influence à distance.

En tout cas, jusqu'à une preuve contraire, la conclusion de Burton et Licklieder que le rayon moyen d'action d'une lettre est environ égal à 30 demeure valable.

5° — Il nous reste une dernière remarque. Si l'on part de l'expression mathématique de l'entropie de deuxième ordre et si l'on prend en considération la manière dont ont été déterminées les valeurs inscrites dans la première colonne du tableau 1, on constate que l'entropie de deuxième ordre n'est rien d'autre que la moyenne pondérée de cette colonne, en prenant comme poids les fréquences des lettres situées sur les lignes respectives.

Après avoir effectué les calculs nous avons obtenu pour le texte fictif étudié une valeur de l'entropie de deuxième ordre égale à 3,26 bits. Naturellement on doit avoir des réserves en ce qui concerne l'exactitude de ce chiffre, étant connue la manière dont a été établie la matrice utilisée. Cependant parce qu'à la base de cette étude s'est trouvé l'inventaire des digrammes de quelques textes corrects de la langue roumaine, il est assez probable que l'entropie de second ordre de la langue ne s'éloigne pas trop de cette valeur. L'existence d'une différence assez réduite par rapport au chiffre analogue établi pour l'anglais écrit (3,32 bits) semble aussi confirmer ces suppositions.

Avec les réserves mentionnées ci-dessus nous notons que la valeur déterminée ici est pour le moment la seule dont on dispose pour l'entropie de deuxième ordre de la langue roumaine.

BIBLIOGRAPHIE

[1] BELLMAN, R.: Introduction to matrix analysis. New York—Toronto—London, Mc Graw-Hill Book Company 1960.

[2] BURTON, N. G.—LICKLIEDER, J. C.: Long range constraints in the statistical structure of printed English. In: American Journal of Psychology, *68*, 1955, nr. 4, Baltimore.

[3] FAURE, R.—KAUFMANN, A.—DENIS-PAPIN, M.: Mathématiques nouvelles. Paris, Dunod 1964.

[4] MARCUS, S.—NICOLAU, E.—STATI, S.: Introducere în lingvistică matematică. București, Editura Științifică 1966.

[5] NICOLAU, E.—SALA, C.—ROCERIC, Al.: Observații asupra entropiei limbii române. In: Studii și Cercetări Lingvistice, X, 1959, p. 35—53.

[6] ROCERIC-ALEXANDRESCU, A.: Fonostatistica limbii române. București, Editura Academiei R. S. R. 1968.
[7] SHANNON, C.: Prediction of entropy in printed English. In: Bell System Technical Journal vol. 30, 1951, p. 50—64.
[8] TOCACIU, O.: Unele date statistice privind frecvența literelor și digramelor în limba (scrisă) contemporană. In: Studii și Cercetări Lingvistice, XVI, 1965, p. 683—722.

Generative Grammars and Document Retrieval Languages

OTTO SECHSER, PRAHA

1. The document retrieval language (DR) as understood in this paper is an object consisting of several semantic and three formal components. We are not going to deal with the semantics of DR languages and concentrate upon their formal properties.

The three formal components are:

a) a set of expressions called "document descriptions", i.e. the so-called "description sublanguage",

b) a set of expressions called "retrieval prescriptions", i.e. the so-called "question sublanguage", and

c) a mapping of the description sublanguage into the set of all subsets of the question sublanguage defined by the so-called "formal relevance relation".

2. To illustrate the somewhat abstract concepts we give an example of a DR language of the coordinate indexing type.

2.1. In the DR language concerned there is a list of indexing terms. They are combined to give document descriptions. If $A_a, A_b, \ldots, A_z$ are indexing terms the following expressions are potential (i.e. possible) document descriptions:

$$A_a, A_k, A_g$$
$$A_p$$
$$A_m, A_q$$
etc.

The set of all expressions of this type form the description sublanguage. They are used to express the content of documents.

2.2. The requirements for relevant literature have the form of Boolean retrieval prescriptions expressed by means of indexing terms formally identical with those used in document descriptions. If we wanted to have all documents written on the problem B_g or B_s we could express it

$$B_g \vee B_s$$

where B_g and B_s are indexing terms used in the question sublanguage, formally identical with A_g and A_s respectively and $\vee$ is the AND/OR sign.

Much more complicated Boolean functions may be used as retrieval prescriptions making use of logical sum (AND/OR), logical product (AND), and negation (NOT).

2.3. Moreover we are able to determine the document descriptions that are formally relevant to any particular retrieval prescription. Evidently, some implicit criterion is involved. This criterion is what we call the "formal relevance relation".

In the example cited the retrieval prescription $B_g \vee B_s$ is satisfied by the document description A_a, A_k, A_g (for A_g has the same form as B_g), which is formally relevant to it, whereas it is not satisfied by the remaining two, which are formally non-relevant.

3. There are plenty of types of DR languages. We are dealing with the so-called file-independent[1] binary[2] DR languages.

4. Our concern is the utilization of the theory of generative grammars in the theory of document retrieval languages.

4.1. In the DR languages of the above-mentioned type both description and question sublanguage can be defined by means of suitably chosen generative grammars.

4.2. We have tried to define the sublanguages of several existing document retrieval languages in terms of generative grammars and obtained a very interesting typology of DR languages. The relations of different DR languages are studied on the basis of the relations of the generative grammars required for their definition.

4.3. For file-independent binary DR languages with context-free and ε-free description and question sublanguages (we call them "context-free DR languages") a general method of definition of the formal relevance relation by means of the so-called correspondence rules has been developed. The correspondence rules operate with the non-terminal and terminal symbols of the generative grammars defining the sublanguages. The correspondence rules are the main object of this paper.

4.4. By means of correspondence rules, the algorithms of matching document descriptions against retrieval prescriptions can be specified. This facilitates the elaboration of the necessary computer routines for computer documentation systems.

4.5. Actual work with complicated DR languages requires much computer time. This can be compensated for by the use of a "screening" DR language. It is a simple DR language employed to eliminate the majority of formally non-relevant documents for which the non-relevance can be seen at the first sight. For the "suspicious" documents the formal relevance relation to the particular retrieval prescription is evaluated using the complicated and more expensive, though finer DR language.

[1] In file-independent *DR* languages, in contrast to the file-dependent ones, the value of the formal relevance relation can be determined for a particular document description and a particular retrieval prescription without any additional knowledge of the actual file of documents.

[2] In binary *DR* languages, in contrast to the so-called scoring ones, the formal relevance relation may assume one of two values, viz. "formally relevant" and "formally non-relevant". The scoring languages distingusih more values.

$G_{ds} = (T_{ds}, N_{ds}, R_{ds}, S_{ds})$

$T_{ds} = \{, , A_a, A_b, \ldots, A_z\}$

$N_{ds} = \{DD, GD, D\}$

$R_{ds} =$
1) $DD \to D, GD$
2a) $GD \to D, GD$
2b) $GD \to D$
3a) $D \to A_a$
3b) $D \to A_b$
⋮
3z) $D \to A_z$

$S_{ds} = DD$

Explanation of symbols

G_{ds} — grammar of the description sublanguage
T_{ds} — set of terminal symbols of G_{ds}
N_{ds} — set of non-terminal symbols of G_{ds}
R_{ds} — set of rewriting rules of G_{ds}
S_{ds} — initial symbol of G_{ds}
, — comma combining descriptors in document descriptions
$A_z \ldots A_a$ — finite set of particular descriptors
DD — document description
GD — group of descriptors
D — descriptor

Fig. 1a. Context-free grammar of the description sublanguage of a *DR* language of the coordinate indexing type.

This method can drastically cut the operating costs of a computer documentation system. Here, generative grammars may be useful.

5. We have listed five ways of applying generative grammars to document retrieval languages. The third item seems to be the most important and interesting. We devote to it the rest of our paper.

6. The starting point of our exposition is the assumption that both the description sublanguage and the question sublanguage are defined by suitable generative grammars. In Fig. 1a and 1b we can see the grammars for DR languages of the coordinate indexing type. In Fig. 2a and 2b, examples of derivation trees generated according to the grammars in Fig. 1a and 1b are shown.

6.1. The correspondence rules are a means of describing the algorithm evaluating the formal relevance relation in steps, using the terminal and non-terminal symbols of the grammars.

6.2. The whole evaluation procedure is split up into a sequence of stages. At the beginning of the sequence correspondence rules using the top (initial) symbols and other general non-terminal symbols are located. At the end of the sequence, there are rules concerning the terminal symbols.

6.3. Each correspondence rule could be transformed into a recursive or non-recursive subroutine (sub-algorithm) of the whole routine (algorithm).

6.4. Each correspondence rule has two parts. On the left-hand side there is a relevance or correspondence statement, on the right-hand side — its explication.

6.5. The statement has the form of a two-place predicate. The first place is occupied by a symbol relating to the document description or to its part, the second place by a symbol of the retrieval prescription or its part. The symbols may refer to terminal strings (Relev) or to derivation trees identified by the symbols at their tops (Corresp).

$G_{qs} = (T_{qs}, N_{qs}, R_{qs}, S_{qs})$

$T_{qs} = \{\neg, \wedge, \vee, (,), B_a, B_b, \ldots, B_z\}$

$N_{qs} = \{RP, DG, Dis, CG, C, T, IT\}$

R_{qs}:

1) $RP \rightarrow DG$
2a) $DG \rightarrow Dis \vee DG$
2b) $DG \rightarrow Dis$
3) $Dis \rightarrow CG$
4a) $CG \rightarrow C \wedge CG$
4b) $CG \rightarrow C$
5a) $C \rightarrow T$
5b) $C \rightarrow C$
6a) $T \rightarrow (DG)$
6b) $T \rightarrow IT$
7a) $IT \rightarrow B_a$
7b) $IT \rightarrow B_b$
⋮
7z) $IT \rightarrow B_z$

$S_{qs} = RP$

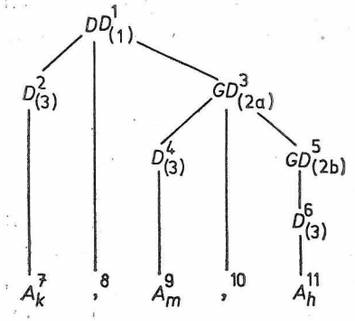
Explanation of symbols

G_{qs} — grammar of the question sublanguage
T_{qs} — set of terminal symbols of G_{qs}
N_{qs} — set of non-terminal symbols of G_{qs}
R_{qs} — set of rewriting rules of G_{qs}
S_{qs} — initial symbols of G_{qs}

$\neg$ — negation symbol
$\wedge$ — AND symbol
$\vee$ — OR symbol
(,) — brackets
$B_a \ldots B_z$ — finite set of particular descriptors

RP — retrieval prescription
DG — disjunctive group
Dis — term of disjunction
CG — conjunctive group
C — term of conjunction (positive or negative)
T — term (positive)
IT — indexing term

Fig. 1b. Context-free grammar of the question sublanguage of a *DR* language of the coordinate indexing type.



The superscripts differentiate the tokens (occurences) of symbols. The figures in brackets refer to the rewriting rules applied at the moment (cf. fig. 1a)

A_k — A_k — cannot be derived

Fig. 2a. Derivation tree of the document description A_k, A_m, A_h.

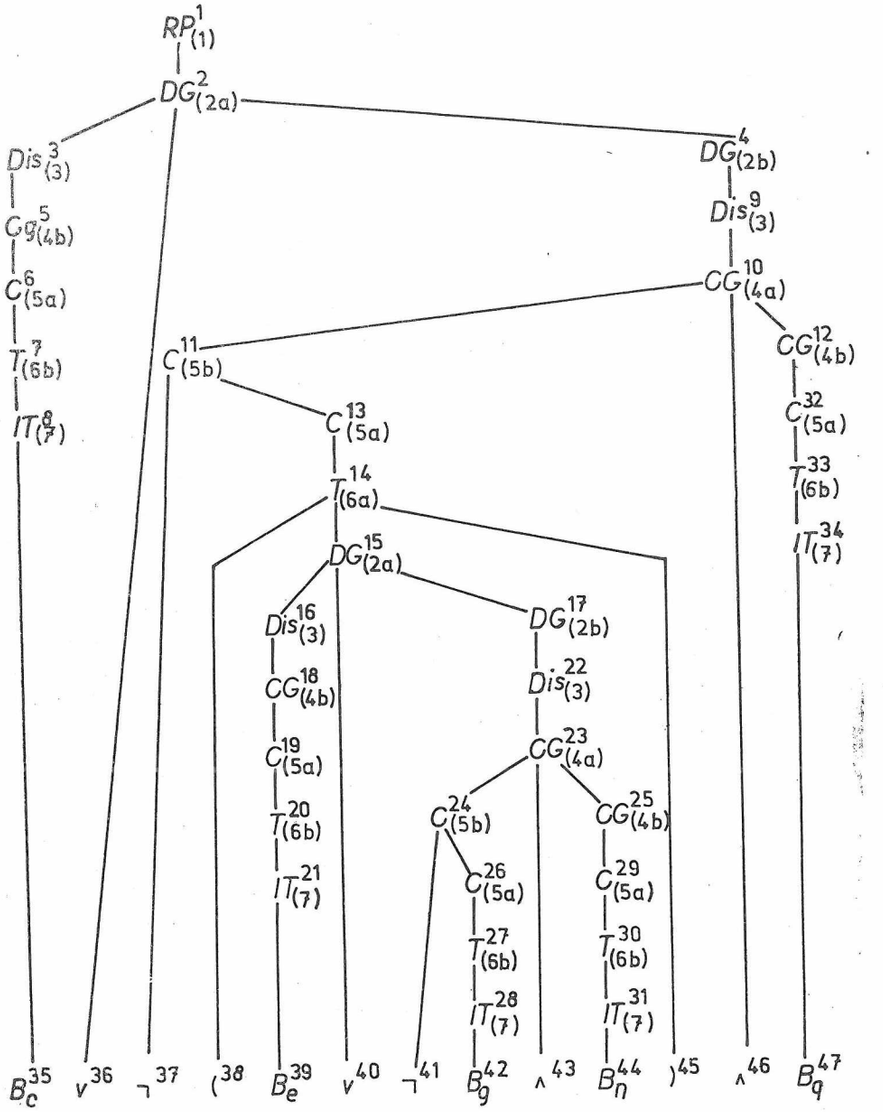


Fig. 2b. Derivation tree of the retrieval prescription $B_c \vee \neg(B_e \vee \neg B_g \wedge B_n) \wedge B_q$.

6.6. If a correspondence rule is employed in practice the statement of relevance or correspondence may be understood as a question about the relation of the indicated objects (terminal strings or derivation trees). In order to reply the question and to decide whether the relation indicated (Relev, Corresp) is fulfilled or not the explication of the statement must be evaluated for the particular document description and the retrieval prescription under investigation. After evaluation of the explication is completed a decision can be made (relevant vs. non-relevant, corresponding vs. not corresponding).

6.7. The explication of a statement of correspondence consists of one or more alternatives, i.e. of a list of all possible ways in which the statement can be fulfilled. The statement is actually fulfilled if at least one alternative is found true, for the given document description and the retrieval prescription.

6.8. In the typical case of a Corresp statement the alternative consists of two parts called *structural condition* and *criterion*. The structural condition specifies the structure of the derivation trees concerned for which the alternative applies. To do this rewriting rules are used to indicate the branching of the trees into subtrees.

6.9. If the structural condition is fulfilled (i.e. if the particular document description or retrieval prescription have the structure described in the structural condition) the criterion may be evaluated. If not, another alternative must be tried.

In evaluating the criterion we may come across two alternative possibilities. Either the criterion can be evaluated without recourse to any further correspondence rule or it can be evaluated only after some other correspondence rule has been applied. In the first case the decision can be made at once; we speak of the so-called *terminal criteria*. In the second case the decision must be postponed. These are the so-called *non-terminal criteria*.

6.10. The operation of the algorithms described in terms of correspondence rules can be generally characterized as follows:

6.10.1. We start with the initial correspondence rule, which involves a Relev statement concerning the terminal strings of the document description and the retrieval prescription concerned. In the explication of this statement the formal relevance is expressed in terms of correspondence of the derivation trees. All criteria are non-terminal so the decision on the relevance or non-relevance has to be postponed, it is reached only after the correspondence rules indicated in the criterion have been applied.

6.10.2. In the next group of correspondence rules the correspondence between the derivation trees (document description and retrieval prescription) is explicated in terms of correspondence of subtrees. The criteria involved are also non-terminal. The derivation trees are thus analysed from left to right and from the top to the bottom. Ultimately the analysis arrives at the bottom of the derivation trees, i.e. to the terminal symbols. Then it is possible to carry out various tests concerning the

terminal symbols of the document description and the retrieval prescription being matched against one another. This is performed according to correspondence rules with terminal criteria.

6.10.3. When the results of application of the terminal criteria to the particular document description and the retrieval prescription are known the analysis may return back, upward, to the preceding correspondence rules; the evaluation of their criteria can be completed and the decision (originally postponed) finally made.

6.10.4. In this way the analysis arrives at the first (initial) correspondence rule and the definitive decision on the formal relevance of the particular terminal strings (document description vs. retrieval prescription) is reached.

7. In Fig. 3 we see the set of correspondence rules for the DR language of the coordinate indexing type the description and question sublanguages of which are defined in Fig. 1a and 1b.

1. Relev (*dd*, *rp*)	↔	Corresp (*DD*, *RP*)	1 → 2
2. Corresp (*DD*, *RP*)	↔	DD^0; $RP^0 \rightarrow DG^1$ Corresp (DD^0, DG^1)	2 → 3
3. Corresp (*DD*, *DG*)	↔ a)	DD^0; $DG^0 \rightarrow Dis\ \vee^2\ DG^3$ Corresp (DD^0, Dis^1) OR Corresp (DD^0, DG^3)	3 → 4 OR 3
	b)	DD^0; $DG^0 \rightarrow Dis^1$ Corresp (DD^0, Dis^1)	3 → 4
4. Corresp (*DD*, *Dis*)	↔	DD^0; $Dis^0 \rightarrow CG^1$ Corresp (DD^0, CG^1)	4 → 5
5. Corresp (*DD*, *CG*)	↔ a)	DD^0; $CG^0 \rightarrow C^1\ \wedge^2\ CG^3$ Corresp (DD^0, C^1) AND Corresp (DD^0, CG^3)	5 → 6 AND 5
	b)	DD^0; $CG^0 \rightarrow C^1$ Corresp (DD^0, C^0)	5 → 6
6. Corresp (*DD*, *C*)	↔ a)	DD^0; $C^0 \rightarrow T^1$ Corresp (DD^0, T^1)	6 → 7
	b)	DD^0; $C^0 \rightarrow \neg^1\ C^2$ NOT Corresp (DD^0, C^2)	6 → 6
7. Corresp (*DD*, *T*)	↔ a)	DD^0; $T^0 \rightarrow (^1\ DG^2)^3$ Corresp (DD^0, DG^2)	7 → 3
	b)	DD^0; $T^0 \rightarrow IT^1$ Corresp (DD^0, IT^1)	7 → 8
8. Corresp (*DD*, *IT*)	↔	$DD^0 \rightarrow D^1,\ ^2\ GD^3$; IT^0 Corresp (D^1, IT^0) OR Corresp (GD^3, IT^0)	8 → 10 OR 9
9. Corresp (*GD*, *IT*)	↔ a)	$GD^0 \rightarrow D^1,\ ^2\ GD^3$; IT^0 Corresp (D^1, IT^0) OR Corresp (GD^3, IT^0)	9 → 10 OR 9
	b)	$GD \rightarrow D^1$; IT^0 Corresp (D^1, IT^0)	9 → 10
10. Corresp (*D*, *IT*)	↔ i)	$D^0 \rightarrow A_i^1$; $IT^0 \rightarrow B_i^1$ $A_i^1 = B_i^1$	10 → DECISION

Fig. 3. Correspondence rules of a *DR* language of the coordinate indexing type (for the grammars of the description and question sublanguages see Fig. 1).

8. The first correspondence rule is of the Relev type. It can be read as follows: "A particular document description *dd* is formally relevant to a particular retrieval prescription *rp* (the relevance statement) if and only if the derivation tree *DD* corresponds to the derivation tree *RP*." (*DD* and *RP* are the initial symbols at the very top of the derivation trees). It is understood that *DD* is the derivation tree of the particular *dd* and *RP* that of the particular *rp*. Evidently the decision on the formal relevance can be made only after the derivation trees are analysed.

9. The second correspondence rule is of the Corresp type. It can be read as follows: "A particular derivation tree *DD* corresponds to a particular derivation tree *RP* if and only if *GD* is the subtree of *RP* (this requirement corresponds to the first rule in Fig. 1b) and the *GD* corresponds to the *DD*."

Before the criterion of the second correspondence rule is evaluated the statement Corresp (*DD*, *GD*) must be analysed. The decision on the correspondence is consequently postponed and the analysis continues with the third correspondence rule.

10. The third correspondence rule is of the Corresp type. The explication of the statement of correspondence consists of two alternatives parallel to the two rewriting rules for the *DG* symbol (see Fig. 1b, rules 2a and 2b). It can be read as follows: "A particular derivation tree *DD* corresponds to a particular derivation tree *DG* if and only if

either (alternative a) the *DT* (referred to more exactly as DG^0) consists of three branches Dis^1, $\vee^2$ and DG^3 (structural condition) and if the DD^0 tree corresponds to the Dis^1 tree and/or the DD^0 tree corresponds to the DG^3 tree (criterion)

or (alternative b) the DG^0 consists of one branch Dis^1 only (structural condition) and if the DD^0 tree corresponds to it (i.e. the Dis^1 tree) (criterion)."

10.1. Several remarks are necessary.

a) Superscripts are used to prevent confusion if one symbol occurs more than once in a rewriting rule. By means of this we can easily differentiate the DG^0 tree from the DG^3 tree (which is a subtree of DG^0).

b) Both criteria are of the non-terminal type. It means the decision on the correspondence statement must be postponed.

c) The criterion of the alternative a) refers to two correspondence rules, i.e. No.4 and No. 3 (see the generative metagrammar in Fig. 3 on the very right). If the first statement [Corresp (DD^0, Dis^1)] is true the second statement [Corresp (DD^0, DG^3)] need not be evaluated since they are linked by the logical sum (OR) symbol. Thus, the second half of the criterion is evaluated only if the first one is found false.

11. In this way the analysis goes on down the derivation trees, with the decisions remaining postponed, until the terminal symbol level (i.e. the bottom level of the derivation tree) is reached.

12. At the terminal symbol level the only correspondence rule with the terminal criterion is used, viz. correspondence rule No. 10. The text specified in the criterion consists in determining if both indexing terms (A_i and B_j) have an identical form.

If so a positive decision on Corresp (*D*, *IT*) is made, if not the decision is negative.

13. At this moment the analysis may return back to the immediately higher level where the postponed decision can be reached. In a similar way the procedure goes upward and downward until the whole derivation trees are checked. By that time the procedure returns to the initial correspondence rule and the definitive decision on the formal relevance or non-relevance is made.

14. The set of correspondence rules in Fig. 3 belongs to a very simple *DR* language (*DR* language of the coordinate indexing type). The merit of the standard form of the

AAAAAA (BB, CC) —
a) $BB^0 \rightarrow$ DDDDDDDDDDDD; $CC^0 \rightarrow$ EEEEEEEE EEEEE FFFFFFFFFFFFFFFFFFFF
b) $BB^0 \rightarrow$ GGGGGGGGGGGG; $CC^0 \rightarrow$ HHHHHHHHHHHHH IIIIIIIIIIIIIIIIII
c) $BB^0 \rightarrow$ JJJJJJJJJJJJ; $CC^0 \rightarrow$ KKKKKKKKKKKKK LLLLLLLLLLLLLLLLLLLL

Explanation

AAAAAA (BB, CC) — left-hand side of the correspondence rule; relevance or correspondence statement

AAAAAA (..,..) — indication of the relation which is the object of the relevance or correspondence statement; "Relev (..,..)" relates to the relation between two terminal strings, "Corresp (..,..)" to that of derivation trees named with their top symbols

BB — symbol relating to the document description; with "Corresp" it is the name of the derivation tree of the document description or its part

CC — symbol relating to the retrieval prescription; with "Corresp" it is the name of the derivation tree of the retrieval prescription or its part

a), b), c) — indication of the alternatives forming the right-hand side of the correspondence rule or the explication of the statement at the left-hand side

BB^0 ...EEEEE — structural condition stating the assumed branching of the derivation tree in the document description derivation tree ("$BB^0 \rightarrow \rightarrow$ DDDDDDDD") and in the retrieval prescription derivation tree ("$CC^0 \rightarrow$ EEEEEEEE"); if no branching is assumed no right-hand side of the rewriting rules is given so that CC^0 or "BB^0" results; the superscripts are used to differentiate individual trees and subtrees

FFFFFFFFFF — criterion; it operates with the symbols in the structural condition; often it includes correspondence statements; then its function is to "call' the lower-level correspondence statements; this gives the correspondence rules the property of recursiveness; it can also include the description of various tests; in such a case no lower-level correspondence statement is called but the test performed

Fig. 4. Schematic representation of the structure of correspondence rules.

correspondence rules (see Fig. 4) lies in the fact that correspondence rules of this type can be employed to specify the formal relevance relation of a wide class of complicated *DR* languages. Some information on the DR languages with brackets has already been published (see the references).

REFERENCES

[1] SECHSER, O.: Korespondenční pravidla. Prague, ÚVTEI 1968. Internal report.
[2] SECHSER, O.: Poznámka k selekčním jazykům se syntaxí. MTI, *10*, 1968, pp. 1—36.

Моделирование языка в связи с задачами информационного поиска

ЭДУАРД Ф. СКОРОХОДЬКО, Киев

В данном сообщении излагается один метод моделирования системы семантических связей в лексике естественного языка. Модель понимается здесь как некоторая аппроксимация действительности (ср. [1], [2]). Рассматриваемая модель предназначена служить основой для построения информационного языка. Из этого общего назначения модели вытекают конкретные требования к ней.

1. Семантические связи между словами должны быть представлены в явном виде, эксплицитно.

Это требование вытекает из стремления к формализации ряда процедур, включающих анализ парадигматических связей между словами (имеется в виду семантическая парадигматика, т. е. связи между значениями слов на уровне языка). Так, при информационном поиске возникает необходимость в установлении наличия и вида парадигматических связей между терминами, входящими в состав информационного запроса, и терминами, входящими в состав документа. Учет парадигматических связей необходим также при автоматическом семантическом анализе текстов, в частности, при установлении синтагматических отношений между словами.

2. Должна быть обеспечена возможность определения количественных параметров, характеризующих систему семантических связей в лексике.

Это требование вытекает из стремления получить адекватные измерители тех свойств лексики, которые определяют выбор типа информационного языка и других компонентов информационно-поисковых систем. Наличие таких измерителей позволит сопоставлять на объективной основе лексику различных сфер человеческой деятельности. Роль таких измерителей особенно возрастает в связи с активизацией процессов целенаправленного воздействия на лексику (прежде всего, упорядочивания терминологии). Благодаря им оказывается возможным контролировать результат этого воздействия.

3. Должна быть обеспечена возможность определения количественных параметров, характеризующих некоторые семантические свойства отдельного слова в лексической системе языка.

Это требование вытекает из стремления получить оценку роли конкретной лексической единицы в лексической системе языка. Определенный таким образом ,,вес" слова или устойчивого словосочетания, наряду с его частотными характеристиками, должен учитываться при составлении словарей, как машинных, предназначенных для автоматической обработки текстов, так и ,,обычных" — толковых, терминологических, переводных.

4. Должна быть обеспечена возможность определения наличия и характера семантической связи между любыми двумя единицами лексической системы языка.

Это требование объясняется тем, что только при его выполнении вся лексика языка (или нужный ее фрагмент) будет представлена как единое целое. В этом случае появится возможность исследовать взаимодействие любого элемента лексики с любым другим элементом или с лексической системой в целом.

В основе модели лежат два понятия: ,,семема" и ,,семантическое отношение". Под семемой понимается семантическая единица языка или речи, соответствующая предмету, а под семантическим отношением — семантическая единица, соответствующая отношению между предметами. Понятия ,,семантическая единица", ,,предмет" и ,,отношение между предметами" считаются исходными, не требующими определения (или, точнее, не требующими определения в рамках данной модели).

План содержания лексики языка можно понимать как совокупность семем и семантических отношений. Естественно считать, что если два предмета объективной действительности соединены некоторым отношением, то соответствующие им семемы в плане содержания языка также соединены определенным семантическим отношением. В результате каждая семема оказывается связанной как с единицей предметной области, так и с другими семемами, т. е. единицами плана содержания языка. Если две семемы соединены семантическим отношением, то слова, значениями которых являются эти семемы, мы считаем семантически связанными.

Подход к плану содержания лексики языка как к совокупности семем, соединенных семантическими отношениями, подсказывает наиболее удобный для наших целей способ представления плана содержания лексики — в форме графа (этот способ применяется и другими исследователями, см., например, [3]).

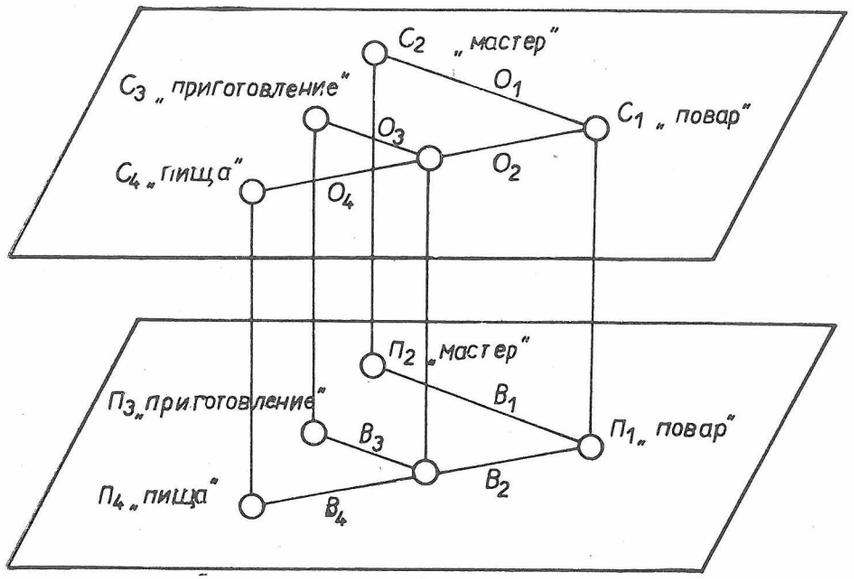
Построим граф, вершины которого соответствуют семемам, а ребра — семантическим отношениям. Назовем его языковой семантической сетью. Для установления семантических отношений можно прибегнуть к различной методике. Так, наличие семантических отношений может быть установлено путем анализа лексической сочетаемости слов. Если два слова часто встречаются с одним и тем же списком слов, то между ними предполагается некоторая семантическая связь [3]. При этом выявляется лишь наличие семантической связи и ее интенсивность, а не ее характер. Это позволяет соединить соответствующие вершины языковой семантической сети ребрами. Языковая семантическая сеть в этом случае представляет собой неориентированный граф с неразмеченными (не имеющими обозначений) ребрами. Если задано описание значений слов в терминах семантических составляющих, то критерием семантической связи между словами является их производность от одной и той же семантической составляющей. Два слова считаются семантически связанными, если одно из них может выступать в качестве семантической составляющей другого (на языковой семантической сети в этом случае от первого слова ко второму направлена стрелка) или если они оба имеют одну и ту же семантическую составляющую (в этом случае к ним ведут стрелки, исходящие из некоторого третьего слова). Такая же методика может применяться и для построения языковой семантической сети на основе описания лексических значений в толковом словаре. Слова, привлеченные для толкования данного слова, выступают как семантические составляющие последнего. Поэтому если одно слово (или его синоним) используется в толковании другого слова, то на языковой семантической сети соответствующие вершины соединяются ребром, направленным от первого слова ко второму. Следует отметить, что в общем случае языковые семантические сети одного и того же фрагмента лексики, полученные на основании разных методик, не совпадают, хотя между ними имеется определенная корреляция.

На рис. 1 условно изображен фрагмент предметной области и соответствующий ему фрагмент плана содержания — языковая семантическая сеть. На этом рисунке символами $П$ обозначены предметы, $С$ — соответствующие им семемы, $В$ — отношения между предметами, $О$ — семантические отношения. Из рисунка видно, что языковая семантическая сеть повторяет предметную. Изоморфизм между ними объясняется тем, что мы учли лишь те связи между предметами, которые нашли отражение в языке.

Представление системы семантических связей в лексике в форме языковой семантической сети удовлетворяет перечисленным выше требованиям.

1. Семантические связи между словами (и устойчивыми словосочетаниями) представлены эксплицитно, в форме ребер или цепочек ребер, независимо от того, как они выражены в словаре языка. Так, семантическая связь между словами *повар* и *мастер* представлена в виде ребра 0_1, а связь между словами *повар* и *пища* — в виде цепочки, состоящей из ребер 0_2 и 0_4. При этом наличие ребра между вершинами свидетельствует о наличии семантической связи между соответствующими лексическими единицами, а его обозначение (0_1, 0_2 и т. п.) — о ее характере.

2. Изображение плана содержания лексики в форме графа позволяет



количественно оценивать этот граф и, тем самым, оценивать план содержания лексики языка. Предлагаются следующие параметры, описывающие систему семантических связей в лексике: степень семантической связности в сильном и слабом смысле, степень семантической компактности, степень семантической плотности и др.

Степень семантической связности в сильном смысле определяется по формуле:

$$S = \frac{2kn}{t(t-1)}, \tag{1}$$

где S — степень семантической связности в сильном смысле,
k — коэффициент связности,
n — число пар семантически связанных слов,
t — общее число слов в рассматриваемом фрагменте лексики.

Коэффициент k принимает значение $k = 1$, если языковая семантическая сеть представляет собой связный граф, и значение $k = 0$, если она представляет собой несвязный граф.

Обширные фрагменты лексики нередко объективно распадаются на ряд относительно независимых лексических систем. При этом отдельные слова, принадлежащие различным системам, иногда не имеют между собой семантической связи. При оценке подобных фрагментов по формуле (1) коэффициент k оказывается равным нулю, и, следовательно, степень семантической связности их будет равна нулю, как велика ни была бы степень связности в пределах отдельных групп слов рассматриваемого фрагмента.

Тем не менее, в подобных случаях может представить интерес и оценка суммарной связности лексики без учета разрывов связности между полярными группами слов. Такая связность в слабом смысле определяется по формуле, которая получается из формулы (1) устранением коэффициента k:

$$S' = \frac{2n}{t(t-1)}. \tag{2}$$

Если учитывать не только сам факт наличия или отсутствия семантической связи между словами, но и их силу, то можно ввести другой семантический параметр лексики — семантическую компактность плана содержания. Семантическая связность характеризует удельный вес числа семантически связанных пар слов во фрагменте лексики. Семантическая же компактность характеризует в обобщенном виде силу семантической взаимосвязи этих пар слов.

Сила семантической связи между парой слов может быть определена по формуле:

$$F = \frac{1}{s+1}, \tag{3}$$

где F — сила семантической связи между двумя словами T_i и T_j,
s — число семем, образующих минимальную по длине цепочку, связывающую в языковой семантической сети сопоставляемые слова T_i и T_j с их общей семантической составляющей.

Силу семантической связи можно определять и по формулам, предложенным для вычисления так называемого „семантического расстояния".

Степень семантической компактности плана содержания можно вычислить по формуле:

$$K = \frac{2F_{\min}\sum_{i=1}^{n} F_i}{t(t-1)}, \tag{4}$$

где K — степень семантической компактности плана содержания лексики,
F — сила семантической связи между словами,
$F_{\min}$ — сила семантической связи между парой наиболее слабо связанных слов рассматриваемого фрагмента плана содержания,
n — число пар семантически связанных слов.

Как само понятие семантической компактности является в известной мере обобщением понятия семантической связности, так и формула (4) является обобщением формулы (1). Если предположить, что сила семанти-

ческой связи может принимать лишь два значения: $F = 1$ и $F = 0$ (связь либо есть, либо она отсутствует), то $F_{\min} = k$, $\sum_{i=1}^{n} F_i = n$ и формула (4) приводится к формуле (1).

В том случае, когда языковая семантическая сеть представляет собой неориентированный граф (см. выше), вместо семантической связности определяется семантическая плотность:

$$P = \frac{2p}{t(t-1)}, \tag{5}$$

где P — степень семантической плотности,

p — число пар слов, непосредственно связанных семантически (на языковой семантической сети такие слова соединены одним ребром),

t — общее число слов во фрагменте лексики.

Предложенные выше параметры, не давая всесторонней характеристики системности плана содержания лексики, тем не менее, дают определенное представление о ее особенностях, позволяют осуществлять типологическое сопоставление различных фрагментов общеупотребительной и терминологической лексики.

3. Представление плана содержания лексики в форме графа позволяет оценивать не только языковую семантическую сеть в целом, но и любые ее элементы. Тем самым обеспечивается возможность получить количественные параметры, характеризующие отдельную лексическую единицу как элемент лексической системы языка.

Основным параметром, характеризующим отдельное слово, является степень его значимости в системе языка. Она определяется по формуле:

$$L = R(Q - Q_{\max}), \tag{6}$$

где L — степень значимости слова в плане содержания лексики,

Q — число вершин языковой семантической сети данного фрагмента лексики,

$Q_{\max}$ — число вершин в наибольшем по величине графе, который образуется из исходной сети в результате удаления из нее вершины, соответствующей слову, степень значимости которого оценивается,

R — число ребер, инцидентных вершине, соответствующей оцениваемому слову.

Величина R характеризует локальную, а $Q - Q_{\max}$ — общелексическую значимость слова (ср. с локальной и общетекстовой значимостью предложения в тексте, о чем идет речь ниже).

Аналогичный метод разработан и для представления системы семантических связей в тексте. Граф, изображающий систему семантических связей между элементами текста (словами, предложениями, абзацами, параграфами), получил название речевой семантической сети. На основании речевой семантической сети можно вычислять ряд количественных параметров, характеризующих план содержания текста в целом и его элементы.

Коэффициент семантической связности текста характеризует семантическую монолитность текста. Он оценивается по формуле, аналогичной формуле (5) или (2):

$$\sum = \frac{2l}{y(y-1)}, \tag{7}$$

где $\sum$ — коэффициент семантической связности текста или фрагмента текста,

l — число пар предложений, семантически связанных друг с другом,

y — общая длина текста (число предложений, образующих текст).

Критерием наличия семантической связи между предложениями считается совпадение хотя бы одной семантической составляющей слов, входящих в эти предложения. Таким образом, обычный критерий семантической связи — совпадение лексики мы рассматриваем лишь как частный случай.

Для устранения паразитных связей вводится понятие релевантности слова. Слово считается релевантным относительно данного текста, если между ним и хотя бы одним другим релевантным словом данного текста имеется семантическая связь на уровне языковых значений.

Вторым параметром, характеризующим план содержания текста, является семантическая компактность текста. Коэффициент семантической компактности текста определяется по формуле, аналогичной формуле (4):

$$\Theta = \frac{2\omega_{\min} \sum_{i=1}^{n} \omega_i}{y(y-1)}, \tag{8}$$

где Θ — коэффициент семантической компактности текста или его фрагмента,

ω — сила семантической связи между предложениями текста, определяемая в простейшем случае по числу слов, входящих одновременно в оба предложения,

$\omega_{\min}$ — сила семантической связи между парой наиболее слабо связанных предложений рассматриваемого текста,

n — число пар семантически связанных предложений в тексте.

Основным параметром, характеризующим значимость предложения (или другого элемента текста — слова, абзаца и т. д.), является функциональный вес слова. Эта величина прямо пропорциональна числу ребер речевой семантической сети, инцидентных вершине, которая соответствует рассматриваемому предложению или слову. Иначе говоря, она определяется, прежде всего, числом семантических связей, в которые вступает в данном тексте предложение или слово. Это характеризует локальную значимость. Кроме того, функциональный вес предложения или слова зависит от изменения речевой семантической сети при удалении этого предложения или слова, что характеризует общетекстовую значимость. Очевидно, что предложение или слово имеет большой функциональный вес, если после его удаления речевая семантическая сеть превращается в несвязный граф. При этом функциональный вес тем выше, чем больше разность между числом вершин исходной речевой семантической сети и максимального по величине графа, образовавшегося из нее после удаления соответствующей вершины. Это означает, что если после удаления предложения или слова текст распадается на фрагменты, не связанные между собой, то функциональный вес такого предложения или слова тем выше, чем меньшая часть текста остается без изменения. Таким образом, функциональный вес тем выше, чем больше влияет данная единица на понимание остального текста.

Функциональный вес предложения или другой единицы в тексте оценивается по формуле:

$$\varphi = N(M - M_{\max}), \tag{10}$$

где φ — функциональный вес предложения или слова в данном тексте или данном фрагменте текста,
N — число дуг, инцидентных рассматриваемой вершине,
M — число вершин речевой семантической сети исходного текста,
$M_{\max}$ — число вершин в наибольшем по величине графе, который образовался в результате удаления из исходной речевой сети вершины, соответствующей оцениваемому предложению, слову или абзацу.

Специфическим для плана содержания текста является параметр, характеризующий среднюю степень семантической связи предложения с предшествующим ему отрезком текста, который состоит из k предложений. Этот параметр получил название линейного коэффициента. Он определяется по формуле:

$$M_{ik} = \frac{j}{k}, \tag{11}$$

где M_{ik} — линейный коэффициент i-го предложения,

k — длина предшествующего отрезка текста (измеряемая числом предложений),
j — число предложений в предшествующем отрезке текста длиной k, семантически связанных с i-м предложением.

Параметры плана содержания текста (как перечисленные, так и некоторые другие) позволяют, в частности, прогнозировать функциональный стиль текста. Коэффициент семантической связности текста определяется функциональным стилем и практически не зависит от авторской манеры письма, т. е. определяется прежде всего адресатом сообщения. Для специальных научных текстов коэффициент семантической связности составляет 0,6—0,8; для научно-популярных 0,4—0,55; для художественных и публицистических 0,15—0,3.

Определение количественных параметров речевой семантической сети позволяет также классифицировать семантическую структуру текстов.

ЛИТЕРАТУРА

[1] АНДРЕЕВ, Н. Д.—ЗИНДЕР, Л. Р.: Основные проблемы прикладной лингвистики. Вопросы языкознания, 1959, № 4.
[2] РЕВЗИН, И. И.: Модели языка. Москва, Издательство АН СССР 1962.
[3] МОСКОВИЧ, В. А.: Статистика и семантика. Москва, Наука 1969.

Problèmes de typologie verbale à l'aide de la théorie des graphes

VICTORIA HOPÂRTEANU, ILEANA LASCU,
DAN MÂRZA, MARIA ŢENCHEA, CLUJ

Une description typologique a toujours à sa base la comparaison. Pour pouvoir comparer des langues assez différentes entre elles, il faut disposer d'un instrument de travail adéquat, d'une méthode d'étude plus générale. La théorie des graphes peut fournir une telle méthode: elle peut être employée en égale mesure dans la description de la typologie nominale ou verbale, tant pour les langues flexionnelles que pour les langues agglutinantes.

En Roumanie cette méthode a été utilisée en phonétique par Solomon Marcus et E. Vasiliu [19], en morphologie par Emese Kis et Felicia Oşianu [12] et en syntaxe par E. Kis en collaboration avec E. Comşulea et I. Anghel [10, 11]. La théorie des graphes a servi de base à J. Horecký [7] qui a étudié la suffixation nominale et verbale dans la langue slovaque, et à Y. Gentilhomme pour décrire la langue russe.

Nous avons essayé d'appliquer la théorie des graphes à l'étude de la morphologie verbale. A l'aide de cette théorie, nous avons étudié les verbes auxiliaires du roumain [3, 5, 6] (*être* et *avoir*), en parallèle avec ceux des langues suivantes: latin [20], italien [13], français [4], espagnol [9], allemand [22], anglais [1], albanais [2], hongrois [23] et russe [16]. Nous avons comparé seulement des faits appartenant à la langue écrite. Le critère sémantique étant exclus, c'est l'aspect formel-quantitatif qui nous a guidé dans la délimitation des éléments constitutifs des formes verbales.

I. Définitions et symboles

On appelle *flectif invariable* tout élément, spécifique d'un mode ou d'un temps, qui se trouve entre deux blancs (f). Sont ainsi des flectifs invariables: la conjonction *să* du roumain (diagramme 4), *che* de l'italien (diagramme 5), *que* du français, *të* de l'albanais.

Le *pronom personnel* est transcrit par un p, suivi de l'indice de personne (de 1 à 6) et, si c'est le cas, de l'indice du genre (a = masculin, b = féminin, c = neutre). Par exemple, en français *il* = p_{3a}, *elle* = p_{3b}; en anglais *he* = p_{3a}, *she* = p_{3b}, *it* = p_{3c}. On a indiqué le genre seulement dans les cas où le verbe change de forme.

Un t note le *thème* du verbe, lorsqu'il est identique pour toutes les personnes. Par

exemple, pour l'imparfait du verbe *avoir* *t* signifie: *habe* (latin) (diagramme 11), *av-* (roumain, italien, français) (diagrammes 7, 8, 10), *hab-* (espagnol) (diagramme 9), *hatt-* (allemand) (diagramme 12), *i-* (albanais) (diagramme 14). S'il y a plusieurs thèmes, on emploie les indices (de 1 à 6) de la personne ou des personnes correspondant à chaque thème. En roumain, le présent de l'indicatif se forme à partir de trois thèmes différents: $t_{1,2,6} = a$-, $t_3 = are$, $t_{4,5} = av$- (diagramme 3).

Le *suffixe morphologique*, que certains linguistes préfèrent appeler infixe ou interfixe (15), est défini comme étant la particule invariable qui apparaît dans le corps du verbe à conjuguer, entre le thème et la désinence, et qui caractérise un temps ou un mode. Il est noté par un *s* (s'il est commun à toutes les personnes). En latin, il y a des enchaînements de suffixes, représentés par *s'*, *s''* etc.

Le *préfixe morphologique* (symbole *r*) n'apparaît que dans certaines langues. *Ge-*, formant du participe en allemand, en est un exemple.

La *désinence*, particule terminale attachée au thème et qui indique l'information concernant la personne, le nombre et le genre, est représentée par un *d* suivi de l'indice de personne. Par exemple, au passé simple $d_1 = i$ (latin, roumain, italien, espagnol), *-s* (français), *-a* (albanais); d_6 = *-ră* (roumain), *-rono* (italien), *-rent* (français).

Les verbes auxiliaires qui apparaissent dans les formes verbales composées, appelés *flectifs mobiles variables*, seront représentés par leurs constituants (thème, désinence, suffixe). Les symboles respectifs sont surmontés d'un ou de deux tirets (si l'auxiliaire apparaît deux fois dans le corps du verbe): $\bar{l}$, $\bar{t}$, $\bar{d}$, $\bar{d}$, $\bar{s}$, $\bar{s}$.

Tous les éléments représentés ci-dessus ($p, f, r, \bar{l}, \bar{s}, \bar{d}, \bar{t}, \bar{s}, \bar{d}, t, s, d$), indifféremment de leur nature et de leur fonction, constituent des *noeuds*, points d'incidence de segments orientés appelés *arcs*.

L'ensemble des noeuds et des arcs correspondant à un mode ou temps constitue un *graphe*.

Les noeuds sont disposés horizontalement en *lignes* et verticalement en *colonnes*. Les lignes correspondent aux différentes personnes (pour les modes personnels) et, en général, aux différentes variantes paradigmatiques. Dans le graphe qui représente l'indicatifs futur immédiat du verbe français *avoir* il y a six lignes (diagramme 2); l'indicatif passé du verbe anglais *to have* a cinq lignes (diagramme 13). Le participe passé du verbe français *avoir* n'a qu'une seule ligne. Les colonnes correspondent aux différents éléments constitutifs de la forme verbale (formants). Le future analytique hongrois a cinq colonnes (diagramme 15), le conditionnel passé en anglais (diagramme 16) et le futur antérieur en albanais sept.

Ce qui nous intéresse ici, ce n'est pas la nature des noeuds, mais leur disposition, la façon dont ils sont structurés. En regardant les graphes des verbes étudiés, on constate qu'il y a des noeuds qui constituent des points d'aboutissement qui n'ont que valences négatives [7] pour plusieurs arcs, ainsi que des noeuds qui forment l'origine de plusieurs arcs et qui ont des valences positives. Les premiers seront appelés *noeuds de convergence*, et les autres *noeuds de divergence*. Il y a aussi des noeuds mixtes,

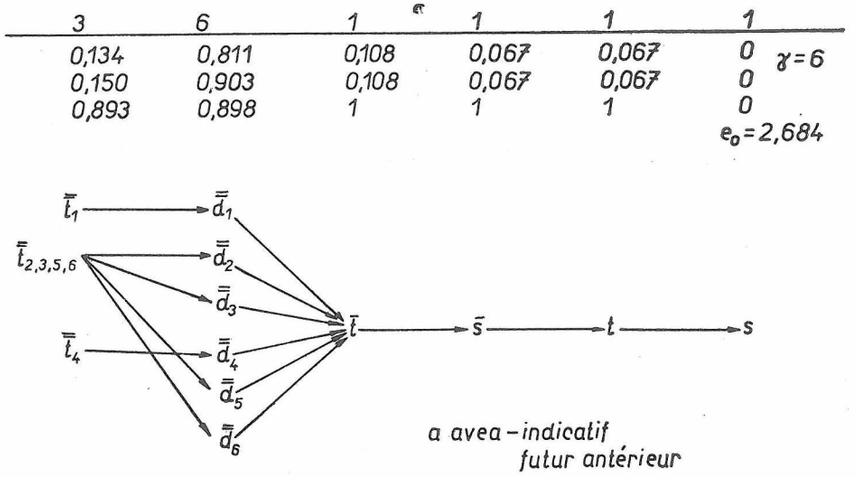


Diagramme 1.

de convergence-divergence. Exemples: noeuds de divergence: $\bar{l}_{2,3,5,6}$ du futur antérieur roumain (diagramme 1), *f* du subjonctif plus-que-parfait en italien (diagramme 17), noeuds de convergence: *t* du subjonctif plus—que—parfait en italien (diagramme 17), *t* du futur immédiat français (diagramme 2); noeuds mixtes: $\bar{t}$ du subjonctif p.q.p. en italien (diagramme 17), $\bar{t}_{4,5}$ du futur immédiat français (diagramme 2).

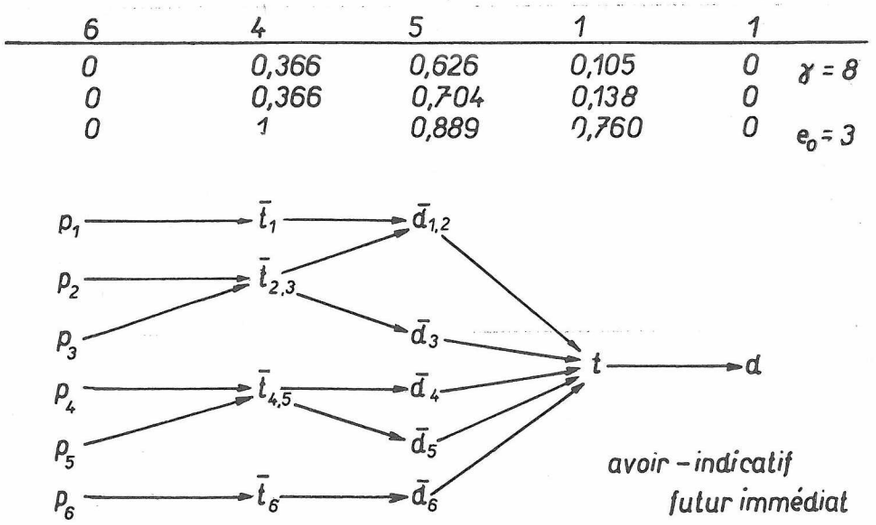


Diagramme 2.

3	2	0	
0	0,276	0	$\gamma = 7$
0	0,276	0	
0	1	0	

$e_0 = 2,807$

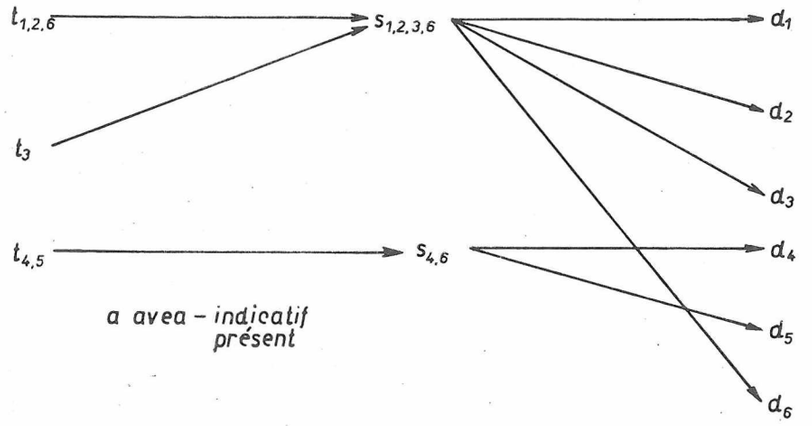


Diagramme 3.

1	1	3	
0	0,152	0	$\gamma = 5$
0	0,152	0	
0	1	0	

$e_0 = 2,318$

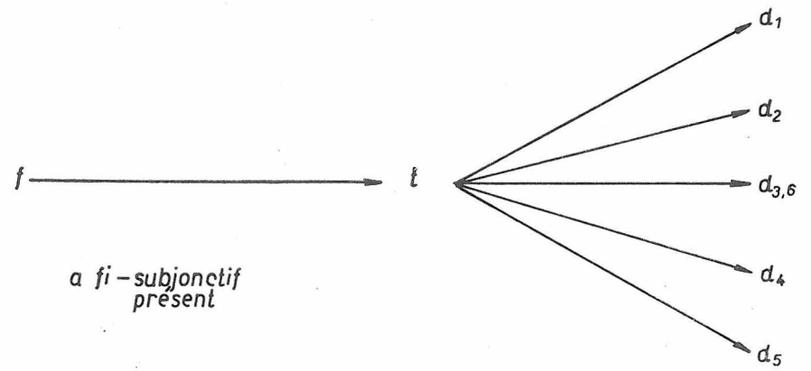


Diagramme 4.

1	6	1	4	
0,121	0,834	0,092	0	$\gamma = 9$
3,124	0,864	0,092	0	$e_0 = 3,169$
0,975	0,965	1	0	

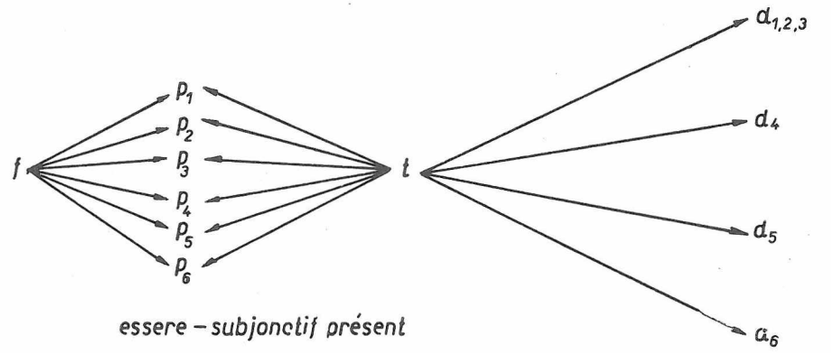


Diagramme 5.

1	1	5	
0	0,152	0	$\gamma = 5$
0	0,152	0	$e_0 = 2,318$
0	1	0	

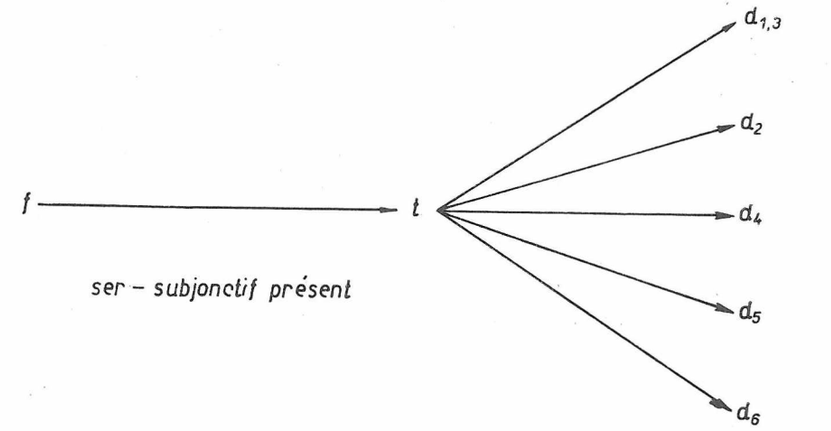


Diagramme 6.

II. Noeuds de divergence et de convergence

Nous avons groupé les graphes de chaque langue dans des tableaux, selon le mode et le temps, en indiquant la fréquence et la place des nœuds de convergence, de divergence et mixtes, et en rapportant leur nombre au nombre total des nœuds d'un graphe. Ce sont les graphes et les tableaux qui nous ont servi de base à l'interprétation des faits.

Conformément au caractère linéaire de la langue, dans toutes les langues étudiées les arcs sont orientés dans un seul sens et les éléments constitutifs des formes verbales, les formants, se succèdent toujours de gauche à droite.

Le nombre des noeuds peut caractériser une langue du point de vue typologique. Dans les langues étudiées, le nombre des noeuds d'un graphe varie de 1 (certains modes non personnels) à 20. Le chiffre maximum va de 10 (en latin, espagnol, hongrois, russe) à 12 (albanais), 14 (roumain, italien), 15 (anglais), 17 (français), 20 (allemand).

Le nombre des formants peut contribuer à définir le caractère analytique d'une langue. Les langues romanes ont plus d'éléments constitutifs que le latin, où le nombre des colonnes va seulement jusqu'à 4 (subjonctif p.q.p.)

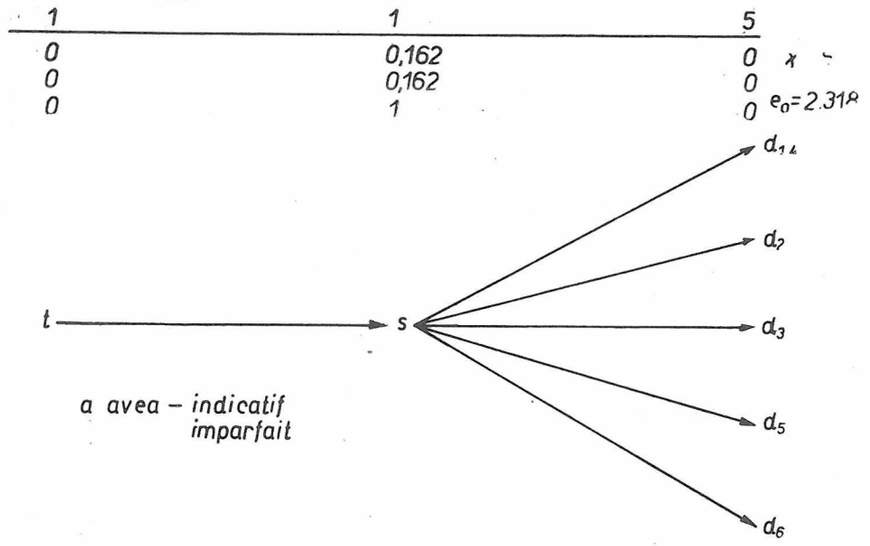


Diagramme 7.

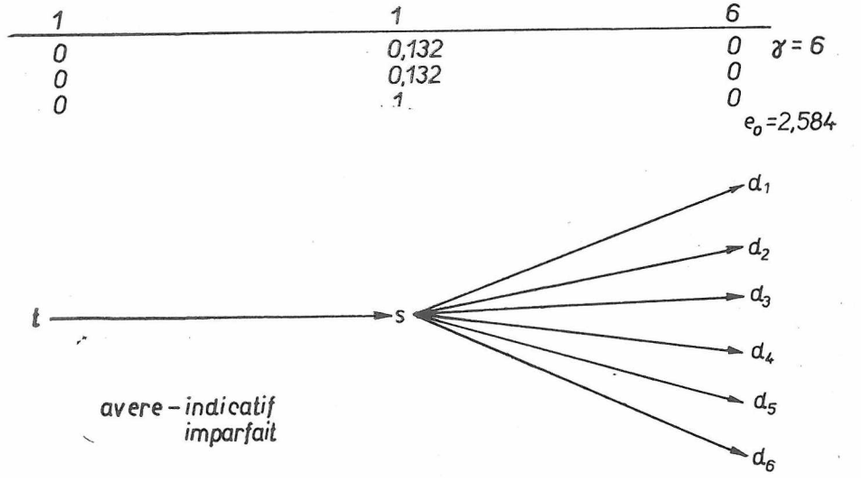


Diagramme 8.

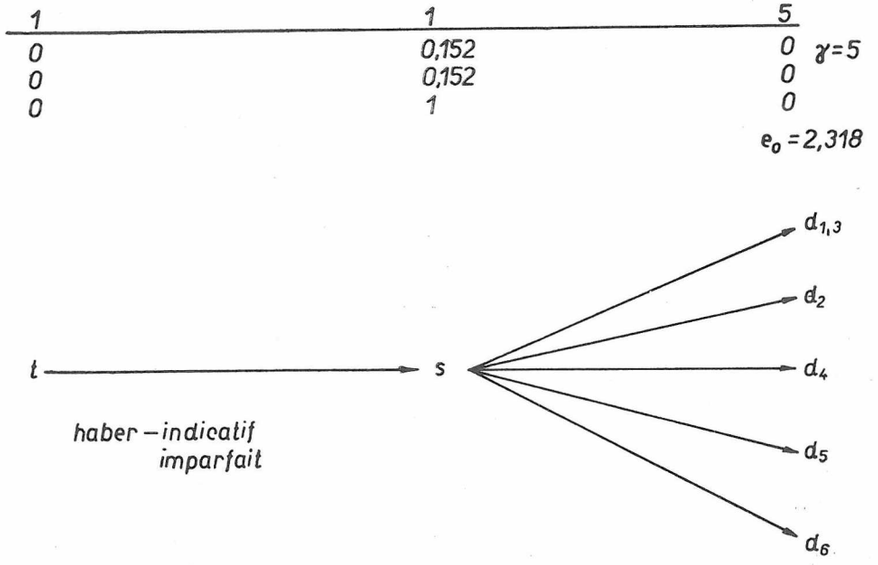


Diagramme 9.

Dans toutes les langues, le nombre des nœuds d'un graphe est inférieur au produit du nombre des lignes multiplié par celui des colonnes. C'est le résultat d'un processus d'abstraction. En roumain, les 6 lignes et 3 colonnes de l'indicatif présent du verbe *avoir* devraient donner 18 nœuds virtuels, dont 11 seulement sont réels (diagramme 3) et en français, sur 18 nœuds possibles, se réalisent 15. C'est une preuve de l'économie [21] dans l'utilisation des moyens linguistiques. Un formant peut cumuler plusieurs fonctions morphologiques. Un thème ou une désinence correspond parfois à plusieurs personnes. La désinence zéro peut fournir plus d'informations morphologiques que les autres formants [14]. A. Lombard [14] remarque le fait qu'une telle économie facilite l'activité cérébrale du locuteur. Plus le nombre des nœuds est réduit, moins l'effort du locuteur est important. — La théorie des graphes peut donc aider à l'étude du rapport entre formant et fonction.

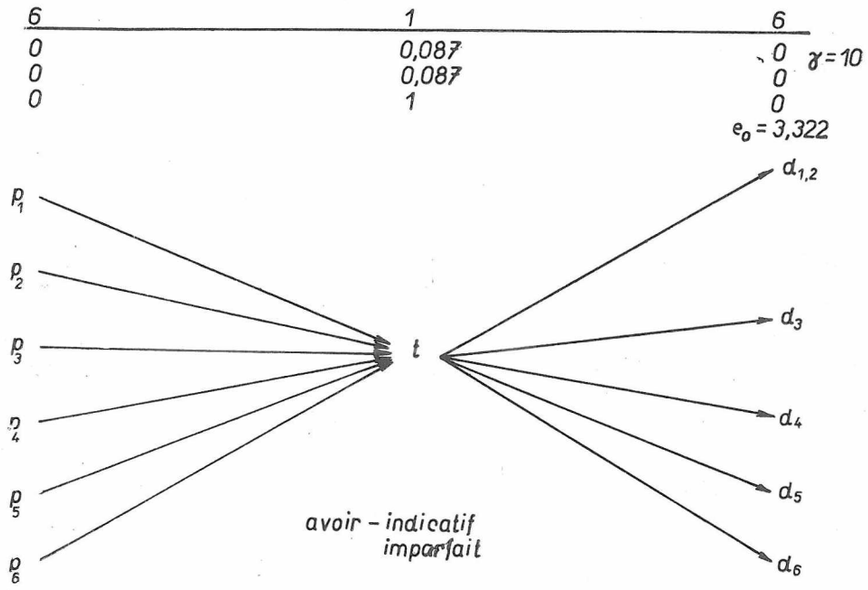


Diagramme 10.

Parmi les formants, ceux qui contribuent le plus à la délimitation du spécifique typologique d'une langue sont les désinences, par leur variété et par leur organisation diverse. Le subjonctif présent du verbe *être* en roumain, italien, espagnol n'a qu'un seul thème (diagrammes 4, 5, 6). Les désinences sont très variées et se groupent différemment dans chacune de ces langues.

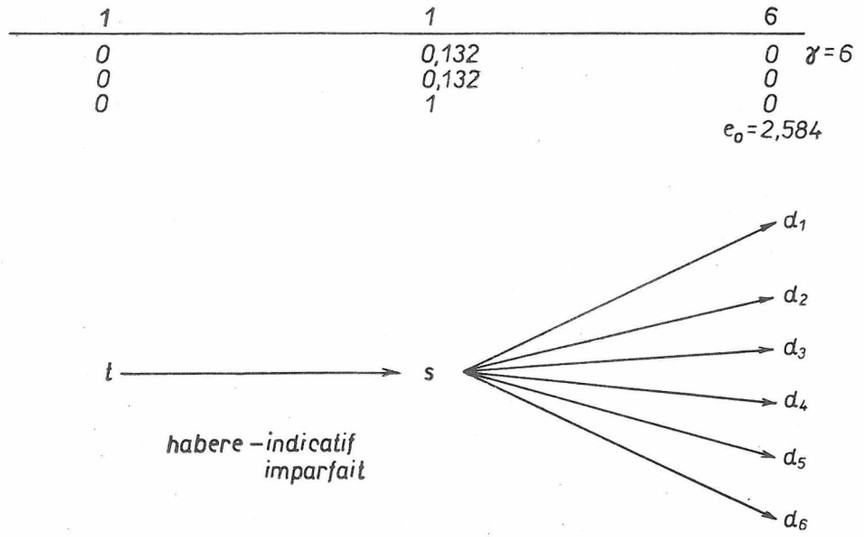


Diagramme 11.

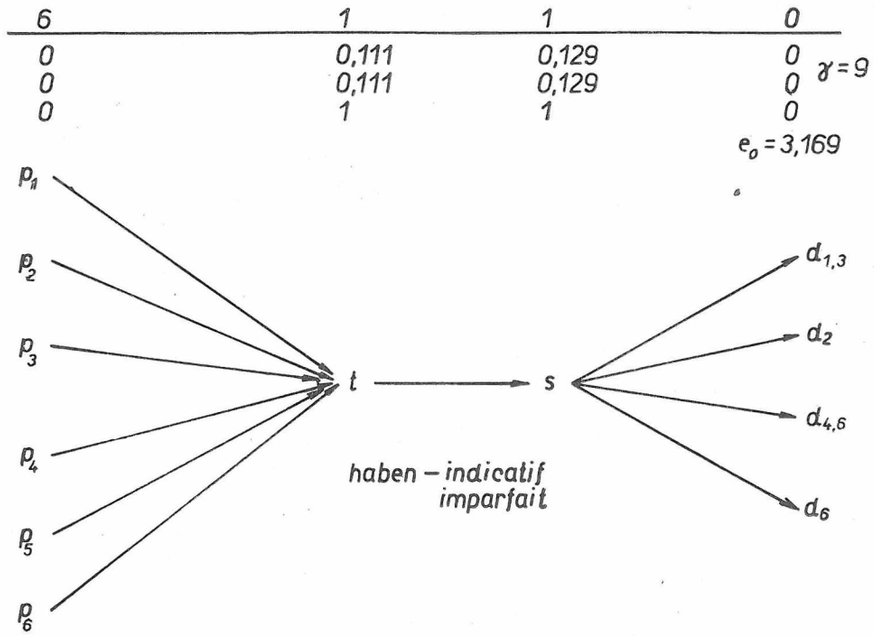


Diagramme 12.

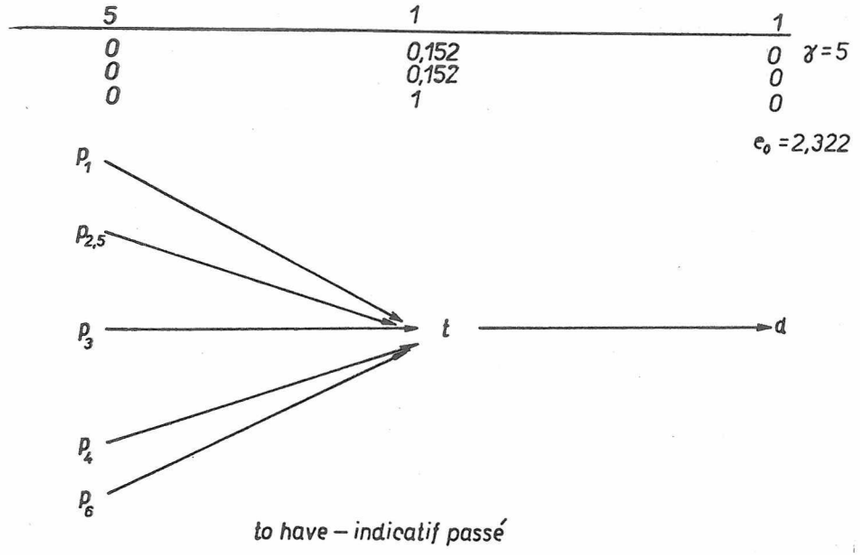


Diagramme 13.

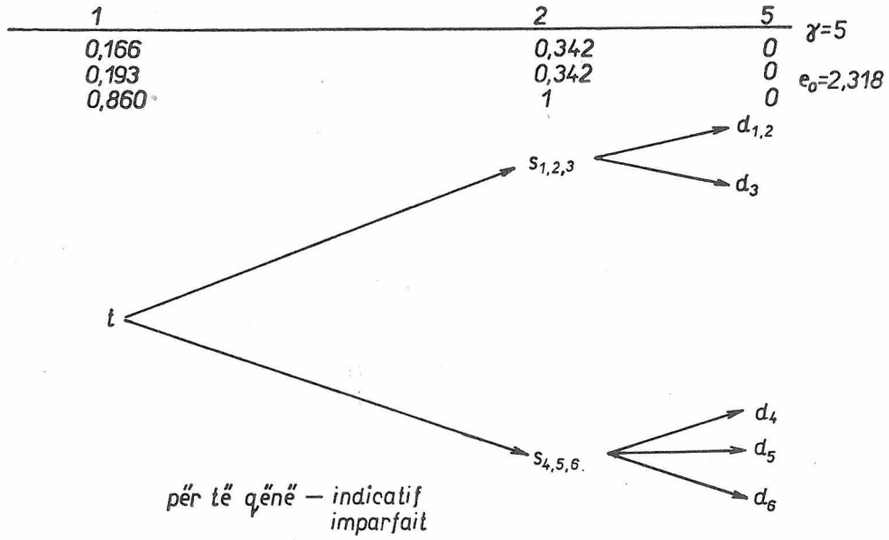


Diagramme 14.

III. La structure morphématique du paradigme verbal: représentation géométrique et matricielle de l'entropie et de la redondance

Nous avons associé à tous les temps de chaque mode un graphe où chaque noeud représente un morphème et les arcs symbolisent la succession d'un morphème à l'autre. Par exemple: le futur antérieur en roumain (diagramme 1): *voi fi avut*: *vo-* = nœud $\overline{\overline{t}}_1$; *-i* = $\bar{d}_1$; *fi* = $\bar{t}$; ∅ = s; *av-* = t; *-ut* = s et l'arc $\bar{l}_1 \rightarrow \bar{d}_1$ représente la succession de $\bar{l}_1$ à $\bar{d}_1$. La route $\bar{l}_1 \rightarrow \bar{d}_1 \rightarrow \bar{t} \rightarrow s \rightarrow t \rightarrow s$ représente la possibilité de passage d'une extrême à l'autre, à savoir de $\bar{l}_1$ (le morphème d'entrée) à s (le morphème de sortie). Il y a quatre nœuds extrêmes: trois nœuds d'entrée: $\bar{l}_1$ $\bar{l}_{2356}$ $\bar{l}_4$ et un seul nœud de sortie: s.

Au-dessus de chaque colonne du graphe nous avons indiqué dans la première ligne le nombre des nœuds (morphèmes [7]) présents dans la colonne respective: $\overline{\overline{t}} = 3$, $\overline{\bar{d}} = 6$, $\bar{t} = 1$, $s = 1$, $t = 1$, $s = 1$. Au-dessous des nombres concernants les morphèmes de chaque colonne on trouve, dans la première ligne, l'entropie relative; dans la deuxième l'entropie conditionnée; dans la troisième le rapport de ces entropies Les derniers nombres représentent le nombre total des routes: $\gamma = 6$ dans ce cas et l'entropie générale de degré zéro du chaque graphe — dans ce cas (diagramme 1) $e_0 = 2{,}318$.

La route γ est calculée à l'aide de la formule établie par M. S. Marcus (19 : 50): $\gamma = A - N + e$ où A = le nombre total des arcs,

N = le nombre des nœuds,
(morphèmes d'après J. Horecký (7 : 86—87)),
e = le nombre des nœuds extrêmes.

Dans le cas où le graphe contient un nœud mixte (réel ou virtuel) la formule proposée pour calculer γ n'est plus opérante (diagrammes 3, 5, 10, 12, 17). L'entropie de degré zéro est donnée par la formule indiquée toujours par M. S. Marcus

$$e_0 = \log_2 \gamma.$$

Dans le cas où la formule de γ n'est pas opérante l'entropie e_0 est fausse.

L'entropie relative a été calculée en partant de la formule [7, 8, 17, 18]

$$H = -\sum_{i=1}^{n} p_i \log p_i$$

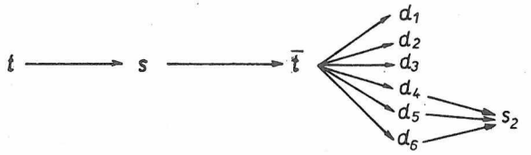
formule qui indique l'information moyenne qu'on peut obtenir envisageant toutes les possibilités. M. Marcus fait une précision en indiquant pour les collectives finites, limitées d'objets la base deux du logarithme, mais qui ne change, en fin des comptes la formule initiale.

La comparaison de deux langues du point de vue de l'entropie peut présenter des difficultés à cause des différences qui existent dans les inventaires morphématiques.

1	1	1	6	1
0	0,150	0,422	0,450	0,159
0	0,150	0,422	0,498	0,263
0	1		0,903	0,604

$\gamma = 3$

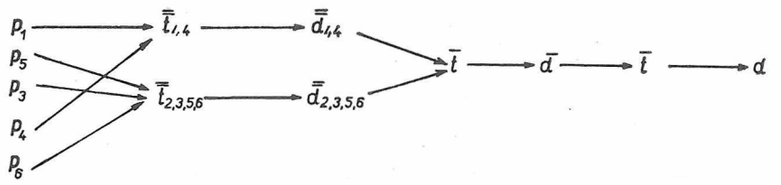
$P_0 = 1{,}584$



lenni – indicatif futur analytique

Diagramme 15.

5	2	2	1	1	1	1	
0	0,277	0,270	0,142	0,135	0,135	0	
0	0,296	0,280	0,152	0,135	0,135	0	$\gamma = 5$
0	0,935	0,937	0,937	1	1	0	$l_o = 2{,}322$

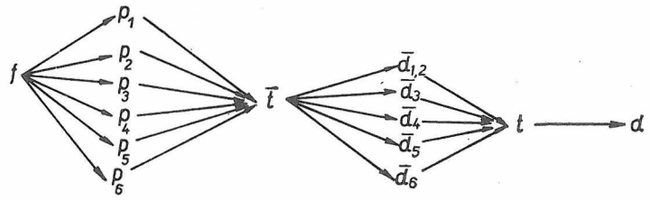


to have – conditionnel passé

Diagramme 16.

Ces différences en créent d'autres entre les entropies des deux langues. C'est pourquoi à la place de H_n (l'entropie absolue d'ordre n) on introduit l'entropie relative d'ordre n h_n qui est définie par le rapport entre l'entropie absolue d'ordre n H_n et l'entropie absolue d'ordre 0 H_0 (cette entropie indique l'information obtenue par la

1	6	1	5	1	1	
0,110	0,762	0,080	0,635	0,110	0	$\gamma = 10$
0,116	0,810	0,080	0,675	0,110	0	$l_o = 3{,}322$
0,948	0,940	1	0,940	1	0	



avere – subjonctif plus-que-parfait

Diagramme 17.

précision d'une variante de n variantes également possibles; elle est égale avec le logarithme de base 2 de n). Donc

$$H_n = -\sum_{i=1}^{n} p_i \quad \log_2 p_i,$$

$$H_0 = \log_2 n,$$

$$h_n = \frac{H_n}{H_0} = -\frac{\sum_{i=1}^{n} p_i \log_2 p_i}{\log_2 n}, \qquad [17 : 240—253]$$

où $p = \frac{1}{k}$, k étant le degré du nœud, donné par le nombre total des arcs qui entrent dans le nœud et qui en sortent. Cette entropie est calculée par colonne. Pour le graphe tout entier on a calculé l'entropie générale e_0 d'après la formule donnée ci-dessus (17 : 240).

L'impossibilité de calculer, pour certains graphes l'entropie e_0, qui est conditionnée par γ indique l'importance de la nature des nœuds (notamment des nœuds mixtes) pour la structure du graphe (et donc du temps ou du mode respectif).

Pour l'entropie conditonnée on a employé la même formule (h_n), mais sans envisager les nœuds linguistiquement redondants.

D'après les valeurs des entropies calculées on peut faire une hierarchie des temps, des modes et des langues, hiérarchie étroitement liée à la charge affective du temps ou du mode respectif.

L'entropie générale de degré zéro prend les plus grandes valeurs, sans tenir comptes des cas où l'entropie est fausse: au futur (il est normal que l'indétermination augmente

quand l'improbabilité de l'action augmente, or c'est notamment le cas des actions exprimées par le futur), et l'imparfait qui en général est très employé dans toutes les langues. Donc, en général la valeur de l'entropie e_0 augmente pour les actions en train de s'accomplir dans le plan du passé ou du futur, donc là où l'improbabilité est plus grande. Mais il y a des différences spécifiques à une certaine langue. Par exemple, la valeur de l'entropie du futur analytique en hongrois n'est pas trop grande.

Après ces temps c'est le subjonctif qui suit dans l'ordre décroissant de la valeur de l'entropie. Il est évident que le subjonctif est plus chargé d'affectivité que les temps précis de l'indicatif. Cela est valable aussi pour les cas mentionnés plus haut car là où l'improbabilité augmente la charge affective augmente elle aussi.

Conclusions

L'analyse des formants peut fournir quelques critériums pour une comparaison typologique des langues. Un formant se caractérise par sa présence ou son absence, par sa fréquence dans les nœuds de divergence, de convergence ou mixtes, par sa fréquence relative (par rapport à d'autres formants).

Aux modes personnels, la présence du pronom exclut l'existence des nœuds de divergence et implique l'existence des nœuds de convergence dans *t* (tableaux 4, 6, 7). L'absence du pronom personnel implique l'absence des nœuds de convergence (tableau 1). Les nœuds mixtes sont plus nombreux dans les langues où l'emploi du pronom est obligatoire. Le français est caractérisé par le plus grand nombre de nœuds mixtes (tableau 4). L'allemand en a aussi beaucoup (tableau 6). Les noeuds de convergence-divergence n'apparaissent pas en latin (tableau 1), anglais (tableau 7), espagnol (tableau 5), albanais (tableau 8), russe (tableau 10), hongrois (tableau 9). En roumain il n'y a que deux nœuds mixtes: indicatif présent, subjonctif présent (tableau 2).

Aux modes non personnels, seuls le latin et le roumain présentent des nœuds de divergence (le roumain en a deux, le latin sept) (tableau 1, 2). Les autres langues n'ont pas de nœuds de convergence, de divergence et mixtes.

Dans toutes les langues, le thème est l'élément le plus fréquent dans les nœuds de convergence, de divergence et mixtes. Le latin fait exception, le suffix étant ici plus fréquent: les suffixes forment, sur 19 nœuds, 13 nœuds de divergence (tableau 1). Après le latin, le plus grand nombre de suffixes se rencontre en roumain (7/16) (tableau 2). Quant à la répartition du thème dans les nœuds de divergence, il y a une monotypie structurale entre le latin (6/19) (tableau 1), le roumain (8/16) (tableau 2) et l'italien (9/19) (tableau 3). Par rapport au latin, le nombre des nœuds de divergence est plus réduit dans les langues romanes, tandis que les nœuds de convergence y sont plus fréquents. Les langues romanes présentent en plus des nœuds mixtes (tableau 2, 3, 4, 5, 11).

Tableau 1

Latin — *avoir*

Mode et temps	Divergence			Convergence			Convergence-divergence		
	Nr.	Place	Rapport	Nr.	Place	Rapport	Nr.	Place	Rapport
Indicatif prés.	1	*t*	1/7	0	0	0/7	0	0	0/7
Indicatif imparfait	1	*s*	1/8	0	0	0/8	0	0	0/8
Indicatif part.	1	*s*	1/8	0	0	0/8	0	0	0/8
Indicatif p.-q.-p.	1	*s*	1/8	0	0	0/8	0	0	0/8
Indicatif futur	1	*s*	1/8	0	0	0/8	0	0	0/8
Futur antér.	1	*s*	1/8	0	0	0/8	0	0	0/8
Subj. prés.	1	*s*	1/8	0	0	0/8	0	0	0/8
Subj. imp.	1	*s*	1/8	0	0	0/8	0	0	0/8
Subj. parf.	1	*s*	1/8	0	0	0/8	0	0	0/8
Subj. p. q. p.	1	*s*	1/8	0	0	0/8	0	0	0/8
Impér. prés.	1	*t*	1/3	0	0	0/3	0	0	0/3
Impér. futur	1	*t*	1/5	0	0	0/5	0	0	0/5
Participe	1	*t*	1/10	0	0	0/10	0	0	0/10
Part. futur	1	*s*	1/5	0	0	0/5	0	0	0/5
Part. passif	1	*s*	1/5	0	0	0/5	0	0	0/5
Inf.	0	0	0/2	0	0	0/2	0	0	0/2
Inf. futur	1	*s*	1/5	0	0	0/5	0	0	0/5
Gérondif	1	*t*	1/5	0	0	0/5	0	0	0/5
Part. futur passif	1	*t*	1/4	0	0	0/4	0	0	0/4
Supin	1	*s*	1/4	0	0	0/4	0	0	0/4

Tableau 2

Roumain — *avoir*

Mode et temps	Divergence			Convergence			Convergence-divergence		
	Nr.	Place	Rapport	Nr.	Place	Rapport	Nr.	Place	Rapport
Indic. prés.	1	$s_{4,5}$	1/18	0	0	0/18	1	$s_{1,2,3,6}$	1/18
Ind. imp.	1	s	1/7	0	0	0/7	0	0	0/7
Passé simple I	1	s'	1/8	0	0	0/8	0	0	0/8
Passé simple II	1	s''	1/9	0	0	0/9	0	0	0/9
Plus-que-parfait	1	s'''	1/10	0	0	0/10	0	0	0/10
Passé composé	1	$\bar{t}$	1/9	1	t	1/9	0	0	0/9
Futur	1	$\bar{t}_{2,3,5,6}$	1/11	1	t	1/11	0	0	0/11
Futur antér.	1	$\bar{t}_{2,3,5,6}$	1/14	1	$\bar{t}$	1/14	0	0	0/14
Subj. prés.	1	f	1/10	0	0	0/10	1	$s_{1,2,3,6}$	1/10
Subj. passé	0	0	0/6	0	0	0/6	0	0	0/6
Condit. prés.	1	$\bar{t}_{2,4,5}$	1/9	2	$\bar{d}_{1,3,6}$ t	2/9	0	0	0/9
Condit. passé	1	$\bar{t}_{2,4,5}$	1/12	2	$\bar{t}$; $\bar{d}$	2/12	0	0	0/12
Présomptif I	1	$\bar{t}_{2,4,5}$	1/11	2	$\bar{t}\bar{d}_{1,3,6}$	2/11	0	0	0/11
Présomptif II	0	0	0/5	0	0	0/5	0	0	0/5
Présomptif III	1	$\bar{t}_{2,3,5,6}$	1/13	1	$\bar{t}$	1/13	0	0	0/13
Impér. aff.	0	0	0/4	0	0	0/4	0	0	0/4
Impér. nég.	1	t	1/6	0	0	0/6	0	0	0/6
Inf. prés.	0	0	0/3	0	0	0/3	0	0	0/3
Inf. passé	0	0	0/6	0	0	0/6	0	0	0/6
Participe	2	s' $s''_{1,3,4}$	2/8	0	0	0/8	0	0	0/8
Gérondif	0	0	0/2	0	0	0/2	0	0	0/2
Supin	0	0	0/4	0	0	0/4	0	0	0/4

Tableau 3

Italien — *avoir*

Mode et temps	Divergence			Convergence			Convergence-divergence		
	Nr.	Place	Rapport	Nr.	Place	Rapport	Nr.	Place	Rapport
Ind. prés.	1	$t_{2,3,6}$	1/10	0	0	0/10	0	0	0/10
Imp.	1	s	1/8	0	0	0/8	0	0	0/8
Passé simple	2	$t_{3,6}$ $t_{2,4,5}$	2/8	0	0	0/8	0	0	0/8
Passé comp.	1	$\bar{t}_{2,3,6}$	1/11	1	t	1/11	0	0	0/11
P. q. p.	1	$\bar{s}$	1/9	1	t	1/9	0	0	0/9
Futur	1	s	1/8	0	0	0/8	0	0	0/8
Futur antér.	1	$\bar{s}$	1/9	1	t	1/9	0	0	0/9
Trapassato remoto	2	$\bar{t}_{1,3,6}$ $\bar{t}_{2,4,5}$	2/9	1	t	1/9	0	0	0/9
Subj. prés.	2	p s	2/12	1	t	1/12	0	0	0/12
Subj. imp.	1	f	1/13	0	0	0/13	1	t	1/13
Subj. passé	2	f $\bar{s}$	2/13	2	$\bar{t}$ t	2/13	0	0	0/13
Subj. p. q. p.	1	f	1/14	1	t	1/14	1	$\bar{t}$	1/14
Condit. prés.	1	t	1/7	0	0	0/7	0	0	0/7
Cond. passé	1	$\bar{t}$	1/8	1	t	1/8	0	0	0/8
Impér.	1	t	1/6	0	0	0/6	0	0	0/6
Inf. prés.	0	0	0/2	0	0	0/2	0	0	0/2
Inf. passé	0	0	0/3	0	0	0/3	0	0	0/3
Part. prés.	0	0	0/2	0	0	0/2	0	0	0/2
Part. passé	0	0	0/2	0	0	0/2	0	0	0/2
Gérondif	0	0	0/2	0	0	0/2	0	0	0/2
Gérondif passé	0	0	0/4	0	0	0/4	0	0	0/4

Tableau 4

Français — *avoir*

Mode et temps	Divergence			Convergence			Convergence-divergence		
	Nr.	Place	Rapport	Nr.	Place	Rapport	Nr.	Place	Rapport
Indic. prés.	0	0	0/15	0	0	0/15	3	$t_{1,6}$ $t_{2,3}$ $t_{4,5}$	3/15
Imp.	0	0	0/13	0	0	0/13	1	t	1/13
Passé simple	0	0	0/13	0	0	0/13	2	$t_{1,2,3,6}$ $t_{4,5}$	2/13
Passé comp.	0	0	0/16	1	$\bar{t}$	1/16	3	$\bar{t}_{1,6}$ $\bar{t}_{2,3}$ $\bar{t}_{4,5}$	3/16
P. q. p.	0	0	0/14	1	t	1/14	1	$\bar{t}$	1/14
Passé antér.	0	0	0/14	1	t	1/14	2	$\bar{t}_{1,2,3,6}$ $\bar{t}_{4,5}$	2/14
Futur	0	0	0/13	0	0	0/13	1	t	1/13
Futur antér.	0	0	0/14	1	t	1/14	1	$\bar{t}$	1/14
Cond. prés.	0	0	0/12	0	0	0/12	1	t	1/12
Cond. passé	0	0	0/13	1	t	1/13	1	t	1/13
Futur périphr. I	0	0	0/17	2	t $\bar{d}_{1,2}$	2/17	2	$\bar{t}_{2,3}$ $\bar{t}_{4,5}$	2/17
Futur périphr. II	0	0	0/14	1	t	1/14	1	$\bar{t}$	1/14
Subj. prés.	0	0	0/14	0	0	0/14	2	$t_{1,2,3,6}$ $t_{4,5}$	2/14
Subj. imp.	0	0	0/14	0	0	0/14	1	$t_{1,2,4,5,6}$	1/14
Subj. passé	0	0	0/15	1	t	1/15	2	$\bar{t}_{1,2,3,6}$ $\bar{t}_{4,5}$	2/15
Subj. p. q. p.	0	0	0/15	1	t	1/15	1	$\bar{t}_{1,2,4,5,6}$	1/15
Impér.	0	0	0/5	0	0	0/5	0	0	0/5
Inf. prés.	0	0	0/2	0	0	0/2	0	0	0/2
Inf. passé	0	0	0/3	0	0	0/3	0	0	0/3
Part. prés.	0	0	0/2	0	0	0/2	0	0	0/2
Part. passé I	0	0	0/1	0	0	0/1	0	0	0/1
Part. passé II	0	0	0/3	0	0	0/3	0	0	0/3

Tableau 5

Espagnol — *avoir*

Mode et temps	Divergence			Convergence			Convergence-divergence		
	Nr.	Place	Rapport	Nr.	Place	Rapport	Nr.	Place	Rapport
Indic. prés.	2	$t_{1,4}$ $t_{2,3,6}$	2/9	0	0	0/9	0	0	0/9
Imp.	1	t	1/6	0	0	0/6	0	0	0/6
Preterito	1	t	1/7	0	0	0/7	0	0	0/7
Futuro	2	$\bar{t}_{1,4}$ $\bar{t}_{2,3,6}$	2/10	2	t $d_{1,3}$	2/10	0	0	0/10
P. q. p.	1	$\bar{t}$	1/8	1	t	1/8	0	0	0/8
Pret. anter.	1	$\bar{t}$	1/9	1	t	1/9	0	0	0/9
Futuro perf.	1	$\bar{t}$	1/9	1	t	1/9	0	0	0/9
Potencial simple	1	t	1/6	0	0	0/6	0	0	0/6
Potencial comp.	1	$\bar{t}$	1/8	1	t	1/8	0	0	0/8
Subj. pres.	1	t	1/6	0	0	0/6	0	0	0/6
Subj. perf.	1	$\bar{t}$	1/8	1	t	1/8	0	0	0/8
Subj. imp. I, II	1	t	1/6	0	0	0/6	0	0	0/6
Subj. p. q. p.	1	$\bar{t}$	1/8	1	t	1/8	0	0	0/8
Subj. futuro imperf.	1	t	1/6	0	0	0/6	0	0	0/6
Subj. futuro perf.	1	$\bar{t}$	1/8	1	t	1/8	0	0	0/8
Imperativo	1	$t_{2,3,5}$	1/8	0	0	0/8	0	0	0/8
Infinitivo	0	0	0/2	0	0	0/2	0	0	0/2
Inf. perf.	0	0	0/4	0	0	0/4	0	0	0/4
Gerundio	0	0	0/2	0	0	0/2	0	0	0/2
Gerundio compuesto	0	0	0/4	0	0	0/4	0	0	0/4
Participio	0	0	0/2	0	0	0/2	0	0	0/2

Tableau 6

Allemand — *avoir*

Mode et temps	Divergence			Convergence			Convergence-diverg.		
	Nr.	Place	Rapport	Nr.	Place	Rapport	Nr.	Place	Rapport
Ind. prés.	0	0	0/17	0	0	0/17	1	*t*	1/17
Imp.	1	*s*	1/16	1	*t*	1/16	0	0	0/16
Passé comp.	0	0	0/20	1	*r*	1/20	1	$\bar{t}$	1/20
P. q. p.	1	$\bar{s}$	1/19	2	$\bar{t}$; *r*	2/19	0	0	0/19
Futur	0	0	0/14	3	$\bar{t}_{1,4,5,6}$ $\bar{t}_{2,3}$ *t*	3/14	0	0	0/14
Futur antér.	0	0	0/17	3	$\bar{\bar{t}}_{1,4,5,6}$ $\bar{\bar{t}}_{2,3}$ *s*	3/17	0	0	0/17
Subj. prés.	1	*s*	1/16	1	*t*	1/16	0	0	0/16
Subj. passé	1	*s*	1/19	2	$\bar{t}$; *r*	2/19	0	0	0/19
Subj. imp.	0	0	0/15	0	0	0/15	1	$\bar{t}$	1/15
Subj. p. q. p.	0	0	0/18	1	*r*	1/18	1	$\bar{t}$	1/18
Subj. futur	0	0	0/17	1	*t*	1/17	1	$\bar{t}$	1/17
Subj. futur antér.	0	0	0/20	1	*r*	1/20	1	$\bar{t}$	1/20
Condit. I	0	0	0/17	1	*t*	1/17	1	$\bar{t}$	1/17
Condit. II	0	0	0/20	1	*r*	1/20	1	$\bar{t}$	1/20
Impératif	1	*t*	1/3	0	0	0/3	0	0	0/3
Impératif nég.	1	*t*	1/4	1	*f*	1/4	0	0	0/4
Infinitif prés.	0	0	0/2	0	0	0/2	0	0	0/2
Infinitif passé	0	0	0/3	0	0	0/3	0	0	0/3
Part. prés.	0	0	0/2	0	0	0/2	0	0	0/2
Part. passé	0	0	0/3	0	0	0/3	0	0	0/3

Tableau 7

Anglais — *avoir*

Mode et temps	Divergence			Convergence			Convergence-diverg.		
	Nr.	Place	Rapport	Nr.	Rapport	Place	Nr.	Place	Rapport
Ind. prés.	0	0	0/11	2	$t_{1,2,4,5,6}$ t_3	2/11	0	0	0/11
Passé composé	0	0	0/13	3	*t*; $\bar{t}_3$ $\bar{t}_{1,2,4,5,6}$	3/13	0	0	0/13
Imp.	0	0	0/9	1	*t*	1/9	0	0	0/9
P. q. p.	0	0	0/9	1	$\bar{t}$	1/9	0	0	0/9
Futur	0	0	0/13	3	*t*; $\bar{t}_{1,4}$ $\bar{t}_{2,3,5,6}$	3/13	0	0	0/13
Futur antér.	0	0	0/15	3	$\bar{t}$; $\bar{\bar{t}}_{1,4}$ $\bar{\bar{t}}_{2,3,5,6}$	3/15	0	0	0/15
Futur du passé	0	0	0/13	3	$\bar{t}$; $\bar{t}_{1,4}$ $\bar{t}_{2,3,5,6}$	3/13	0	0	0/13
Futur ant. du passé	0	0	0/15	3	$\bar{t}$; $\bar{\bar{t}}_{1,4}$ $\bar{t}_{2,3,5,6}$	3/15	0	0	0/15
Subj. prés.	0	0	0/9	1	*t*	1/9	0	0	0/9
Subj. passé	0	0	0/11	1	$\bar{t}$	1/11	0	0	0/11
Cond. prés.	0	0	0/13	3	*t*; $\bar{t}_{1,4}$ $\bar{t}_{2,3,5,6}$	3/13	0	0	0/13
Cond. passé	0	0	0/15	3	$\bar{t}$; $\bar{\bar{t}}_{1,4}$ $\bar{\bar{t}}_{2,3,5,6}$	3/15	0	0	0/15
Impératif	0	0	0/2	0	0	0/2	0	0	0/2
Infinitif prés.	0	0	0/3	0	0	0/3	0	0	0/3
Infinitif passé	0	0	0/5	0	0	0/5	0	0	0/5
Part. prés.	0	0	0/2	0	0	0/2	0	0	0/2
Part. passé I	0	0	0/2	0	0	0/2	0	0	0/2
Part. passé II	0	0	0/4	0	0	0/4	0	0	0/4

Tableau 8

Albanais — *avoir*

Mode et temps	Divergence			Convergence			Convergence-diverg.		
	Nr.	Place	Rapport	Nr.	Place	Rapport	Nr.	Place	Rapport
Ind. prés.	2	$t_{1,6}$, $t_{2,4,5}$	2/9	0	0	0/9	0	0	0/9
Imparfait	1	s	1/8	0	0	0/8	0	0	0/8
Parfait	2	$\tilde{t}_{1,3,6}$ $\tilde{t}_{2,4,5}$	2/10	1	t	1/10	0	0	0/10
Aoriste I	1	t	1/6	0	0	0/6	0	0	0/6
Aoriste II	1	$\tilde{t}$	1/10	1	t	1/10	0	0	0/10
P. q. p.	1	$\tilde{s}$	1/10	1	t	1/10	0	0	0/10
P. q. p. II	2	$\tilde{\tilde{t}}_{1,3,6}$ $\tilde{\tilde{t}}_{2,4,5}$	2/12	1	$\tilde{t}$	1/12	0	0	0/12
Futur	1	t	1/9	0	0	0/9	0	0	0/9
Futur II	1	$\tilde{t}$	1/11	1	t	1/11	0	0	0/11
Subj. prés.	1	t	1/8	0	0	0/8	0	0	0/8
Conj. passé	1	$\tilde{t}$	1/10	1	t	1/10	0	0	0/10
Condit. imp.	1	s	1/10	0	0	0/10	0	0	0/10
Condit. p. q. p.	1	$\tilde{s}$	1/12	1	t	1/12	0	0	0/12
Admir. prés.	1	s	1/8	0	0	0/8	0	0	0/8
Admir. passé	1	$\tilde{s}$	1/10	1	t	1/10	0	0	0/10
Subj. adm. imp.	2	f $t_{1,2,4,5,6}$	2/9	0	0	0/9	0	0	0/9
Subj. adm. p. q. p.	2	f $\tilde{t}_{1,2,4,5,6}$	2/11	1	t	1/11	0	0	0/11
Optatif	1	$t_{1,2,4,5,6}$	1/8	0	0	0/8	0	0	0/8
Impératif	1	t	1/3	0	0	0/3	0	0	0/3
Infinitif prés.	0	0	0/4	0	0	0/4	0	0	0/4
Infinitif passé	0	0	0/6	0	0	0/6	0	0	0/6
Part. prés.	0	0	0/2	0	0	0/2	0	0	0/2
Part. passé	0	0	0/4	0	0	0/4	0	0	0/4
Gérondif prés.	0	0	0/3	0	0	0/3	0	0	0/3
Gérondif passé	0	0	0/5	0	0	0/5	0	0	0/5

Tableau 9

Hongrois — *avoir*

Mode et temps	Divergence			Convergence			Convergence-diverg.		
	Nr.	Place	Rapport	Nr.	Place	Rapport	Nr.	Place	Rapport
Ind. prés.	1	p	1/9	2	$s_{4,5,6}$ t	2/9	0	0	0/9
Imparfait	1	p	1/9	2	$s_{4,5,6}$ t	2/9	0	0	0/9
Parf.	1	p	1/9	2	$s_{4,5,6}$ t	2/9	0	0	0/9
Futur	1	p	1/9	2	$s_{4,5,6}$ t	2/9	0	0	0/9
Subj. prés.	1	p	1/9	2	$s_{4,5,6}$ t	2/9	0	0	0/9
Subj. passé	1	p	1/10	1	$s_{4,5,6}$	1/10	0	0	0/10
Optatif pr.	1	p	1/9	1	$s_{4,5,6}$	1/9	0	0	0/9
Optatif parf.	1	p	1/9	2	$s_{4,5,6}$ t	2/9	0	0	0/9
Infinitif prés.	0	0	0/2	0	0	0/2	0	0	0/2
Part. prés.	0	0	0/2	0	0	0/2	0	0	0/2
Part. passé	0	0	0/2	0	0	0/2	0	0	0/2
Gérondif	0	0	0/2	0	0	0/2	0	0	0/2

Tableau 10

Russe — *avoir*

Mode et temps	Divergence			Convergence			Convergence-diverg.		
	Nr.	Place	Rapport	Nr.	Place	Rapport	Nr.	Place	Rapport
Ind. prés.	1	f	1/10	1	t	1/10	0	0	0/10
Passé	1	f	1/10	0	0	0/10	0	0	0/10
Futur	1	f	1/10	1	t	1/10	0	0	0/10
Participe	0	0	0/2	0	0	0/2	0	0	0/2

Tableau 11

Divergence			Convergence		Mixtes		Total
Roumain	16	8 *t* 7 *s*, *f*	10	7 *t* 3 *d*	2	*s*	28/181
Latin	19	6 *t* 13 *s*	0		0		19/127
Italien	19	9 *t* 6 *s* 1 *p* 3 *f*	9	2	2	*t*	30/160
Espagnol	19	*t*	9	8 *t* 1 *d*	0		28/143
Français	1	*t*	10	9 *t* 1 *d*	25	*t*	36/242
Allemand	6	2 *t* 4 *s*	19	11 *t* 6 *r* 1 *s* 1 *f*	8	*t*	33/261
Anglais		0	27	*t*	0		27/164
Albanais	23	16 *t* 5 *s* 2 *f*	9	*t*	0		32
Hongrois	8	7 *f* 1 *s*	14	8 *s* 6 *t*	0		22/81
Russe	3	*f*	2	*t*	1	*t*	6/34

Il est possible d'établir un rapport direct entre la complexité d'une microstructure morphématique du verbe et le degré de l'indétermination représenté par l'entropie interne. L'entropie augmente proportionellement avec l'affectivité.

L'analyse de la fréquence et de la structuration des formants verbaux permet de comparer des langues très différentes. La description de la conjugaison des verbes auxiliaires à l'aide de la théorie des graphes offre une méthode unitaire pour l'étude typologique de la flexion verbale et de la morphologie en générale.

Symboles

f — flectif invariable
p — pronom personnel
t — thème du verbe
s — suffixe morphologique (s' s'')
d — désinence
r — préfixe morphologique
$\left.\begin{matrix} \bar{t}\,\bar{t} \\ \bar{s}\,\bar{s} \\ \bar{d}\,\bar{d} \end{matrix}\right\}$ les verbes auxiliaires

BIBLIOGRAPHIE

[1] BĂDESCU, A. L.: Gramatica limbii engleze. București 1963.
[2] CIPO, K.: Gramatika shqipe. Tiranë 1949.
[3] Gramatica limbii române. Vol. I. București 1963.
[4] Grammaire Larousse du français contemporain. 1964.
[5] GUȚU-ROMALO, V.: Morfologia structurală a limbii române. București 1968.
[6] GUȚU-ROMALO, V.: Pour une description de la flexion verbale du roumain. Cahiers de linguistique théorique et appliquée, II, 1965, p. 71—114.
[7] HORECKÝ, J.: Morfematická štruktúra slovenčiny. Bratislava 1964.
[8] IAGLOM, A. M.—IAGLOM, I. M.: Probabilitate și informații. București 1962.
[9] JORDAN, J.—DUHĂNEANU, C.: Curs de gramatica limbii spaniole. București 1963.
[10] KIS, E.—ANGHEL, I.—COMȘULEA, E.: Description d'un aspect syntaxique de la langue roumaine à l'aide de la théorie des graphes. In: Revue roumaine de linguistique, XI, 1965, nr. 5, p. 469—479.
[11] KIS, E.—COMȘULEA, E.—ANGHEL, I.: The order of the syntactic elements of principal sentences in the Romanian language. In: Computational Linguistics, V, 1967, p. 431—442.
[12] KIS, E.—OȘIANU, F.: Cu privire la indicativul prezent al verbelor auxiliare în limbile romanice. In: Buletinul Institutului Pedagogic Baia Mare. Vol. II, 1967.
[13] LĂZĂRESCU, G.: Curs de gramatica limbii italiene. București 1963.
[14] LOMBARD, A.: Tradition latin et tradition slave. Le roumain, résultat de leur fusion. In: Acta Congressus Madvigiani—Proceedings of the Second International Congress of Classical Studies, Copenhagen 1964; Vol. V, Copenhagen 1957, p. 115.
[15] MALKIEL, J.: Los interfijos hispánicos. Problemas de linguistica historica y estructural. II Miscelánea homenaje a André Martinet 1958.
[16] MANGUL, N.—VASCENCO, V.—OIȚĂ, I.: Limba rusă literară contemporană. București 1963.
[17] MARCUS, S.—NICOLAU, E.—STATI, S.: Introducere în lingvistica matematică. București 1966.
[18] MARCUS, S.: Lingvistică matematică. Mode matematice în lingvistică. București 1963.
[19] MARCUS, S.—VASILIU, E.: Teoria grafelor și consonantismul limbii române. In: Fonetică și dialectologie, III, 1962.
[20] MOISIL, Gr. C.: Probleme puse de traduceres automată. Conjugarea verbelor în limba română. Studii și cercetări lingvistice, vol. 11, 1960; Revue roumaine de linguistique, V, 1960, nr. 2.

[21] PÎRLOG, M.: Gramatica limbii latine. Bucureşti 1966.
[22] PUŞCARIU, S.: Le morphonème et l'économie de la langue. Études linguistiques, Cluj—Bucureşti 1937.
[23] SAVIN, E.—ABAGER, B.—ROMAN, A.: Gramatica practică a limbii germane. Bucureşti 1968.
[24] TOMPA, J.: Leiró magyar nyelvtan. Budapest 1964.

La structure algébrique des adverbes des langues romanes

EMESE KIS, CLUJ

Cet exposé veut présenter une restriction de la théorie générale du langage en tant que structure algébrique dans le sens donné par S. Marcus [18, 17] et L. Kalmár [15], restriction valable pour les langues naturelles et en corrélation avec la typologie déductive proposée par V. Skalička [29, 16]. Le but est de mettre en évidence une microstructure algébrique spécifique pour la morphologie des dix langues néolatines: 1. le roumain, 2. le dalmate, 3. l'italien, 4. le sarde, 5. les dialectes rhétoromans, 6. le français, 7. le provençal, 8. le catalan, 9. l'espagnol, 10. le portugais.

Nous considérons le hexaplet $(M_i, K_i, \Psi_i, R_i, \Pi_i, \Phi_i)$, où M_i est l'ensemble des mots $m_i \in M_i$ de ces dix langues: l'indice $i = \overline{1,10}$ correspond à chaque langue mentionnée. K_i est l'ensemble des contextes $k_i \in K_i$ qui peuvent être attachés à chaque mot $m_i \in M_i$ à l'aide des applications (fonctions) ψ_i, $\psi_i \in \Psi_i$, Ψ_i constituant l'ensemble de ces applications. Convenons à distinguer deux sous-ensembles de l'ensemble K_i des contextes $K_i : K_{1i}$ et K_{2i}. Soient deux mots m_i', et $m_i'' \in M_i$. Si un contexte k_{1i} est compatible soit avec m_i', soit avec m_i'', on dit $k_{1i} \in K_{1i}$. Soit le contexte k_{2i}, si k_{2i} est compatible tant avec m_i' qu'avec m_i'', on a $k_{2i} \in K_{2i}$. Nous disons que m_i' est equivalent à m_i'' du point de vue contextuel (ou nous disons que m_i' est l'équivalent contextuel de m_i''), si et seulement si le rapport entre la somme K_1 de tous les éléments de K_{1i} et la somme K_2 de tous les éléments de l'ensemble K_{2i} attachés à m_i' et à m_i'' tend vers zéro: $\frac{K_1}{K_2} \to 0$. Cette équivalence contextuelle, notée par R_i, veut dire que m_i' et m_i'' peuvent se substituer réciproquement dans presque tous leurs contextes et si un autre mot m_i''' peut se substituer à m_i'', il peut se substituer également à m_i'. Dans ce qui suit le symbole „=" indique cette équivalence contextuelle R_i. Π_i est l'ensemble des parties du discours π_i, $\pi_i \in \Pi_i$ qui appartiennent à chaque langue donnée. Φ_i est l'ensemble des applications (fonctions) φ_i qui transposent M_i en Π_i.

Ce qui nous intéresse ici c'est la structure algébrique de sous-ensemble A_i, $A_i \in \in (M_i, K_i, \Psi_i, R_i, \Pi_i, \Phi_i)$. Le sous-ensemble A_i des adverbes des langues néolatines est muni d'un ensemble des opérations Ω_i: (A_i, Ω_i). Toutes les opérations de Ω définies en A sont partielles. Nous y distinguons une opération interne: la juxta-

position, notée par „o", et trois opérations externes: l'enclise „①", la proclise „②"; l'exoclise ou la comparaison „③", donc nous avons la structure algébrique (A, o, ①, ②, ③).

1. La juxtaposition

L'ensemble des adverbes A_i est muni de l'opération de la juxtaposition (A_i, o) qui possède les propriétés suivantes:

1. Soit $a_i, b_i \in A_i$ et l'opération de la juxtaposition „o"; si a_i o $b_i = c_i$, nous aurons $c_i \in A_i$.

2. De même: soit $a_i \in A_i$ et si a_i o $a_i = a_i'$ il y en a $a_i \in A_i$.

3. Si on a i = o, c'est-à-dire dans la structure (A_0, Ω_0) pour quelques a_0^*, $a_0^* \in A_0$ il existe au moins un élément e_0, $e_0 \in A_0$ qui jouit de la propriété que l'équation a_0^* o $e_0 = a_0^*$ a au moins une solution. Si on a $i = 1$, c'est-à-dire dans la structure (A_1, Ω_1), pour quelques a_1^*, $a_1^* \in A_1$ il existe au moins un élément e_1, $e_1 \in A_1$ qui s'enjouisse de la propriété que l'équation a_1^* o $e = a_1^*$ a au moins une solution.

4. Dans les conditions $a_i, b_i \in A_i$ et a_i o $b_i \in A_i$, b_i o $a_i \in A_i$, a_i o $b_i = b_i$ o a_i si et seulement si $a = b$.

Exemples:

1. $a_i, b_i \in A_i$, a_i o $b_i \in A_i$

$i = 0$ lat. *ubi* „là" + *primum* „pour la première fois" = *ubi primum* „dès que, aussitôt", *nunc* = „maintenant" + *iam* „déjà" = *nunc iam* „juste à ce moment".

$i = 1$ roum. *cînd* „quand" + *colo* „là" = *cînd colo* „d'autre part", „en temps que".

$i = 3$ it. *ben* + *volontieri* = *ben volontieri*
ben + *felice* = *ben felice* [31 : 241].

$i = 6$ fr. *bien* + *tôt* = *bientôt*; *avant* + *hier* = *avant-hier*; *après* + *demain* = *après demain*, *là-bas*, *là-haut* [1 : 373].

$i = 7$ provençal *avans-ier* [25 : 126].

$i = 8$ catalan *tant* „tant" + *poch* „moins" = *tampoch* „pas de tout; *res* „rien" + + *mes*, *pus* „moins" = *resmes*, *respus* „rien de plus" [7 : 126].

$i = 9$ span. *siempre* „toujours" + *jamás* „jamais" = *siempre jamás* „absolument, toujours" [21, III: 553].

2. Si $a_i \in A_i$ et a_i o $a_i = a_i'$, $a_i' \in A_i$

$i = 1$ roum. *mai* „encore" + *mai* = *mai-mai* „presque", *aproape* „proche" + *aproape* = *aproape-aproape* „approximativement, presque" [8: 309].

$i = 8$ cat. *xano* „peu" + *xano* = *xano-xano* „peu à peu" [7: 132].

3. Si $a_i^* \in A_i$, $e_i \in A_i$, on a a_i^* o $e_i = a_i^*$

si et seulement si

$$i = 0, \qquad i = 1.$$

$i = 0$ lat. *vix* „à peine" + *dum* „même" = *vixdum* „à peine"; *vixdum* + *etiam* „encore" = *vixdum etiam* „à peine"; c'est-à-dire $a_0^0 e_0' = a_0 = a_0^0 e_0^0 e_0'' = a_0$.

$i = 1$ roum. *ici* „ici" + roum. régional *şa* (lit. *aşa* „ainsi") = roum. rég. *ici-şa*; *colo* „là" + *şa* = roum. rég. *colo-şa* „là", *amu* „maintenant" = *amu -şa* „maintenant", *acu* „maintenant" = *acuşi* „maintenant".

Il est à observer que la relation entre les éléments de type a_i et e_i n'est pas une relation de sous-ordination contextuelle. Par exemple lat. *Illi montes vix aperiebantur* „ces montagnes apparaissaient à peine" est aussi correcte dans le latin que *illi montes dum aperiebantur* „ces montagnes étaient même apparus" ou *illi montes etiam aperiebantur* „ces montagnes apparaissaient aussi". Si la présence d'un terme en absence de l'autre conduisait à un énoncé incorrecte on pourrait parler d'une relation de sous-ordination. Il est à mentionner que bien que *vix*, *dum*, *etiam* ne sont pas des équivalents contextuels, toutefois *vix*, *vixdum*, *vixdum etiam* le sont. À cause de cela: *Illi montes vix aperiebantur* = *Illi montes vixdum aperiebantur* = *Illi montes vixdum etiam aperiebantur* „ces montagnes apparaissaient à peine".

4. Si $a_i, b_i \in A$, a o $b \in A$, b o $a \in A$

$$a_i \text{ o } b_i = b_i \text{ o } a_i \Leftrightarrow \text{a} = \text{b}$$

$i = 1$ roum. *mult aproape* „très proche"; *aproape mult* „presque beaucoup"; *mult-mult* „beaucoup beaucoup"; *aproape* — *aproape* „approximativement, tout proche".

2. L'enclise

Soit l'ensemble B_i des scalaires β_i, $\beta_i \in B_i$ et $B_i \in (K_i, \Psi_i, R_i, M_i, \Pi_i, \Phi_i)$. B est l'ensemble des suffixes et des postpositions dans la terminologie linguistique. Ce $\beta \in B$ est un élément enclitique qui assure la façon d'être ou d'agir [1: 367] du contenu informationnel de l'adverbe auquel il se rattache. Il munit d'assurance (prive d'assurance) ou intensifie (prive d'intensité) [26: 295], autrement dit, il modifie quantitativement l'information de l'adverbe auquel il est accouplé donnant aussi un adverbe. L'assurance et l'intensité pourriont être représentées comme les deux axes d'un système de référence d'un espace bidimensionnel. En effet ce β_i est un élément qui est sous-ordoné à l' a_i auquel il est attaché. Cela veut dire que a_i encadré dans son contexte k_i conduit à une entité correcte dans la langue donnée et (a_i ① β_i) situé dans le même contexte k_i est correct aussi, mais la présence de β_i en absence de a_i dans le même contexte k_i conduit à un énoncé incorrect. Comparée à la juxtaposition — une opération additive — l'enclise (on la note ①) est une opération de

composition à droite, elle est multiplicative [20, 24] si l'on considère l'équivalence contextuelle assez souvent de $a_i \circ a_i = a_i \textcircled{1} \beta_i$.

Ainsi les propriétés de l'enclise sont:

1. Si $a_i \in A_i$, $\beta_i \in B_i$ et on a $a_i \textcircled{1} \beta_i = a_i'$, $a' \in A_i$.

2. Si $a_i \in A_i$, $a_i \circ a_i = a_i'$ et $a_i' \in A_i$ et $\beta_i \in B_i$, on a $a_i \circ a_i = a_i' = a_i \textcircled{1} \beta_i$ et $a_i \textcircled{1} \beta_i \in A_i$.

3. Considérons (A_0, Ω_0), $a_0^* \in A_0$, $\eta_0 \in A_0$. Il y a quelques a^*, pour lesquels existe au moins un élément η_0 qui jouit de la propriété que l'équation: $a_1^* \textcircled{1} \eta_0 = a_0^*$ a au moins une solution. De même en (A_1, Ω_1), $\eta_1 \in A_1$. Il y a quelques $a_1^* \in A_1$ pour lesquels existe au moins un élément η_1 qui jouit de la propriété que l'équation $a_1^* \textcircled{1} \eta_1 = a_1^*$ a au moins une solution.

Exemples

1. Si $a_i \in A_i$, $\beta_i \in B_i$ et $a_i \textcircled{1} \beta_i = a_i' \Leftrightarrow a_i' \in A_i$

$i = 0$ lat. *dulcis* + *issime* = *dulcissime* „le plus doucement", *rarissime* „le plus rarement".

$i = 1$ roum. *bine* „bien" + *işor* = *binişor* „presque bien"; *departe* „loin" + *ior* = = *depărtiş* „un peu loin"; *legal* „légalement" + *ment* = *legalmente* " „suivant absolument les lois"; *unde* „où", *cum* „comment", *cînd* „quand", + *-va* = = *undeva* „quelque part", *cumva* „qui sait comment", *cîndva* „une fois".

$i = 3$ it. *sicuro* „sûrement" + *-mente* = *sicuramente* „absolument sûr"; *forte* + *-mente* = *fortamente* [31: 240—1].

$i = 6$ fr. *comme, comment, quasi, quasiment* [1: 367—8].

$i = 8$ cat. *cert* „sûrement" + *-ment* = *certament* [7: 126].

$i = 9$ sp. *cerca* „proche" + *-ita* = *cerquita* „assez proche", *lejos* „éloigné" + *-itos* = *lejitos* „assez éloigné", *mucho* „beaucoup" + *-azo* = *muchazo* „beaucoup au sens péjoratif" [14: 138].

2. Si $a_i \in A_i$, $\beta_i \in B_i$, $(a_i \circ a_i) \in A_i (a_i \textcircled{1} \beta_i) \in A_i$

on a

$$a_i \circ a_i = a_i \textcircled{1} \beta_i = a_i' \text{ et } a_i' \in A_i$$

$i = 3$ it. *certo-certo* = *certamente* „absolument sûr", *chiaro-chiaro* = *chiaramente* „très évidemment",

3. Soit $a_i \in A_i$, $\eta_i \in a_i$ et on a $a_i \textcircled{1} \eta_i = a_i$ et $i = 0$, $i = 1$

$i = 0$ lat. *illuc* „là" + *que* = *illoque* „là", *dum* „même" + *que* = *dumque* „même", *nunc* „au présent, maintenant" *nunquam* „jamais" [33, 21 III: 552].

$i = 1$ L'élément unité envers l'enclise est en roum *-a* une entité déictique dans les parlers régionaux il existe même *a* = *-le*, = *-lea* qui ne sont pas commutatives:

aci „ici" + *-a* = *acia* „ici"; *aci* + *-le* = *acile* „ici"; *aci* + *-lea* = *acilea* „ici" [8: 301], *altminteri* „d'autre part" = *altminterea* „d'autre part".

3. La proclise

Pour doter cet ensemble A_i des adverbes romanes avec une deuxième opération externe on accepte d'abord le terme et la notion de la préposition délimitative. Soit l'ensemble C_i des scalaires γ_i, $\gamma_i \in C_i$, $C_i \in (K_i, \Psi_{i\,i}, R_i, M_i, \Pi_i, \Phi_i)$. C_i contient des prépositions délimitatives [9, 30: 209, 27: 26] qui bornent l'information fondamentale fournie par a_i [26: 206—23, 3: 456]. Cette délimitation peut être statique [11: 321, 21 III: 486—7] (inessive, subessive, superessive, adessive) ou dynamique [10, 13, 28, 32, 26]: centripète (illative, sublative, allative, translative, instrumentale, terminale), centrifuge (ellative, délative, ablative, causale). En tous cas elle permet d'être représentée dans un espace tridimensionnel [26: 313—330] par des voisinages sphériques ou cubiques, où chaque scalaire agit plutôt en direction de l'un des trois axes soit parallèlement soit en coïncidence avec eux.

Il n'est pas sans intérêt que la délimitation centrifuge s'exprime analytiquement [12]: à l'aide d'un scalaire centripète précédé d'un scalaire séparatif ou privatif: scalaire élatif = scalaire séparatif + scalaire illatif; scalaire délatif = scalaire séparatif + scalaire allatif. Cette expression analytique et la nécéssité de préciser la position simultanément envers tous les trois axes de l'espace tridimensionnel explique le cumul des scalaires proclitiques. La concaténation proclitique de m scalaires est possible à gauche d'un a_i dans les langues romanes où m varie selon le spécifique de la langue donnée, et $\gamma_i^1, \gamma_i^2, \ldots, \gamma_i^m \in C_i$, $a_i \in A_i$.

Exemples

$i = 1$, $m = 3$ roum. *asupra* „au-dessus", *de-asupra* „au-dessus" sublatif, *de deasupra* „au-dessus" délatif; *afară* „hors, dehors", *în afară* illatif, *din afară* ellatif, *de din afară* ellatif + délatif.

$i = 3$, $m = 3$ it. *giù, laggiù, per lagiù*; *dietro, indietro, all'indietro*; même $i = 3$, $m = 4$, parce que: *dall' indietro*.

Il est à observer que les dialectes rhétoromans excellent dans le cumul des éléments proclitiques: $i = 5$, $m = 7$: „Lorsqu'un endroit se trouve derrière une éminence, on indique d'abord la direction jusqu'à cette éminence, et puis la direction à partir de cette éminence: *keu — vi — ed — or e — si — sum devor — vi — ed — or ed d'u* „là de l'autre côté dessus tout en haut derrière dehors en bas" [21, III: 538].

$i = 6$, $m = 3$ fr. *avant, par avant, auparavant.*

$i = 7$, $m = 3$ prov. *ounte, mounte* „où, là", *amount, paramount.*

$i = 8$, $m = 2$ cat. *arrera, darrera* „derrière", *en arrera* „en derrière".

Ce scalaire $\gamma_i \in C_i$ semblablement au scalaire $\beta_i \in B_i$ est un élément sous-ordiné à l' a_i auquel il est attaché. Cela signifie que a_i, $a_i \in A_i$, encadré dans son contexte k_i conduit à une entité correcte dans la langue romane respective, $\gamma_i \textcircled{2} a_i$ situé dans le même contexte k_i conduit à une entité correcte aussi: $(\gamma_i \textcircled{2} a_i) \in A_i$; mais la présence de γ_i en absence de a_i dans le même contexte k_i conduit à un énoncé incorrect.

Nous considérons la proclise en A_i comme une opération externe de composition à gauche notée par $\textcircled{2}$, ayant les propriétés suivantes:

1. Soit $a_i \in A_i$ et $\gamma_i \in C_i$ et l'opération „$\textcircled{2}$" de composition à gauche. S'il y a $\gamma_i \textcircled{2} a_i = b_i$ on a $b_i \in A_i$.

2. Parmi les éléments de C_i il existe des scalaires comme sont les corréspondants de la préposition lat. *de* dans quelques langues romanes, nous le notons γ_i^*, qui jouit de la propriété de l'idempotence: Soit $a_i \in A_i$, $\gamma_i^* \in C_i$; si $(\gamma^* \textcircled{2} a_i) \in A_i$

$$a_i \neq \gamma^* \textcircled{2} a_i, \text{ mais } \gamma_i^* \textcircled{2} a_i = \gamma_i^* \textcircled{2} \gamma_i^* \textcircled{2} a_i = \\ = \gamma_i^* \textcircled{2} \underbrace{\ldots\ldots}_{n \text{ fois}} \textcircled{2} \gamma_i^* \textcircled{2} a_i.$$

On peut noter aussi $\gamma_i^* \textcircled{2} a_i = \ldots = \gamma_i^{*n-1} \textcircled{2} a_i = \gamma^{*n} \textcircled{2} a_i$. Les scalaires γ^* sont les éléments idempotents d'un groupoïde partiel. L'ordre n de ces scalaires $\gamma_i^* \in C_i$ ne dépasse pas $n = 3$ dans les langues néolatines.

3. En opposition avec la propriété 2 on peut mettre en évidence l'existence d'un élément ε_i. Soit $\varepsilon_i \in C_i$, $a_i \in A_i$.

Si $(\varepsilon_i \textcircled{2} a_i) \in A_i$ on a $\varepsilon_i \textcircled{2} a_i = a_i$ ce scalaire de type ε_i n'est pas idempotent.

4. Soit l'élément $\varepsilon_i \in C_i$, les scalaires $\gamma_i', \gamma_i'' \in C_i$ et $\gamma_i' \neq \gamma_i''$, a_i', $a_i'' \in A_i$, $a' \neq a''$. Si $((\varepsilon_i \textcircled{2} \gamma_i') \textcircled{2} a_i') \in A_i$, on y a $((\gamma_i' \textcircled{2} \varepsilon_i) \textcircled{2} a_i') \notin A_i$. Mais si $((\gamma_i'' \textcircled{2} \varepsilon_i) \textcircled{1} a_i'') \in A_i$, nous avons $((\varepsilon_i \textcircled{2} \gamma_i'') \textcircled{2} a_i'' \notin A_i$.

Exemples

1. $i = 0$ lat. *post* „après" + *cras* „demain" = *postcras* „lendemain",
   $i = 1$ roum. *mîine* „demain", *poimîine* „lendemain";
   $i = 3$ it. v. *poscrai*, napolit. *pescraye*;
   $i = 5$ engad. *puschmaun* „lendemain" [22].
2. $i = 1$, $n = 3$ roum. *de de departe* = *de departe* „de loin", de „de" + parte „côté" [3: 62—5];
   $i = 6$, $n = 3$ fr. *de dedans* = *dedans* = *de* + *dans* lat. *de intus* [6: 151—3];
   $i = 7$, $n = 2$ prov. *de dinz*.
3. $i = 1$ roum. *a* + *ici* „ici" = *aici* „ici" = *ici*, *a* + *colo* „là" = *acolo* „là" = *colo*; *a* + *şa* lat. „sic" = *aşa* „sic" — *şa*, roum. reg. *păi* = *apăi* „puris";
   $i = 3$ it. *qui*, *aqui*;

   $i = 5$ engad. *maun* „demain" *demaun* „demain" [22];
   $i = 6$ fr. *ci*, *ici*;
   $i = 7$ prov. *ounte*, *mounte* „où, là";
   $i = 8$ cat. *arrera*, *darrera* „derrière" [7].
4. $i = 1$ roum. *degeaba* „en vain" *pe degeaba* „en vain" $\in A_1$, mais **de pe geaba* $\notin A_1$; *gratis* „sans rémunération" *pe de gratis* „sans rémunération" $\in A_1$, mais **de pe gratis* $\notin A_1$ [3: 62—4];
   $i = 10$ port. *diante*, *perante*, *perdante* „en avant" $\in A_{10}$ tandis que **de per ante* $\notin A_{10}$ [23: 356].

4. L'exoclise ou la comparaison

Un sous-ensemble $\mathscr{A}_i$ des adverbes $A_i \subset \mathscr{A}_i$, peut jouer aussi le rôle des scalaires multiplicatifs, un rôle similaire à celle des scalaires enclitiques $\beta_i \in B_i$. Il s'agit des adverbes de comparaison α_i, $\alpha_i \varepsilon \mathscr{A}_i$ qui modifient l'intensité des adverbes auxquels ils sont attachés.

Considérons l'ensemble A_i muni d'une opération externe à gauche notée par $\textcircled{3}$, nous avons $(A_i, \textcircled{3})$. Cette opération est définie sur l'ensemble $\mathscr{A}_i$ des scalaires et est nommée comparaison. On peut la nommer aussi exoclise parce que du point de vue syntaxique ces adverbes-scalaires peuvent être considérés comme des compléments de l'adverbe et du point de vue morphologique comme des entités exocentriques [16] qui ne participent pas à la structure endocentrique morphologique de l'adverbe déterminé. Un scalaire α_i, $\alpha_i \in \mathscr{A}_i$, est un élément sous-ordoné a l' a_i, $a_i \in A_i$ auquel il est attaché. Si a_i encadré dans son contexte k_i conduit à une entité correcte dans une langue romane, on y a $(\alpha_i \textcircled{3} a_i) \in A_i$ et $(\alpha_i \textcircled{3} a_i)$ situé dans le même contexte k_i conduit à une entité correcte aussi. Mais la présence de α_i dans le contexte k_i en absence de a_i dans le même contexte k_i conduit à un énoncé incorrect.

Les propriétés de l'exoclise (de le comparaison) des adverbes sont les suivantes:

1. Soit $\alpha_i \in \mathscr{A}$, $a_i \in A_i$ et l' opération „$\textcircled{3}$". Si l' $(\alpha_i \textcircled{3} a_i)$ existe, on y a $\alpha_i \textcircled{3} a_i = a_i'$ et $a_i' \in A_i$.

2. Soit $\alpha_i \in \mathscr{A}_i$, $a_i \in A_i$, $\beta_i \in B_i$ et les opérations „o", „$\textcircled{2}$", „$\textcircled{3}$". Si $(a_i \text{ o } a_i) \in A_i$, $(a_i \textcircled{1} \beta_i) \in A_i$, $(\alpha_i \textcircled{3} a_i) \in A_i$ nous avons $a_i \text{ o } a_i = a_i \textcircled{2} \beta_i = \alpha_i \textcircled{3} a_i = a_i'$, $a_i' \in A_i$.

3. Soit le scalaire α_i, $\alpha_i \in \mathscr{A}_i$ et un scalaire γ_i, $\gamma_i \in C_i$. Si $\alpha_i \textcircled{3} \gamma_i \textcircled{2} a_i = b_i$, $b_i \in A_i$, mais $(\gamma_i \textcircled{2} \alpha_i \textcircled{3} a_i) \notin A_i$.

Pour les exemplifier il suffit de parcourir n'importe quelle grammaire d'une langue romane et voir [6, 3, 2, 19].

5. Conclusions

Cette approximation à l'aide d'un modèle algébrique permet une comparaison et une meilleure compréhension des systèmes des adverbes romans et offre la possibilité de les réduire et de les confronter à d'autres systèmes similaires.

1. La structure algébrique des adverbes romans A_i se distingue du point de vue typologique de celle des langues nonindoeuropéennes par le caractère „à droite" de l'enclise, „à gauche" de l'exoclise et l'organisation bidimensionnelle des scalaires enclitiques et exoclitiques.

2. Ce qui délimite les adverbes romans de ceux du type germanique par exemple, c'est le caractère „à gauche" de la proclise et l'organisation de ces scalaires proclitiques dans un espace tridimensionnel.

3. À la différence des langues romanes examinées dans cet exposé, le latin (A_0, Ω_0) et le roumain (A_1, Ω_1) se caractérisent par l'existence des éléments du type e_0, e_1 par rapport à l'opération de juxtaposition et des autres du type η_0, η_s par rapport à l'opération d'enclise.

4. Si l'on considère deux structures algébriques fixées de type „in-put" et du type „out-put" on peut formuler à l'aide des propriétés algébriques ce qui c'était passé avec les formes non-attestées des éléments, au cours de l'évolution des langues naturelles.

BIBLIOGRAPHIE

[1] CHEVALIER, J. C.—BLANCHE-BENVENISTE, C.—ARRIVÉ, M.—PEYTARD, J.: Grammaire Larousse du français contemporain. Paris 1964.

[2] CIOBANU, F.: Unele aspecte ale corespondenței dintre elementele prepoziționale și cele conjuncționale cu referire specială la locuțiuni. În: Studii de gramatică, III, 1961, p. 67—68.

[3] CIOBANU, F.: Valorile prepozițiilor în construcție cu adverbele. In: Studii de gramatică, III, 1961, p. 43—66.

[4] COUSIN, J.: Évolution et structure de la langue latine. Paris 1944.

[5] ERNOUT, A.—THOMAS, F.: Syntaxe latine. Paris 1953.

[6] FAHLIN, C.: Étude sur l'emploi des prépositions *en*, *a*, au sens local. Uppsala-Leipzig-Haag-Cambridge 1942.

[7] FRISONI, G.: Grammatica, esercizi pratici e dizionario della lingua catalana. Milano 1912.

[8] Gramatica limbii române. Vol. 1. ed. a II-a revăzută și adăugită. București 1963.

[9] GRAUR, Al.: *Ab*, *ad*, *apud* et *cum* en latin de Gaule. In: Bulletin de la Societé de Linguistique. Vol. 33, Paris 1932, p. 1—76.

[10] GRAUR, Al.: Studii de lingvistică generală. Variantă nouă. București 1960.

[11] HAMP, K.: Die zusammengesetzten Präpositionen für lateinische Lexicographie und Grammatik. Vol. 5, Leipzig 1888.

[12] HUGHES, J. P.: The science of language. An introduction to linguistics. New York, ed. III-a, 1963.

[13] IORDAN, I.: Note sintactice. 1. Adverbe de loc du sens temporal. In: Studii și cercetări lingvistice, 1950, p. 269—274.

[14] IORDAN, I.—DUHĂNEANU C.: Curs de gramatica limbii spaniole. București 1963.

[15] KALMÁR, L.: Le langage comme structure algébrique. In: Cahiers de linguistique théorique et appliquée, IV, 1967, p. 73—82.

[16] KIS, E.: Comparaison typologique de deux langues non apparentées du point de vue génético-structural. In: Theoretical problems of typology and the northern eurasian languages, 1969,p. 107—110. KIS, E.: The variability of typological features in the evolution of Romanian. In: Résumés des communications — X^ème congrès international des linguistes, Bucharest, 28 août — 2 septembre 1967, p. 185.

[17] MARCUS, S.: Algebraic linguistics. New York 1967.

[18] MARCUS, S.: Typologie des langues et modèles logiques. In: Acta math. Academiae Scientiarum Hungaricae, XIV, 1963, 269—281.

[19] MELANDER, J.: Études sur *magis* et les adversatives dans les langues romanes. 1916.

[20] MOISIL, GR. C.: Introducere în algebră. I. Inele și ideale. Bucarest 1954.

[21] MEYER—LÜBKE: Grammaire des langues romanes. Traduction française par Auguste Doutrepont et G. Doutrepont. Paris 1895—1906. Vol. II. Morphologie, 1895. Vol. III. Syntaxe, 1900.

[22] MEYER—LÜBKE: Romanisches etymologisches Wörterbuch. Heidelberg, ed. III-a 1935.

[23] NUNES, J. J.: Compêndio de Gramática Histórica Portuguesa. Fonética e morfologia. Lisboa. ed. II-a, 1930.

[24] PIC, Gh.: Algebră superioară. București 1966.

[25] PORTAL, E.: Grammatica provenzale (lingua moderna) e dizionarietto provenzale-italiano. Milano 1914.

[26] POTTIER, B.: Systématique des éléments de relations. Étude de morphosyntaxe structurale romane. Paris 1962.

[27] PUȘCARIU, S.: Limba română. Vol. I. Bucarest, Partea generală 1940.

[28] SÄVBORG, T.: Étude sur le rôle de la préposition *de* dans les expressions de lieu relatives. 1941.

[29] SKALIČKA, Vl.: Ein typologisches Konstrukt. In: Travaux linguistiques de Prague, II. Prague 1966, pp. 157—163.

[30] TIKTIN, H.: Gramatica română, I. Etimologia. Iași 1893.

[31] TRABALZA, C.—ALDOLI, E.: La grammatica dell'italiano. Firenze, ed. a II-a, 1934.

[32] VASILIU, L.: Schiță de sistem al propozițiilor limbii române. În: Studii de gramatică, III, 1961, pp. 11—42.

[33] WEIGAND, G.: In: Jahresbericht des Instituts für rumänische Sprache zu Leipzig, XII, 1965.

Wortformenklassensysteme und ihre Optimierung

JÜRGEN KUNZE, WALTER PRIESS, BERLIN

Es sei A eine beliebige endliche, nicht-leere Menge von unzerlegbaren Zeichen, die als Wortformen interpretiert werden. Mit $W(A)$ bezeichnen wir die von A erzeugte freie Halbgruppe. Eine beliebige nicht-leere Teilmenge L von $W(A)$ heißt eine Sprache über A.

Ist L eine Sprache, so nennen wir ein System Δ von Teilmengen von A (d.h. eine Teilmenge der Potenzmenge von A) ein *distributives Wortformenklassensystem für L*, falls Δ die folgenden beiden Bedingungen erfüllt:

H_5: Zu jeder Kette $\varphi \in L$ gibt es eine Folge $X_1, \ldots, X_n$ von Mengen aus Δ, so daß

$$\varphi \in \Pi(X_1, \ldots, X_n) \subseteq L$$

gilt. (Π bezeichnet dabei das Cartesische Produkt.)

H_6: Jede Menge $X \in \Delta$ ist abgeschlossen.

(Zur Rechtfertigung von H_5 und H_6 vgl. Versuch eines objektivierten Grammatikmodells II, Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung *21*, 1968, S. 421—466, insbesondere die Seiten 445—454, die Definition des Begriffs „abgeschlossen" wird hier weiter unten gegeben.)

Wir erklären zunächst noch einige weitere Begriffe, bevor wir zu den eigentlichen Betrachtungen übergehen.

Mit $\alpha(L, a)$ bezeichnen wir die Menge aller Kontexte der Wortform $a \in A$ in der Sprache L. Wir beschränken uns hier auf nichtparasitäre Wortformen a, d. h. auf solche, für die $\alpha(L, a) \neq \varnothing$ ist. Eine Wortform a heißt ein *reines Homonym in L*, falls

$$\alpha(L, a) = \bigcup_{x \in A} \alpha(L, x)$$

$$\alpha(L, x) \subset \alpha(L, a)$$

gilt, andernfalls (nach Dobrušin) eine *Wurzel in L*.

Ferner sei für beliebige Wortformenmengen X mit $\varnothing \subset X \subseteqq A$

$$\alpha_1(L, X) = \text{Def} \bigcap_{x \in X} \alpha(L, x),$$

$$\alpha_2(L, X) = \text{Def} \bigcup_{x \in X} \alpha(L, x).$$

Mit $P(L)$ bezeichnen wir die Menge aller Wurzeln in L. $P(L)$ ist offenbar eine Teilmenge der Wortformenmenge A, die wir im folgenden der Deutlichkeit halber mit $A(L)$ bezeichnen wollen. Man kann leicht zeigen, daß stets $P(L) \neq \varnothing$ gilt. Für beliebige Mengen X von Wortformen sei $P(L, X)$ die Menge der in X enthaltenen Wurzeln in L:

$$P(L, X) = \text{Def } X \cap P(L).$$

Sei $X \subseteqq A(L)$ eine beliebige Menge von Wortformen. Die *Hülle* $H(L, X)$ von X in L wird folgendermaßen definiert:

$$H(L, X) = \text{Def } \{y \mid y \in A \text{ und } \alpha(L, y) \supseteqq \alpha_1(L, X)\}.$$

Die Operation H besitzt (bei festem L) die folgenden Eigenschaften:

$$X \subseteqq H(L, X), \tag{1}$$

$$H(L, H(L, X)) = H(L, X), \tag{2}$$

$$\text{wenn } X \subseteqq Y, \text{ so } H(L, X) \subseteqq H(L, Y). \tag{3}$$

Eine Wortformenmenge heißt *abgeschlossen in* L, falls $H(L, X) = X$ gilt.

Für eine beliebige vorgegebene Sprache gibt es im allgemeinen mehrere Systeme Δ, die H_5 und H_6 erfüllen, und es erweist sich in vielen Fällen als wünschenswert, mittels geeigneter Verfahren aus diesen die „besseren" auszusondern. Das kann einmal geschehen, indem man an die Wortformenklassensysteme zusätzliche, in jeder Sprache erfüllbare sowie untereinander (und natürlich mit H_5 und H_6) verträgliche Forderungen stellt. Eine andere Möglichkeit besteht darin, geeignete Umwandlungsverfahren auf Klassensysteme anzuwenden, die zu jedem Wortformenklassensystem Δ durch effektive Konstruktion ein Klassensystem $\omega(\Delta)$ oder eine Menge von Systemen $\Omega(\Delta)$ liefern. Dabei sind $\omega(\Delta)$, bzw. die Systeme aus $\Omega(\Delta)$ in einer heuristisch begründbaren Hinsicht optimaler als Δ selbst, sofern nicht schon $\Delta = \omega(\Delta)$, bzw. $\Delta \in \Omega(\Delta)$ gilt. Durch die Bedingungen $\Delta = \omega(\Delta)$ bzw. $\Delta \in \Omega(\Delta)$ erreicht man also wieder eine Auswahl aus der Vielfalt der möglichen Wortformenklassensysteme, wobei natürlich nachgewiesen werden muß, daß es Klassensysteme Δ gibt, die diese Einschränkungen erfüllen.

Wir wollen nun zwei solche Optimierungsverfahren für Wortformenklassensysteme betrachten. Das erste beruht auf der linguistisch motivierbaren Erfahrungstatsache, daß die reinen Homonyme in L für die Struktur der Sprache L unwesentlich sind. (Wie oben schon bemerkt wurde, kann $A(L)$ niemals nur reine Homonyme enthalten!) Diese Erfahrungstatsache läßt sich formal-distributiv so erfassen:

Ist L eine Sprache und $P(L)$ die Menge der Wurzeln in L, so sei

$$p(L) = \text{Def } L \cap W(P(L)).$$

$p(L)$ besteht also aus genau den Ketten von L, in denen kein reines Homonym vorkommt. Es läßt sich beweisen, daß für die Operation p neben der trivialen Selbstinklusion ($p(L) \subseteqq L$) auch die Idempotenz gilt: Für beliebige Sprachen L ist

$$p(p(L)) = p(L).$$

Die folgenden Sätze liefern die Rechtfertigung für das anschließend beschriebene Optimierungsverfahren:

Satz 1. Für beliebige Sprachen L gilt:
Ist Δ_p ein Wortformenklassensystem für $p(L)$, so ist

$$\overline{\omega}_p(\Delta_p) = \{H(L, X) \mid X \in \Delta_p\}$$

ein Wortformenklassensystem für L.

Satz 2. Für beliebige Sprachen L gilt:
Ist Δ ein Wortformenklassensystem für L, so ist

$$\overline{\omega}_p(\Delta) = \{H(p(L), P(L, X)) \mid X \in \Delta\}$$

ein Wortformenklassensystem für $p(L)$.

Aus diesen beiden Sätzen ergibt sich leicht die folgende Aussage: Für beliebige Sprachen gilt: Ist Δ ein Wortformenklassensystem für L, so ist

$$\omega_p(\overline{\omega}_p(\Delta)) = \{H(L, H(p(L), P(L, X))) \mid X \in \Delta\}$$

ebenfalls ein Wortformenklassensystem für L.

Man kann sich davon überzeugen, daß

$$\omega_p(\overline{\omega}_p(\Delta)) \neq \Delta$$

durchaus möglich ist. Es gilt jedoch der folgende

Satz 3. In jeder Sprache L existiert wenigstens ein Wortformenklassensystem Δ, bei dem für alle $X \in \Delta$

$$X = H(L, H(p(L), P(L, X)))$$

gilt (und damit natürlich auch $\omega_p(\overline{\omega}_p(\Delta)) = \Delta$).

Die Gültigkeit dieses Satzes ergibt sich aus der Tatsache, daß die Operation $D_1(L, X) = \text{Def } H(L, H(p(L), P(L, X)))$ idempotent ist, d. h., es gilt

$$D_1(L, D_1(L, X)) = D_1(L, X).$$

Die Bedingung

$$D_1(L, X) = X \text{ für alle } X \in \Delta$$

ist aus den genannten Gründen somit als Optimierungsbedingung geeignet, während der Übergang von Δ zu $\omega_p^1(\Delta)$, wobei

$$\omega_p^1(\Delta) = \text{Def}\ \{D_1(L, X) \mid X \in \Delta\}\ (= \omega_p(\overline{\omega}_p(\Delta)))$$

ist, ein idempotentes Optimierungsverfahren darstellt. Man kann dieses Verfahren folgendermaßen interpretieren: Zu jeder Klasse $X \in \Delta$ bildet man zunächst $X^* = H(p(L), P(L, X)$. Dies ist die der Klasse X in natürlicher Weise entsprechende Klasse in $p(L)$. Sie enthält keine reinen Homonyme. Durch die Operation $H(L, X^*)$ kommen im allgemeinen zwar wieder gewisse reine Homonyme hinzu, aber dies geschieht durch die Hüllenbildung gewissermaßen „automatisch" in dem Sinne, daß $H(L, X^*)$ durch X^* bereits eindeutig bestimmt ist.

Die Bedingung $\omega_p^1(\Delta) = \Delta$ ist die formal-distributive Präzisierung der Forderung, daß die reinen Homonyme für die Struktur von L (und damit für die Struktur der Klassen aus Δ) unwesentlich sind. Auf Beispiele und auf ausführliche linguistische Motivierungen müssen wir hier verzichten.

Die Erklärung der zweiten Optimierungsbedingung erfordert einige weitere Definitionen:

Ist $x \in A(L)$ eine nicht-parasitäre Wortform, so sei die *Fundamentalmenge* $\mathfrak{M}(L, x)$ von x in L die folgende Kontextmenge:

$$\mathfrak{M}(L, x) = \text{Def}\ \alpha(L, x) \setminus \bigcup_{\alpha(L, y) \subset \alpha(L, x)} \alpha(L, y).$$

Aus der Definition der Wurzeln und der reinen Homonyme ergibt sich unmittelbar: $\mathfrak{M}(L, x) \neq \varnothing$ genau dann, wenn $x \in P(L)$. $\mathfrak{M}(L, x)$ enthält genau die Kontexte, in denen x „maximal eindeutig" ist. Wie üblich kürzen wir die Beziehung $\alpha(L, y) \subseteq \alpha(L, x)$ durch $y \underset{L}{\rightarrow} x$ ab. Für $\alpha(L, y) \subset \alpha(L, x)$ verwenden wir die Bezeichnung $y \underset{L}{\rightarrow} x$.

Die Sprache $u(L)$ wird folgendermaßen definiert: $u(L)$ ist die größte der Teilsprachen L^* von L, für die

$$\alpha(L^*, x) \subseteq \mathfrak{M}(L, x)$$

gilt. Es läßt sich zeigen, daß $u(L)$ durch diese Bedingung eindeutig bestimmt ist. Ferner gilt $u(L) \neq \varnothing$, falls $L \neq \varnothing$. Aus der Definition folgt außerdem, daß die reinen Homonyme in L in der Sprache $u(L)$ parasitäre Wortformen sind. Umgekehrt gilt jedoch auch $\alpha(u(L), x) \neq 0$ für alle Wurzeln in L. Damit haben wir: $\alpha(u(L), x) \neq 0$ genau dann, wenn $x \in P(L)$. Es ist jedoch zu bemerken, daß allgemein $\alpha(u(L), x) \subset \mathfrak{M}(L, x)$ ist.

Für $u(L)$ gelten folgende Aussagen:

Satz 4. (1) $u(p(L)) = u(L)$; $u(L) \subseteq p(L)$.

(2) $L \subseteq u(L) \subseteq u(u(L)) \subseteq \ldots$

(3) Für alle natürlichen Zahlen n und k gilt:
Wenn $u^{n+1}(L) = u^n(L)$, so $u^{n+k}(L) = u^n(L)$.

(4) Zu jeder Sprache L gibt es eine natürliche Zahl n mit $u^n(L) = u^{n+1}(L)$.

(5) Es ist $u(L) = L$ genau dann, wenn $A(L)$ nur initiale Wortformen a enthält, d. h. solche, zu denen kein $b \in A(L)$ mit $b \underset{L}{\rightarrow} a$ existiert.

Satz 5. Für jede Kette $\varphi = x_1 \ldots x_n \in L$ gilt:
Es ist $\varphi \in u(L)$ genau dann, wenn

$$x_1 \ldots, x_{\nu-1}{}^* x_{\nu+1} \ldots x_n \in \mathfrak{M}(L, x_\nu)$$

für alle ν mit $1 \leqq \nu \leqq n$.

Satz 6. Für jede Kette $\varphi = x_1 \ldots x_n \in L$ gilt:
Entweder ist $\varphi \in u(L)$ oder es gibt eine Kette $\psi = y_1 \ldots y_n \in u(L)$ mit $y_\nu \underset{L}{\rightarrow} x_\nu$ für alle ν mit $1 \leqq \nu \leqq n$, wobei für wenigstens ein ν sogar $y_\nu \underset{L}{\rightarrow} x$, gilt.

Der letzte Satz zeigt eine wesentliche Eigenschaft von $u(L)$: $u(L)$ ist eine Sprache (und in einem gewissen Sinne sogar die kleinste), bei der sich durch Anwendung der Substitution $\underset{L}{\rightarrow}$ auf die in ihr enthaltenen Ketten genau die in L enthaltenen Ketten ergeben. (Daß sich dabei keine Ketten ergeben können, die nicht in L liegen, folgt aus $u(L) \subseteq L$ und den allgemeinen Eigenschaften von $\underset{L}{\rightarrow}$.)

Analog dem Satz 1 für $p(L)$ gilt für $u(L)$:

Satz 7. Für beliebige Sprachen L gilt:
Ist Δ_u ein Wortformenklassensystem für $u(L)$, so ist

$$\omega_p(\Delta_u) = \{H(L, X) \mid X \in \Delta_u\}$$

ein Wortformenklassensystem für L.

Für den Übergang von Wortformenklassensystemen in L zu solchen in $u(L)$ konnte bisher keine für $u(L)$ spezifische Operation gefunden werden, wie sie Satz 2 für $p(L)$ angibt. Man hat daher ein entsprechendes Optimierungsverfahren etwas anders zu formulieren. Die Optimierungsbedingung, bei der $u(L)$ verwendet wird, läßt sich daher vorläufig nur so aussprechen: Es gibt ein Wortformenklassensystem Δ_u in $u(L)$, das die Bedingung

$$\Delta = \omega_p(\Delta_u)$$

erfüllt.

Es ist leicht zu sehen, daß es in jeder Sprache L ein Wortformenklassensystem Δ gibt, das diese Forderung erfüllt. Ferner gibt es Sprachen, in denen gewisse Wortformenklassensysteme dieser Bedingung nicht genügen. Somit ist diese Forderung nicht trivial.

Die beiden skizzierten Optimierungsbedingungen für Wortformenklassensysteme bezogen sich auf Teilsprachen von L (nämlich auf $p(L)$ und $u(L)$). Natürlich gibt es auch Kriterien, die auf anderen Prinzipien beruhen. Die wichtigste Frage bei allen derartigen Optimierungskriterien ist ihre Verträglichkeit, d. h. der Beweis dafür, daß sie in jeder Sprache von gewissen Wortformenklassensystemen gleichzeitig erfüllt werden. Derartige Betrachtungen sind jedoch recht kompliziert.

Die hier dargestellten Ergebnisse sind enthalten in den Teilen III und IV der schon zitierten Arbeit *Versuch eines objektivierten Grammatikmodells*, die ebenfalls in der Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung erscheinen werden.

Die grammatischen Konfigurationen im Modell der Abhängigkeitsgrammatik

GERDA KLIMONOW, BERLIN

Es hat sich als günstig erwiesen, zum Zweck der Automatisierung des Übersetzungsvorgangs ein Abhängigkeitsschema zur Darstellung der syntaktischen Beziehungen im Satz zu benutzen. Es ist bekannt, daß nach dem Modell der Abhängigkeitsgrammatik jede Texteinheit[1] eines Satzes (außer einer einzigen, die nicht untergeordnet wird) einer anderen Texteinheit desselben Satzes untergeordnet (bzw. schematisch durch einen Pfeil als von ihr abhängig dargestellt) wird. Auf diese Weise entsteht eine Art Baum, in dem jede Texteinheit nur von *einer* anderen Texteinheit abhängig sein darf und in dem es keine geschlossenen Wege geben darf. Die Nummer der übergeordneten Texteinheit im Satz wird Unterordnungsindex *(UI)* genannt. Des weiteren erhält jedes Paar von abhängiger und übergeordneter Texteinheit der Quellensprache eine Unterordnungscharakteristik *(UC)*, d. h. eine Charakteristik der Art der syntaktischen Beziehung zwischen dem jeweiligen Paar von Texteinheiten. Die *UC* sind die entscheidende Information für die syntaktische Synthese der Zielsprache, d. h. für die Herstellung der Reihenfolge ihrer Wörter im Satz.[2]

Die Bausteine eines Abhängigkeitsbaumes sind also die Texteinheiten (meist sind es zwei), die in sinngemäße und syntaktische Beziehungen miteinander treten (nach I. A. Mel'čuk отношения непосредственной доминации),[3] bilden grammatische Konfigurationen. Der Begriff *grammatische Konfiguration* (im folgenden einfach *Konfiguration* genannt) soll nun näher erläutert werden.

[1] *Texteinheit* wird nicht so verstanden wie im traditionellen Sinne *Wortform*, sondern dazu gehören z. B. auch die Satzpunktzeichen und solche Texteinheiten, die zusammen analytische Wortformen bilden.

[2] Über den Mechanismus der syntaktischen Synthese s. J. Kunze, *Zur syntaktischen Synthese*. Kybernetika, *1*, 1965, Nr. 1, S. 85—101.

[3] И. А. Мельчук, Автоматический синтаксический анализ. Кибернетика в монографиях. Новосибирск 1964, S. 18 ff.

In dem Sinne, wie wir diesen Begriff gebrauchen wollen, wurde er von T. N. Mološnaja geprägt: Die grammatische Konfiguration ist eine Kombination bestimmter Wortklassen, die in einer bestimmten Reihenfolge angeordnet sind und eine bestimmte grammatische Form haben.[4] Die Konfiguration $A_{G'K'N'} + N_{G'K'N'}$ kann z. B. die konkreten Formen высокой горе, большие дома, ночной сторож usw. haben, aber nicht z. B. равный площади, da hier nicht die Beziehung der Kongruenz vorliegt, sondern eine Rektion $A(K') + N_{K'}$.

Wir stimmen der Begründung von T. N. Mološnaja zu, warum statt *grammatische Konfiguration* nicht einer der üblichen Termini *Wortfügung* (словосочетание) oder *Syntagma* gewählt wurde: Eine grammatische Konfiguration unterscheidet sich von einer Wortfügung dadurch, daß sie nicht unbedingt eine Fügung von Wörtern sein muß (z. B. красив als Variante von быть красивым). Sie unterscheidet sich aber auch vom Syntagma, nämlich dadurch, daß sie nicht nur subordinative, sondern auch koordinative Beziehungen erfaßt (интегрирование и дифференцирование[5]). In der Definition von Mološnaja bleiben einige Fragen offen. Beginnen wir mit der Anzahl der Elemente einer Konfiguration. Mološnaja selbst schreibt, daß die Größe der Konfigurationen von der Betrachtungsweise abhängt. Z. B. könne man in Wortfolgen wie называть фигуру треугольником *eine* Konfiguration sehen ($V^{\text{называть/считать}} + N_{\text{косв. падеж}}$) oder zwei ($V^{\text{назыв./счит.}} + N_4$ und $V^{\text{назыв./счит.}} + + N_5$), je nach gröberer oder feinerer Unterteilung der Wortklassen. Und überhaupt sei das Moment der Willkür hier derart groß, daß es ausreiche, sämtliche grammatischen Beziehungen zwischen den Wörtern als identisch zu erklären (denn die grammatischen Beziehungen innerhalb aller konkreten, ein und dieselbe Konfiguration verkörpernden Wortfolgen müssen identisch sein), um zu dem Ergebnis zu gelangen, daß es in der Sprache nur eine einzige Konfiguration gibt.[6] Wir wollen versuchen, das Moment des Willkürlichen etwas einzuschränken, indem wir von der Folge der Elemente einer Konfiguration verlangen, daß sie von minimaler Länge ist. Unter „minimal" verstehen wir „aus zwei konfigurationsbildenden Texteinheiten bestehend und also nicht in zwei oder mehr Konfigurationen zerlegbar" (oder anders: „Es gibt zu der Menge der Elemente einer Konfiguration keine echte Untermenge"). An dieser Stelle muß bestimmt werden, welche Texteinheiten konfigurationsbildend wirken und welche nicht. Betrachten wir zur Illustration drei Beispielgruppen:

[4] Т. Н. Молошная, О понятии грамматической конфигурации. In: Структурно-типологические исследования, Москва 1962, S. 48.

[5] Vgl. T. N. Mološnaja, op. cit., S. 46.

[6] Ebenda, S. 48 f.

(1a) Он здесь;
(1b) Он был/будет там;
(2a) Он учитель;
(2b) Он был/будет учителем;
(3a) Он будет приходить;
(3b) Он был/будет послан;
(3c) Он пришел бы.

Es ist eine Sache der Absprache, welche Texteinheit als Spitze des Baumes betrachtet werden soll: Nach unserer Betrachtungsweise ist es immer das Satzpunktzeichen (Punkt, Semikolon, Fragezeichen, Ausrufungszeichen). Das Prädikat eines Satzes (bzw. Hauptsatzes) oder sein finiter Teil, falls es sich um ein zusammengesetztes Prädikat handelt, wäre als ein Element der der Spitze am nächsten liegenden Ebene dieser unterzuordnen. Im Text nicht explizit ausgedrückte Strukturbestandteile (so das Nullprädikat in (1a) und die Nullkopula in (2a)), müssen sichtbar gemacht werden, um als Knoten eines Baumes fungieren zu können. Das geschieht mit Hilfe der Konfigurationen. Die Beispiele (1a) bis (2b) bestehen jeweils aus zwei Konfigurationen:

(1a) Он здесь:
$Pp_{m1}\ \varnothing$ und $\varnothing\ Av_{\text{det/loc}}$;

(1b) Он был/будет там:
$Pp_{1m}V^b(\text{loc})_f$ und $V^b(\text{loc})_f\ Av_{\text{det/loc}}$;

(2a) Он учитель:
$Pp_{m1}\ \varnothing$ und $\varnothing\ N_{ms1}$;

(2b) Он был/будет учителем:
$Pp_{m1}\ V^b(5)_f$ und $V^b(5)_f\ N_{ms5}$.

Die Sätze unter (3) dagegen bestehen nur aus einer einzigen Konfiguration:

(3a) Он будет приходить:
$Pp_{m1}\quad V^b_f V_i$;

(3b) Он был/будет послан:
$Pp_{m1}\quad V^b_f$Apk;

(3c) Он пришел бы:
$Pp_{m1}\quad V_{v\text{konj}}$.

Die linguistische Argumentation des qualitativen Unterschiedes zwischen den Formen von быть in (1b), (2b) einerseits und (3a), (3b) andererseits liegt auf der Hand:

1. Быть in (3a)—(3c) hat rein grammatische und keinerlei semantische Bedeutung und ist durch keine andere Wortform zu ersetzen, während быть in (2a), (2b) kopulatives Verb (und durch andere kopulative Verben zu ersetzen) und somit Träger der allgemeinsten Semantik, nämlich der des Seins, ist. Быть in (1a), (1b) ist Vollverb und durch andere Vollverben substituierbar.

2. Die Prädikate in (3a)—(3c) sind (aus jeweils zwei Texteinheiten bestehend) analytische Wortformen und damit immer nur *ein* Element einer Konfiguration, während был учителем bzw. был там nicht etwa Vergangenheitsformen von учитель oder там sind; die konfigurationsbildende Wirkung der Formen von быть ist bei den Beispielen (1a)—(3c) bereits dargestellt. Andererseits können aber auch nichtkonfigurationsbildende Elemente in Konfigurationen zusätzlich (fakultativ) vorhanden sein, die explizit im Text ausgedrückt sind. Als solche fakultativen Elemente müssen z. B. die determinativen Adverbien behandelt werden, da sie sich semantisch und syntaktisch nicht auf eine konkrete Texteinheit, sondern auf den Satz als Ganzes beziehen. Sie befinden sich im Verhältnis der bedingten Unterordnung gegenüber bestimmten Texteinheiten (dem obersten Knoten des jeweiligen Prädikats). Beispiele:

(4a) *Поэтому* строить переводческий алгоритм путем установления соответствий между английскими структурами, выделенными Ч. Фрайсом, и структурами русского языка нецелесообразно (*Проблемы кибернетики*, вып. 3, стр. 210).

Der Konfiguration $M(i)\ V_i$ (нецелесообразно строить) ist das Adverb поэтому in bedingter Unterordnung angegliedert: $Av_{\text{det}}\ M(i)\ V_i$ (поэтому нецелесообразно строить).

(4b) *Везде*, где удается проследить генез автоматов, они оказываются прижизненно обретенными, т. е. ... (*Проблемы кибернетики*, вып. 6, стр. 147).

Hier ist das Adverb везде den Konfigurationen bedingt angegliedert, deren ein Element (оказывается) die Spitze des Baumes bildet:

$Av_{\text{det}}\ Pp_{1_p}\ V_f^{\text{kop.}}$ (везде они оказываются) und

$Av_{\text{det}}\ V_f^{\text{kop.}}\ Apl_5$ (везде оказываются приобретенными).

Weiterhin muß u. E. unbedingt folgender Aspekt in einer Definition der Konfiguration zum Ausdruck kommen: Die Wortklassencharakteristika der Elemente einer Konfiguration müssen eindeutig sein, da sonst die syntaktischen (subordinativen oder koordinativen) Beziehungen zwischen den Elementen ebenfalls nicht eindeutig[7] sind oder die konfigurationsmäßig zusammengehörigen Elemente überhaupt nicht aufgefunden werden können. Z. B. kann man bei der Folge $[Av/M(i)/Ak]\ V_i$ nicht von einer Konfiguration sprechen, wohl aber gibt es nach Anwendung einer Prozedur zur Lösung der Wortklassenmehrdeutigkeiten entweder die Konfiguration $M(i)\ V_i$ (интересно [было бы] преподнести ...) oder die Konfiguration $Av\ V_i$ ([нельзя] интересно преподнести). Wir erhalten also folgende Arbeitsdefinition der Konfiguration: *Eine grammatische Konfiguration ist eine minimale Folge eindeutiger Wortklassen, die in einer bestimmten Reihenfolge angeordnet und so fein in Unterklassen eingeteilt sind, daß sie syntaktisch eindeutige Beziehungen untereinander klar erkennen lassen.*

Die Elemente einer Konfiguration sind im Analysesatz nicht immer kontinuierlich angeordnet.

Eine Konfiguration kann frei (kontextfrei) oder gebunden (kontextgebunden) sein, je nachdem, ob ihre Elemente in Bezug auf die Wortklasse eindeutig sind und in der angegebenen Reihenfolge stehen oder nicht. Die oben genannten Konfigurationen $M(i)\ V_i$ und $Av\ V_i$ wären z. B. nicht als frei zu bezeichnen, da sie aus der Folge $[Av/M(i)/Ak]\ V_i$ ermittelt wurden. Die Kenntnis der freien Konfigurationen ist für uns von praktischem Wert: Bei jedem Vorkommen einer solchen Konfiguration werden mit Hilfe algorithmischer Regeln ohne zusätzliche Prüfungen sofort die entsprechenden Unterordnungen vorgenommen.

Die Elemente der syntaktischen Konfigurationen sind also auf keinen Fall mit den traditionellen Satzgliedern identisch, sie sind (in der überwiegenden Zahl der Fälle) eine feinere Einteilung (in einigen Fällen eine gröbere) der

[7] „Syntaktisch eindeutig" wird verstanden wie bei E. Agricola, *Syntaktische Mehrdeutigkeit (Polysyntaktizität) bei der Analyse des Deutschen und des Englischen*, Berlin 1968, S. 58 f. d. h. es gibt für jede Texteinheit nur eine mögliche Abhängigkeitsrelation und einen möglichen Abhängigkeitstyp.

Sätze in sinngemäß und syntaktisch miteinander in Beziehung stehende Texteinheiten und widerspiegeln den syntaktischen Aufbau des Satzes präziser als die Einteilung in Satzglieder.

Verzeichnis der verwendeten Symbole

A	— Adjektiv
$A(K')$	— Adjektiv, das einen bestimmten Kasus regiert
Ak	— Kurzform des Adjektivs
Apk	— Kurzform des Partizips
Apl	— Langform des Partizips
Av	— Adverb
Av_{det}	— Determinativadverb
$Av_{\text{det/loc}}$	— Adverb, das immer einer der semantischen Unterklassen „determinativ" oder „lokal" angehört
M	— Prädikativ
$M(i)$	— Prädikativ, das einen Infinitiv regiert
Pp	— Personalpronomen
V	— Verb
V_f	— finite Verbform
V_i	— Infinitiv
V^b	— das Verb быть
$V^{\text{kop.}}$	— kopulatives Verb
$V_{v\text{konj.}}$	— Verb im Konjunktiv
$\varnothing$	— Zero-Wortform

Indices rechts unten:

1,2,...,6	— Nominativ, Genitiv, ..., Präpositiv
s	— Singular
p	— Plural
m	— Maskulinum
f	— Femininum
n	— Neutrum
G'	— ein bestimmtes Genus
K'	— ein bestimmter Kasus
N'	— ein bestimmter Numerus
[ ]	— schließen Wortklassenmehrdeutigkeiten ein
/	— Disjunktion
⌊——↑	— Unterordnung innerhalb von Konfigurationen
⌊- - -↑	— bedingte Unterordnungen

Sur quelques problèmes d'analyse algébrique contextuelle

CONSTANTIN V. CRĂCIUN, BUCAREST

1. Introduction

Le but du présent article est l'étude, du point de vue formel, des opérateurs contextuels G, φ, ψ, λ et μ ([3], [12], [9], [10], [11], [13], [1], [14] et [2]). L'accent sera mis sur les notions de caractère algébrique. Nous nous occupons de deux problèmes posés par S. Marcus.

En étudiant la permutabilité de ces opérateurs nous avons démontré que les seuls paires d'opérateurs permutables sont: (G, φ), (G, λ) et (φ, λ). Si $\mathscr{F}$ est la classe des catégories morphologiques qui sont des fermetures de Sestier, nous allons montrer que ses éléments ont la forme: $G(X) = \varphi(X)$ [Théorème 2]. Il sera aussi montré que: 1. pour tout vocabulaire V et pour tout langage L sur V on a: $\lambda(X) \in \mathscr{F}$ et $\mu(X) \in \mathscr{F}$ (ou $X \subseteq V$); 2. la classe $\mathscr{F}$ est fermée par rapport à l'intersection mais elle n'est pas fermée par rapport à la réunion.

2. Definitions et notations

Les définitions, les notations et la terminologie sont celles de [7], [8], [9], [14], [1] et [2]. Considérons un vocabulaire fini V. Toute suite finie d'éléments de V est une *phrase* sur V. On appelle *langage* (sur V) toute collection de phrases (sur V). Toute paire ordonnée $\langle u, v\rangle$ de phrases sur V est une *contexte* sur V. On dira que x domine y et on écrira $x \to y$, si pour toute contexte $\langle u, v\rangle$ sur V, la relation $uxv \in L$ implique la relation $uyv \in L$ [3].

On a donc *citește* $\to$ *studiază* (en roumain), *beaux* $\to$ *gras* (en français) et окно $\to$ $\to$ солнце (en russe).

On n'a pas *gras* $\to$ *beaux* car la phrase *Cet objet est gras* est une phrase de la langue française et le remplacement de *gras* par *beaux* conduit à la phrase *Cet objet est beaux* qui n'est pas une phrase de la langue française.

Dans un langage naturel la relation $x \to y$ a l'interprétation suivante: l'homonymie morphologique qui se manifeste parmi les formes flexionnelles du mot x est plus pauvre ou égale à celle qui se manifeste parmi les formes flexionnelles du mot y.

Soient $A \subseteqq V$ et $B \subseteqq V$. On dit que A *domine* B et on écrit $A \to B$, si pour tout mot $a \in A$ et pour tout mot $b \in B$, on a $a \to b$.

L'ensemble des mots y tels que $x \to y$ et $y \to x$ constitue la *classe de distribution* ou la *famille* du mot x [4].

Si $S(x)$ est la classe de distribution de x, on a $S(x) = \{y \in V: x \leftrightarrow y\}$.

Exemple. Soit $V =$ le vocabulaire de la langue française et $L =$ l'ensemble des phrases françaises écrites correctement construites, alors *différent* $\leftrightarrow$ *nul* et donc *différent* $\in S$ (*nul*) [7].

L'ensemble $A \subseteqq V$ est dit *initial* s'il n'éxiste aucun mot n'appartenant pas à A et qui domine chaque mot de A [7]. Par *catégories morphologique* engendrée par un ensemble A de mots on comprend la réunion de A avec les mots dominés par tous le mots de A. Cette catégorie sera désignée par $G(A)$. Si $A_1 = \{x \in V: A \to x\}$, on a $G(A) = A \cup A_1$ [7].

Dans le cas particulier où A est une famille, $G(A)$ est, par définition, une catégorie morphologique élémentaire [3]. (Pour illustrations linguistiques voir [7].)

Si L est un langage sur V et $x \in V$, on dit que le contexte $\langle u, v\rangle$ *accepte* le mot x (ou que x est *accepté* par $\langle u, v\rangle$) si $uxv \in L$ ([12], [8]). Désignons par $\alpha(x)$ l'ensemble des contextes sur V qui acceptent le mot x (toujours par rapport au même langage L sur V) et par $\alpha^{-1}(c)$ l'ensemble des mots acceptés par le contexte c.

Si $X \subseteqq V$ et $\mathscr{C}$ est un ensemble de contextes, posons ([8], [13], [2]):

$$\alpha(X) = \bigcap_{x \in X} \alpha(x),\ \beta(X) = \bigcup_{x \in X} \alpha(x), \text{ pour } X \neq 0 \text{ et } \beta(0) = 0.$$

$$\alpha^{-1}(\mathscr{C}) = \bigcap_{c \in \mathscr{C}} \alpha^{-1}(c),\ \beta^{-1}(\mathscr{C}) = \bigcup_{c \in \mathscr{C}} \alpha^{-1}(c) \text{ pour } \mathscr{C} \neq 0 \text{ et } \beta^{-1}(0) = 0.$$

On déduit qus $\alpha(X)$ est l'ensemble des contextes qui acceptent tous les mots $x \in X$, $\beta(X)$ est l'ensemble des contextes qui acceptent moins un mot $x \in X$, $\alpha^{-1}(\mathscr{C})$ est l'ensemble des mots qui sont acceptés par tous les contextes $c \in \mathscr{C}$ et $\beta^{-1}(\mathscr{C})$ est l'ensemble des mots qui sont acceptés moins un contexte $c \in \mathscr{C}$.

Posons pour $X \subseteq V$:

$$\varphi(X) = \alpha^{-1}[\alpha(X)],\ \psi(X) = \beta^{-1}[\beta(X)].$$

$$\lambda(X) = \beta^{-1}[\alpha(X)] \text{ et } \mu(X) = \alpha^{-1}[\beta(X)].$$

L'ensemble $\varphi(X)$ est la *fermeture* de Sestier de l'ensemble X [12]. S. Marcus a introduit dans [10] les opérateurs λ et μ et Toma-Marcu dans [14] l'opérateur ψ.

J. Kunze ([5], [6]) a introduit une relation de domination, notée $\Rightarrow$, plus générale que la relation de Dobrušin, de la manière suivante: pour $X \subseteqq V$, $Y \subseteqq V$ (où V est le vocabulaire du langage considéré) on a $X \Rightarrow Y$ et $\alpha(X) \subseteq \alpha(Y)$. Comme il est montré dans [11], en utilisant la relation $\Rightarrow$ on peut présenter la fermeture de Sestier sous un aspect nouveau. Précisément: la définition de la fermeture de Sestier s'obtient en remplacant dans la définition de la catégorie morphologique, la relation de domination au sens de Dobrušin par la relation de domination au sens de Künze. Ce fait

est très utile parceque l'étude des fermetures de Sestier peut profiter ainsi des méthodes utilisées dans l'étude des catégories morphologiques ([7], chapitre 5).

La permutabilité des operateurs contextuels G, φ, Ψ, λ, et μ

Les opérateurs contextuels T et S sont permutables si pour tout ensemble non-vide $X \subseteq V$ on a:

$$T[S(X)] = S[T(X)].$$

En vertu du théorème 1 de [10], nous avons:

$$G[\varphi(X)] = \varphi(X). \tag{1}$$

Nous avons démontré ([2], Proposition 2) que pour tout ensemble $X \subseteqq V$ on a:

$$\varphi[G(X)\, J = \varphi(X). \tag{2}$$

Les égalités (1) et (2) nous montrent que $G[\varphi(X)] = \varphi[G(X)]$, c'est à-dire les opérateurs G et φ sont permutables.

À une sugestion donnée par S. Marcus, nous avons étudié cette propriété pour tous les opérateurs contextuels: G, φ, ψ, λ et μ. Voici le résultat obtenu:

Théorème 1. Les seules paires d'opérateurs permutables sont: (G, φ), (G, λ), *et* (φ, λ) (Les théorèmes 1 et 2 de [2]).

Catégories morphologiques qui sont des fermetures de Sestier

On sait que toute fermeture de Sestier est une catégorie morphologique mais pas réciproquement. (Les théorèmes 1 et 2 de [10].)

Dans le cas où l'ensemble X est une classe de distribution, les notions de catégories morphologique et de fermeture de Sestier coincident [8]. Il est possible que pour un ensemble X qui n'est pas une classe de distribution, les notions de catégorie morphologique et de fermeture de Sestier soient identiques [8].

Une catégorie morphologique $G(X)$ est une fermeture de Sestier, s'il existe un ensemble $Y \subseteqq V$ tel que $G(X) = \varphi(Y)$.

Nous allons donner maintenant quelques exemples des catégories morphologiques qui sont des fermetures de Sestier, obtenus à l'aide des opérateurs contextuels envisagés.

Proposition 1. *Pour tout ensemble* $X \subseteqq V$ *et pour tout langage* L *sur* V, *les ensembles* $\lambda(X)$ *et* $\mu(X)$ *sont catégories morphologiques et des fermetures de Sestier.*

Démonstration: On sait [10] que

$$G[\lambda(X)] = \lambda(X) \text{ et } G[\mu(X)] = \mu(X). \tag{3}$$

Dans [1] nous avons démontré que les ensembles $\lambda(X)$ et $\mu(X)$ sont des fermetures de Sestier. Précisément on a:

$$\varphi[\lambda](X)] = \lambda(X) \text{ et } \varphi[\mu(X)] = \mu(X). \tag{4}$$

Les égalités (3) et (4) nous montrent qu'on a

$$G[\lambda(X)] = \varphi[\lambda(X)] \text{ et } G[\mu(X)] = \varphi[\mu(X)]$$

c'est-à-dire que $\lambda(X)$ et $\mu(X)$ sont des catégories morphologiques et des fermetures de Sestier. La proposition 1 est ainsi démontrée.

Problème. (S. Marcus [11].)

Trouver une caractérisation des catégories morphologiques qui sont des fermetures de Sestier. Dans cet ordre d'idées on a le

Théorème 2. Afin que la catégorie morphologique $G(X)$ soit une fermetures de Sestier il faut et il suffit que $G(X) = \varphi(X)$.

Démonstration. Evidemment, la condition est suffisante. La condition est nécéssaire. Soit $G(X) = \varphi(Y)$. On a:

$$\varphi[G(X)] = \varphi[\varphi(Y)]. \tag{5}$$

Dans [2] nous avons démontré que $\varphi[G(X)] = \varphi(X)$.

On sait que $\varphi[\varphi(Y)] = \varphi(Y)$ ([12], [13]). De (5) on déduit que $\varphi(X) = \varphi(Y)$ et donc $G(X) = \varphi(X)$. Le théorème 2 est démontré.

Désignons par $\mathscr{F}$ la classe des catégories morphologiques qui sont des fermetures de Sestier. En vertu de la proposition 1, il résulte: *quel que soit le vocabulaire V et quel que soit le langage L sur V, $\lambda(X) \in \mathscr{F}$ et $\mu(X) \in \mathscr{F}$.*

Proposition 2. L'intersection de deux catégories morphologiques qui sont des fermetures de Sestier est une catégorie morphologique qui est fermeture de Sestier.

Démonstration. Soient $G(X) \in \mathscr{F}$ et $G(Y) \in \mathscr{F}$. Il s'ensuit donc qu'ils existent deux ensembles U' et V' tels que $G(X) = \varphi(U')$ et $G(Y) = \varphi(V')$ et donc $G(X) \cap G(Y) = \varphi(U') \cap \varphi(V')$.

On sait que l'intersection de deux fermetures de Sestier est aussi une fermeture de Sestier [9]. On déduit qu'il existe un ensemble $W \subseteq V$ tel que $G(X) \cap G(Y) = \varphi(W)$.

Mais, toute fermeture de Sestier est une catégorie morphologique (Théorème 1 de [10]); donc il existe un ensemble Z tel que $\varphi(W) = G(Z)$.

Remarques. 1. La réunion de deux ensembles de $\mathscr{F}$ n'est pas un ensemble de $\mathscr{F}$ parceque la réunion de deux fermetures de Sestier n'est pas toujours une fermeture de Sestier [9].

2. L'intersection de deux catégories morphologiques n'est pas toujous une catégorie morphologique [9]. La proposition 2 montre que cette propriétée est vraie pour les catégories morphologiques de $\mathscr{F}$.

BIBLIOGRAPHIE

[1] CRĂCIUN, C. V.: Quasi-catégories morphologiques et opérateurs contextuels. Bulletin mathématique de la Société des sciences mathématiques de la République Socialiste de Roumanie, *12*, (60), 1968 2, p. 29—39.

[2] CRĂCIUN, C. V.: Unele proprietăți ale operatorilor contextuali din lingvistica algebrică. Studii și cercetări matematice, *21*, 1970, 3, p. 419—434.

[3] ДОБРУШИН, Р. Д.: Математические методы в лингвистике. Приложение. Математическое просвещение, *6*, 1961, p. 52—59.

[4] КУЛАГИНА, О. С.: Об одном способе определения грамматических понятии на базе теории. множеств. Проблемы кибернетики, *1*, 1958, p. 203—214.

[5] KUNZE, J.: Versuch eines objektivierten Grammatikmodells I. Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung, *20*, 1967, p. 415—418.

[6] KUNZE, J.: Versuch eines objektivierten Grammatikmodells II. Arbeitsstelle für mathematische und angewandte Linguistik und automatische Übersetzung der Deutschen Akademie der Wissenschaften zu Berlin, Mai 1968, p. 68.

[7] MARCUS, S.: Introduction mathématique à la linguistique structurale. Paris, Dunod 1967.

[8] MARCUS, S.: Catégories de Dobrušin, fermetures de Sestier et voisinage de Sakai. Glossa, *1*, 1967, p. 59—67.

[9] MARCUS, S.: Catégories morphologiques et analyse contextuelle dans la linguistique algébrique. Deuxième Conférence internationale pour le traitement automatique des langues. Grenoble 23—25 août 1967. Communication, *39*, p. 1—8.

[10] MARCUS, S.: Opérateurs contextuels et catégories morphologiques. Bulletin mathématique de la Société des sciences mathématiques de la République Socialiste de la Roumanie, *12* (60), 1968, No. 3, p. 65—72.

[11] MARCUS, S.: Sur la domination au sens de Kunze dans la linguistique algébrique. Revue roumaine de mathématiques pures et appliquées, *14*, 1969, No. 3, p. 377—398.

[12] SESTIER, A.: Contribution à une théorie ensembliste des classifications linguistiques. Actes du Premier Congrès de l'AFCAL. Grenoble 1960, Paris 1961, p. 293—305.

[13] TOMA-MARCU, E.: Proprietăți ale închiderilor Sestier din teoria algebrică a gramaticii. Studii și cercetări matematice, *19*, 1967, 9, p. 1383—1390.

[14] TOMA-MARCU, E.: Asupra unui operator contextual inrudit cu operatorul lui Sestier din analiza algebrică contextuală. Studii și cercetări matematice, *21*, 1969, No. 3, p. 499—508.

The n-Derivative of a Partition

STEPHAN YLAN SOLOMON, BUCHAREST

In this work, the concept of n-derivative for the partitions of the vocabulary of a language is introduced and some of its properties are studied.

Thus, it proves to have many analogies with the usual concept of derivative introduced by Kulagina [2]. Both the usual and the n-derivative are particular cases of the concept of generalized derivative which belongs to Trybuleč [12].

Then, the different relations and connections existing between the derivative, n-derivative and asterisk operations are followed.

For the moment, it was not laid stress on the applications in natural languages of the shown concepts. That is to be the topic of a future work.

1. Partitions. Generalities

1.1. **Definition.** Let V be a non-empty set and ϱ a binary relation on V. ϱ is called an equivalence relation if it is reflexive, symmetric and transitive. The set of all equivalence relations on V will be denoted by $E(V)$. If $\varrho \in E(V)$ and $\varrho(x) = \{y \in V \mid x\varrho y\}$, then for every $u, v \in V$ we have either $\varrho(u) = \varrho(v)$ or $\varrho(u) \cap \varrho(v) = \varnothing$ and besides $\bigcup_{x \in V} \varrho(x) = V$. Therefore an equivalence relation is also called a partition of V.

We denote by 1 the unit partition ($x\, 1\, y$ if $x = y$, or $1(x) = \{x\}$ for any $x \in V$) and by ∞ the improper partition ($x \infty y$ for any $x, y \in V$, or $\infty(x) = V$ for each $x \in V$) of the set V. The set of all cells $\varrho(x)$ will be denoted by V/ϱ.

1.2. **Definition.** Let ϱ_1, ϱ_2 be binary relations on V. We shall say that ϱ_1 is finer that ϱ_2 (and we shall write $\varrho_1 \leqq \varrho_2$) if $x\varrho_1 y$ implies $x\varrho_2 y$ for any $x, y \in V$.

If $\varrho_1, \varrho_2 \in E(V)$, then $\varrho_1 \leqq \varrho_2$ if $\varrho_1(x) \subseteq \varrho_2(x)$ for each $x \in V$.

Obviously, $1 \leqq \varrho \leqq \infty$ for any $\varrho \in E(V)$.

1.3. **Definition.** Let $\varrho_1, \varrho_2 \in E(V)$. We shall denote by $\varrho_1 \vee \varrho_2$ the binary relation on V thus defined: $x(\varrho_1 \vee \varrho_2)\, y$ if there exist a natural number n and some elements $a_1, a_2, a_3, \ldots, a_{2n-1} \in V$ so that: $x\varrho_1 a_1, a_1\varrho_2 a_2, a_2\varrho_1 a_3, \ldots, a_{2n-2}\varrho_1 a_{2n-1}$ and $a_{2n-1}\varrho_2 y$.

The relation $\varrho_1 \vee \varrho_2$ is an equivalence relation and represents the upper limit of the relations ϱ_1 and ϱ_2 with respect to the order relation $\leqq$ introduced in $E(V)$.

2. Languages. Operation asterisk

2.1. **Definition.** Let V be a non-empty set called vocabulary and whose elements are named words. We shall call sentence over V a finite sequence (string) of words from V. By a language over the vocabulary V, we shall understand a set (finite or not) of sentences over V.

Let L be a language over V and $x_1, x_2, \ldots, x_n \in V$. If the sentence $x_1x_2 \ldots, x_n \in L$, it will be called a marked (correct) one.

2.2.2. **Definition.** Let L be a language over V and $\varrho \in E(V)$. We shall call factor language the language L/ϱ over the vocabulary V/ϱ for which $\varrho(x_1)\,\varrho(x_2) \ldots \varrho(x_n)$ is a marked sentence (i.e. belongs to L/ϱ) if there exist such $x_1', x_2', \ldots, x_n' \in V$ that $x_1'x_2' \ldots x_n' \in L$ and $x_i'\varrho x_i$ for $i = 1, 2, \ldots, n$.

2.3. **Definition.** Let L be a language over the vocabulary V, ϱ a partition of V and x, y two words of V. We say that x ϱ-dominates y with respect to L ($x \overset{\varrho}{\to} y$) if from $x_1x_2 \ldots x_{i-1}x_{i+1} \ldots x_n \in L$ there results the existence of such $x_1', x_2', \ldots, x_{i-1}', x_{i+1}', \ldots, x_n' \in V$ that $x_1' \ldots x_{i-1}'yx_{i+1}' \ldots x_n' \in L$ and $x_j'\varrho x_j$ for $j = 1, 2, \ldots i-1, i+1, \ldots n$. The ϱ-domination relation is a quasi-order relation, i.e. reflexive and transitive.

2.4. **Definition.** Let L be a language over V and $\varrho \in E(V)$. The asterisk of the partition ϱ will be the binary relation ϱ^* defined as follows: $x\varrho^*y$ if both $x \overset{\varrho}{\to} y$ and $y \overset{\varrho}{\to} x$ hold in L. Obviously, $\varrho^* \in E(V)$.

2.5. **Proposition.** If $\varrho_1, \varrho_2 \in E(V)$, then $\varrho_1 \leqq \varrho_2$ implies $\varrho_1^* \leqq \varrho_2^*$.

2.6. **Definition.** Let L be a language over V and $\varrho \in E(V)$. ϱ is said to be a saturated partition of L if for every marked sentence $\varrho(x_1) \ldots \varrho(x_n)$ of L/ϱ and for every i there exist such $x_1', \ldots, x_{i-1}', x_{i+1}', \ldots, x_n' \in V$ that $x_1' \ldots x_{i-1}'x_ix_{i+1}' \ldots x_n' \in L$ and $x_j'\varrho x_j$ for $j = 1, \ldots, i-1, i+1, \ldots, n$.

2.7. **Theorem.** The necessary and sufficient condition for a partition ϱ to be saturated is that $\varrho \leqq \varrho^*$.

2.8. **Proposition.** If the partitions ϱ_1 and ϱ_2 are saturated, then the partition $\varrho_1 \vee \varrho_2$ is saturated, too.

2.9. In a language L over V, there exists a greatest saturated partition, let us denote it by Z. That, since V being finite, the number of the partitions of V (and so much the more of its saturated partitions) is finite. Z is thus the result of the repeated application of the proposition 2.8.

2.10. **Theorem.** For any language L there is a natural number r so that $Z = \infty^{r*}$. The smallest number having this property will be called the saturated degree of L. (By the notation ∞^{n*} we understand: $\infty^{0*} = \infty$ and $\infty^{(n+1)*} = (\infty^{n*})^*$).

2.11. **Example.** Let $V = \{a, b, c, d, e, f, g, h, i\}$, $L = \{ae, bf, bg, cf, ch, dg, di, eh, fi, gi, hh, ii\}$. In this case, the partition ∞^* has the following classes: $\{a, b, c, d\}$, $\{e, f, g, h, i\}$. Further, we have:

$\infty^{**} = \langle\{a, b, c, d\}, \{e, f, g\}, \{h, i\}\rangle$;
$\infty^{3*} = \langle\{a, b\}, \{c, d\}, \{e, f, g\}, \{h, i\}\rangle$;
$\infty^{4*} = \langle\{a, b\}, \{c, d\}, \{e\}, \{f, g\}, \{h, i\}\rangle$;
$\infty^{5*} = \langle\{a\}, \{b\}, \{c, d\}, \{e\}, \{f, g\}, \{h\}, \{i\}\rangle$;
$\infty^{6*} = \langle\{a\}, \{b\}, \{c\}, \{d\}, \{e\}, \{f, g\}, \{h\}, \{i\}\rangle$;
$\infty^{7*} = 1$.

The saturation degree of L is 7. By using Kulagina's notation $S = 1^*$, we have $S = Z = 1$.

If we put $L' = L - \{cf, dg\}$, then the saturation degree of L' becomes 3. This time, $S = \langle\{a\}, \{b\}, \{c\}, \{d\}, \{e\}, \{f, g\}, \{h\}, \{i\}\rangle$ and $Z = \langle\{a, b\}, \{c, d\}, \{e, f, g\}, \{h, i\}\rangle$.

Remarks

Definitions 2.1. and 2.2. belong to Kulagina [2]. The relation defined in 2.3. generalizes the Dobrušin's [1] and Sestier's domination relation and that from [10] of Trybuleč ("ϱ-zamestimo na").

2.4. appears in [11] in two equivalent definitions. The same for the proposition 2.5.

Kulagina uses S for 1^* and Trybuleč N for ∞^*.

Items 2.6., 2.7., 2.8., 2.9., appear in [10].

Theorem 2.10. is proved in [9]. It represents the basis of an algorithm for reaching the partition Z in r steps (like in 2.11.).

3. Derivative and asterisk of a partition

3.1. **Definition.** Let L be a language over V and $\varrho \in E(V)$. We shall define the derivative ϱ' of the partition ϱ as the binary relation on V, for which we have $x\varrho'y$ if $\varrho(x)\,1^*\varrho(y)$ in L/ϱ. Obviously, ϱ' is also a partition and we have $\varrho \leqq \varrho'$.

Let $\varrho_1, \varrho_2 \in E(V)$. We shall say that ϱ_1 is regularly finer than ϱ_2 (with respect to L) if $\varrho_1 \leqq \varrho_2 \leqq \varrho_1'$.

3.2. **Theorem.** Let L be a language over V and $\varrho_1, \varrho_2 \in E(V)$ so that $\varrho_1 \leqq \varrho_2$. Then, $\varrho_1' = \varrho_2'$ if ϱ_1 is regularly finer than ϱ_2.

3.3. **Corollary.** For every $\varrho \in E(V)$ we have $\varrho'' = \varrho'$.

3.4. **Theorem.** The necessary and sufficient condition for two partitions ϱ_1 and ϱ_2 have the same derivative is the existence of a partition ϱ so that both ϱ_1 and ϱ_2 be regularly finer than ϱ.

3.5. **Theorem.** Let L be a language over V and $\varrho \in E(V)$. In order that $\varrho' = \varrho^*$ it is necessary and sufficient that ϱ is saturated.

3.6. **Corollary.** If $\varrho \in E(V)$ is saturated, then $\varrho^{**} = \varrho^*$.

3.7. **Theorem.** If $\varrho \in E(V)$, then the relations $\varrho \leqq \varrho'^*$, $\varrho' \leqq \varrho'^*$ and $\varrho' = \varrho'^*$ are pairwise equivalent.

3.8. **Definition.** A language provided with a paradigmatic structure, that is a system (V, ϱ, L) where L is a language over V and $\varrho \in E(V)$ is called adequate if $1^* \leqq \varrho'$.

3.9. **Proposition.** The language (V, ϱ, L) where the paradigm ϱ is saturated, is adequate.

3.10. **Theorem.** Let (V, ϱ, L) be a language and $R = \varrho \vee 1^*$. Then, the following conditions are equivalent:

$$1^* \leqq \varrho' \text{ (the language is adequate)}, \tag{1}$$

$$R \leqq \varrho', \tag{2}$$

$$R' = \varrho'. \tag{3}$$

Remarks

3.1. and 3.3. appear in an equivalent form in [2].

Theorems 3.2. and 3.4. belong to Uspenskii [13] (without proof) and Marcus [3, 4].

The concept of adequate language (3.8.) is due to Uspenskii [13]. The partition R (in mixed cells) was introduced by Revzin [7]. Theorem 3.10. belongs to Marcus [3, 4].

The other results (except 3.7.) are proved in a more general framework in [9].

4. The n-derivative

4.1. **Definition.** Let L be a language over V and $\varrho \in E(V)$. We shall define the n-derivative ϱ^n of the partition ϱ as the binary relation on V, for which we have $x \varrho^n y$ if $\varrho(x)\ \infty^n{}^* \varrho(y)$ in L/ϱ [n is a non-negative integer and $\infty^n{}^* \in E(V/\varrho)$].

Here are some immediate properties:

$$\varrho^0 = \omega, \tag{1}$$

$$\varrho^{n+1} \leqq \varrho^n, \tag{2}$$

$$1^n = \infty^n{}^*, \tag{3}$$

$$\varrho \leqq \varrho' \leqq \varrho^n. \tag{4}$$

Let $\varrho_1, \varrho_2 \in E(V)$. We shall say that ϱ_1 is n-regularly finer than ϱ_2 (with respect to L) if $\varrho_1 \leqq \varrho_2 \leqq \varrho^n$.

4.2. **Theorem.** Let L be a language over V and $\varrho_1, \varrho_2 \in E(V)$ so that $\varrho_1 \leqq \varrho_2$. Then, $\varrho_1^n = \varrho_2^n$ if ϱ_1 is n-regularly finer than ϱ_2.

4.3. **Corollary.** For any non-negative integer n, we have $(\varrho')^n = \varrho^n$.

4.4. **Corollary.** If $m \geqq n$, then $(\varrho^m)^n = \varrho^n$.

4.5. **Corollary.** If $m \geqq n$, then $(\varrho^{m*})^n = \infty^n{}^*$.

4.6. **Theorem.** The necessary and sufficient condition for two partitions ϱ_1 and ϱ_2 have the same n-derivative is the existence of a partition ϱ so that both ϱ_1 and ϱ_2 be n-regularly finer than ϱ.

4.7. **Proposition.** For any non-negative integer n, we have $(\varrho^n)' = \varrho^n$.

4.8. **Lemma.** For any natural number n, we have $(\varrho^*)^n \leqq (\varrho \vee (\varrho^*)^{n-1})^*$.

4.9. **Proposition.** If $n \geqq m$, then $(\varrho^{m*})^n \leqq \infty^m{}^*$.

4.10. **Proposition.** We have $(1^m)^n = 1^{\min(m, n)}$, for any two non-negative integers m and n.

4.11. **Definition.** Let L be a language over V, $\varrho \in E(V)$ and n a natural number. We shall say that the partition ϱ is n-saturated if the inequality $\varrho \leqq (\varrho^{n-1})^*$ holds in L.

A saturated partition is n-saturated for every n.

4.12. **Proposition.** If the partition ϱ is n-saturated, then we have $\varrho^m = \infty^m{}^*$ for $m = 0, 1, 2, \ldots, n$.

4.13. **Corollary.** If ϱ is a saturated partition, then $\varrho^n = \infty^n{}^*$ for all non-negative integers n.

4.14. **Definition.** A language provided with a paradigmatic structure (V, ϱ, L) is called n-adequate if $\infty^n{}^* \leqq \varrho^n$.

4.15. **Proposition.** The language (V, ϱ, L) where the paradigm ϱ is n-saturated is m-adequate for $m = 0, 1, 2, \ldots, n$.

4.16. **Theorem.** Let (V, ϱ, L) be a language, $R_n = \varrho \vee \infty^n{}^*$ and $\bar{R}_n = \varrho' \vee \infty^n{}^*$. Then, the following conditions are pairwise equivalent:

$$\infty^n{}^* \leqq \varrho^n \text{ (the language is } n\text{-adequate)} \tag{1}$$

$$R_n \leqq \varrho^n, \tag{2}$$

$$R_n^n = \varrho^n, \tag{3}$$

$$\bar{R}_n \leqq \varrho^n, \tag{4}$$

$$\bar{R}_n^n = \varrho^n. \tag{5}$$

4.17. **Example.** Let $V = \{a_1, a_2, b_1, b_2, c_1, c_2, d_1, d_2, e_1, e_2, f, g, h_1, h_2, i_1, i_2\}$, $L = \{a_1e_2, a_2e_2, b_1f, b_2g, c_1h_2, c_2h_2, d_1i_2, d_2i_2, e_1h_2, fi_2, gi_2, h_1h_2, i_1i_2\}$ and $\varrho = \langle a = \{a_1, a_2\}, b = \{b_1, b_2\}, c = \{c_1, c_2\}, d = \{d_1, d_2\}, e = \{e_1, e_2\}, f = \{f\}, g = \{g\}, h = \{h_1, h_2\}, i = \{i_1, i_2\}\rangle$.

According to 2.2., $L/\varrho = \{ae, bf, bg, ch, di, eh, fi, gi, hh, ii\}$.

Like in 2.11., we have in L/ϱ:

∞^* $= \langle\{a, b, c, d\}, \{e, f, g, h, i\}\rangle$;

∞^{**} $= \langle\{a, b, c, d\}, \{e, f, g\}, \{h, i\}\rangle$;

∞^{***} $= \langle\{a, b\}, \{c, d\}, \{e, f, g\}, \{h, i\}\rangle = Z$;

1^* $= \langle\{a\}, \{b\}, \{c\}, \{d\}, \{e\}, \{f, g\}, \{h\}, \{i\}\rangle = S$.

According to 4.1., we have in L:

$\varrho^1 = \{a \cup b \cup c \cup d, e \cup f \cup g \cup h \cup i\} = \langle\{a_1, a_2, b_1, b_2, c_1, c_2, d_1, d_2\}, \{e_1, e_2, f\, g, h_1, h_2, i_1, i_2\}\rangle$;

$\varrho^2 = \{a \cup b \cup c \cup d, e \cup f \cup g, h \cup i\} = \langle\{a_1, a_2, b_1, b_2, c_1, c_2, d_1, d_2\}, \{e_1, e_2, f, g\}, \{h_1, h_2, i_1, i_2\}\rangle$;
$\varrho^3 = \{a \cup b, c \cup d, e \cup f \cup g, h \cup i\} = \langle\{a_1, a_2, b_1, b_2\}, \{c_1, c_2, d_1, d_2\}, \{e_1, e_2, f, g\}, \{h_1, h_2, i_1, i_2\}\rangle$.
According to 3.1., $\varrho' = \langle\{a_1, a_2\}, \{b_1, b_2\}, \{c_1, c_2\}, \{d_1, d_2\}, \{e_1, e_2\}, \{f, g\}, \{h_1, h_2\}, \{i_1, i_2\}\rangle$.
According to 2.4., we have: $\varrho^* = \langle\{a_1, a_2\}, \{b_1\}, \{b_2\}, \{c_1, c_2, e_1, h_1\}, \{d_1, d_2, i_1\}, \{e_2\}, \{f, g\}, \{h_2\}, \{i_2\}\rangle$;
$\varrho^{**} = \langle\{a_1, a_2\}, \{b_1, b_2\}, \{c_1, c_2, e_1, h_1\}, \{d_1, d_2, i_1\}, \{e_2\}, \{f\}, \{g\}, \{h_2\}, \{i_2\}\rangle$ and generally $\varrho^{(2n+1)*} = \varrho^*$, $\varrho^{(2n+2)*} = \varrho^{**}$ for any natural number n.

Remarks

The definition of the n-derivative is analogous of that of Kulagina's derivative, by using the partition $1^n = \infty^{n*}$ instead of the partition $S = 1^*$. Both the partitions 1^* and 1^n are particular cases of syntactic partitions, concept introduced by Trybuleč [10]. Thus, the n-derivative is another particular illustration of the concept of generalized derivative [12].

Definitions 3.1. and 4.1. are particular cases of definition 4 of [12]. Theorems 3.2. and 4.2. are inclosed in theorem 3 of [12]. The same, a part of corollary 4.4. $[(\varrho^n)^n = \varrho^n]$, as well as the relation $\varrho \leqq \varrho^n$ [4.1., (4)], analogous of 3.3. and 3.1.

From 4.3. and 4.7. it can be seen that the n-derivative is stronger than the usual derivative.

From the previous paragraphs, as well as from 4.10. and 4.13., it can be found that the partitions 1^* and 1^n ($n = 1, 2, \ldots, r$) are the single ones that can be obtained in a general language from the unit and improper partitions, by means of the asterisk, derivative and n-derivative operations.

Corollary 4.13. can be used for a more rapid determination of the partition Z. Instead of determining the partitions ∞^{n*} of V, one can work with the language L/ϱ over the more limited vocabulary V/ϱ; when the sequence of partitions ∞^{n*} becomes stationary (in L/ϱ), one comes back to L, obtaining the partition Z.

The analogy between the concepts of n-derivative and derivative allows the introduction of the concept of n-adequate language (4.14.). Thus, items 4.15. and 4.16. correspond to the items 3.9. and 3.10., if we associate to R, R_n or $\bar{R}_n$ and to the concept of saturated partition, the concept of n-saturated partition.

REFERENCES

[1] DOBRUSHIN, R. L.: Matematicheskie metodi v lingvistike. Prilozhenie. Matematicheskie prosveshchenie, *6*, 1961, pp. 52—59.

[2] KULAGINA, O. S.: Ob odnom sposobe opredelenia gramaticheskikh ponyatii na baze teorii mnozhestv. Problemi kibernetiki, *1*, pp. 203—214.

[3] MARCUS, S.: Asupra unui model logic al părtii de vorbire. St. cerc. matem., *13*, 1962, 1, pp. 37—62.

[4] MARCUS, S.: Algebraic Linguistics. Analytical Models. New York, Academic Press 1967.

[5] NOVOTNÝ, M.: On some algebraic concepts of mathematical linguistics. Prague Stud. in math. linguistics, *1*, 1966, pp. 125—140.

[6] REVZIN, I. I.: Modeli jazika. Moscow 1962.

[7] REVZIN, I. I.: On some aspects of the contemporary theoretic researches concerning mechanical translation. Byul. obedin. probl. masinnogo perevoda, *7*, 1958, pp. 1—12.

[8] SESTIER, A.: Contribution à une théorie ensembliste des classifications linguistiques. Actes du premier congrès de l'AFCAL, Grenoble 1960. Paris 1961, pp. 293—305.

[9] SOLOMON, S. Y.: Lingvistica algebrică si teoria modelelor. St. cerc. mat., *21*, 1969, pp. 1107—1134.

[10] TRYBULECH, A.: Ob odnom klasse sintaksicheskikh otnoshenii. Nauchno-tekhnicheskaya inform., *9*, 1967, pp. 34—37.

[11] TRYBULECH, A.: Razbienie slovnika v jazikakh s zadanimi paradigmami. Nauchno-tekhnicheskaya inform., *12*, 1967, pp. 40—44.

[12] TRYBULECH, A.: Obobshchenie ponyatia proizvodnogo razbienia. Nauchno-tekhnicheskaya inform., *5*, 1969, pp. 18—21.

[13] USPENSKII, V. A.: K opredeleniyu chasti rechi v teoretikomnozhestvennoj sisteme yazika. Byul. obedin. probl. massinogo perevoda, *5*, 1957, pp. 22—26.

Quelqes résultats concernant les ensembles homologues

GABRIEL ORMAN, BRAŞOV

1. Avant toute chose nous allons rappeler les principales définitions, comme elles sont données en [4].

Soit Γ un ensemble d'éléments de nature quelconque. Toute suite finie d'éléments de Γ sera une *chaîne*. Une chaîne β est une *souschaîne* d'une chaîne α s'il existe deux chaînes μ et ν ayant la propriété $\mu\beta\nu = \alpha$.

On appelle *sous-chaîne initiale*, d'une chaîne α, toute chaîne β telle qu'il existe une chaîne γ ayant la propriété $\beta\gamma = \alpha$. On appelle *sous-chaîne finale*, d'une chaîne α, toute chaîne δ telle qu'il existe une chaîne η ayant la propriété $\alpha = \eta\delta$. Une sous-chaîne qui n'est ni initiale, ni finale s'appelle une *sous-chaîne médiane*.

Une paire ordonnée de chaînes s'appelle *opposition ordonnée*.

On appelle *base initiale* de l'opposition (α, β) la sous-chaîne initiale maximale commune à α et à β. La *sous-chaîne* finale maximale commune à α et à β, s'appelle la *base finale* de l'opposition (α, β). Une sous-chaîne médiane maximale, commune à α et à β, sera appelée *une base médiane* de l'opposition (α, β).

Une opposition ordonée (α, β) s'appelle *initiale*, *finale* ou *médiane* selon qu'elle est considérée par rapport à sa *base initiale*, à sa *base finale* ou par rapport à une de ses bases médianes.

Désignons l'opposition initiale par $(\alpha, \beta)_i$, l'opposition finale par $(\alpha, \beta)_f$ et l'opposition médiane par $(\alpha, \beta)_\gamma$, où γ est la base médiane à raison de laquelle nous avons considéré cette opposition.

Nous définirons une opération de composition des chaînes dans la manière suivante: étant données deux chaînes α et β, l'opération de composition conduit à la chaîne $\alpha\beta$, obtenue par la juxtaposition de la chaîne β à la droite de α. Evidemment, cette opération peut être définie pour un nombre fini, quelconque, de chaînes.

Soit $(\alpha, \beta)_i$ une opposition initiale, et γ la base initiale correspondante. Alors, il existe deux chaînes α' et β', uniquement déterminées, telle que $\alpha = \gamma\alpha'$ et $\beta = \gamma\beta'$. Les chaînes α' et β' s'appellent les *sous-chaînes différentielles* de l'opposition initiale $(\alpha, \beta)_i$. On peut distinguer les types suivants d'oppositions initiales:

a) L'opposition $(\alpha, \beta)_i$ sera appelée *opposition zéro* si $\alpha = \beta$, c'est-à-dire si les sous-chaînes différentielles sont vides: $\alpha' = \beta' = 0$;

b) L'opposition $(\alpha, \beta)_i$ sera appelée *privative à gauche* si α' est la chaîne vide, tandis que β' n'est pas vide;

c) L'opposition $(\alpha, \beta)_i$ sera appelée *privative à droite* si β' est la chaîne vide, tandis que α' n'est pas vide.

Une opposition privative à gauche ou privative à droite est appelée *opposition privative.*

d) L'opposition $(\alpha, \beta)_i$ sera appelée *disjonctive* si la base initiale correspondante est la chaîne vide: $\gamma = 0$;

e) L'opposition $(\alpha, \beta)_i$ sera appelée *équipollente* si elle n'est ni zéro, ni privative, ni disjonctive;

f) L'opposition $(\alpha, \beta)_i$ sera appelée *propre* si $\alpha \neq 0 \neq \beta$; si non elle sera appelée *impropre.*

Évidemment, on peut définir ces types d'oppositions tant pour les oppositions finales que pour les oppositions médianes.

Considérons, maintenant, deux oppositions initiales $(\alpha_1, \beta_1)_i$ et $(\alpha_2, \beta_2)_i$ et nous désignons par $\alpha_1', \beta_1', \alpha_2', \beta_2'$ les sous-chaînes différentielles des oppositions et par γ_1 et γ_2 leurs bases initiales. Les oppositions initiales $(\alpha_1, \beta_1)_i$ et $(\alpha_2, \beta_2)_i$ s'appellent *proportionnelles* si leurs sous-chaînes différentielles sont égales: $\alpha_1' = \alpha_2'$ et $\beta_1' = \beta_2'$. Si l'on a seulement la relation $\alpha_1' = \alpha_2'$, elles s'appellent *proportionnelles à gauche*, tandis que si l'on a seulement la relation $\beta_1' = \beta_2'$ elles seront appelées *proportionnelles à droite.*

Nous rappelerons, aussi, la définition de la notion de partition d'un ensemble donné. Soit M un ensemble et $M_1, M_2, M_3, \ldots, M_n, \ldots$ les sous-ensembles de M tel que n'existe pas un élément appartenant à deux sous-ensembles. Si

$$M = M_1 \cup M_2 \cup M_3 \cup \ldots \cup M_n \cup \ldots$$

alors, on dit que cette relation définit une *partition* de M.

Enfin, nous désignons par Ω un ensemble précis de chaînes. Les éléments de Ω sont, par définition, *les mots permis*. Considérons aussi, une partition Π de Ω : $\Omega = \bigcup_i \Omega_i$. Soit Φ un ensemble précis de chaînes dont les éléments appartiennent à Ω. Les éléments de Φ sont, par définition, *les phrases permises*. L'ensemble $L = \{\Omega, \Pi, \Phi\}$ s'appelle *langue.*

On dit que deux ensembles Ω_i et Ω_i de la partition Π sont *homologues* si l'on peut établir entre Ω_i et Ω_j une corespondance biunivoque ainsi que, si $\alpha_1 \in \Omega_i$, $\beta_1 \in \Omega_i$, $\alpha_2 \in \Omega_j$, $\beta_2 \in \Omega_j$ et si α_2 correspond à α_1 et β_2 correspond à β_1, les oppositions initiales $(\alpha_1, \beta_1)_i$ et $(\alpha_2, \beta_2)_i$ sont proportionnelles. Évidemment, la relation d'homologie est une relation d'équivalence; les classes d'équivalence obtenues de cette manière sont *les classes d'homologie.*

Nous rappelons qu'une relation R, définie entre les éléments d'un ensemble M, s'appelle *relation d'équivalence* si les propriétés suivantes sont accomplies: a) Pour

tout $x \in M$ on a xRx; b) Si $x \in M$, $y \in M$ et on a xRy, alors on a aussi yRx; c) Si $x \in M$, $y \in M$, $z \in M$ et on a xRy et yRz, alors on a xRz.

2. Une autre fois (voir [7]), nous avons essayé d'étudier les ensembles homologues à l'aide des oppositions finales, en partant de la proposition suivante: étant donnés deux ensembles homologues Ω_i et Ω_j si $\alpha_1 \in \Omega_i$, $\beta_1 \in \Omega_i$, $\alpha_2 \in \Omega_j$, $\beta_2 \in \Omega_j$ et si α_2 correspond à α_1 tandis que β_2 correspond à β_1, les oppositions finales $(\alpha_1, \alpha_2)_f$ et $(\beta_1, \beta_2)_f$ sont proportionnelles. (Cette proposition est énoncée en [4].)

A cette occasion nous avons établi quelques propositions qui nous ont permis d'énoncer, à la fin, une condition nécessaire et suffisante pour que deux ensembles soient homologues. On a pu conclure donc, qu'on a une modalité de vérifier toujours si deux ensembles sont ou ne sont pas homologues en vérifiant seulement la proportionnalité des paires d'oppositions finales, ce qui est très simple, leur nombre étant inférieur au nombre des paires d'oppositions initiales proportionnelles.

3. Dans ce qui suit nous nous proposons de trouver les rapports de dépendance entre la paire d'oppositions initiales proportionnelles utilisées pour définir les ensembles homologues et la paire d'oppositions finales proportionnelles correspondantes.

Soit Ω_i et Ω_j les deux ensembles homologues définis plus haut. Désignons les bases initiales des oppositions $(\alpha_1, \beta_1)_i$ et $(\alpha_2, \beta_2)_i$ respectivement par γ_1 et γ_2; désignons aussi par α_1', β_2' les sous-chaînes différentielles de l'opposition $(\alpha_1, \beta_1)_i$ et par α_2', β_2' les sous-chaînes différentielles de l'opposition $(\alpha_2, \beta_2)_i$. Nous avons alors:

$$\begin{matrix} \alpha_1 = \gamma_1\alpha_1' \\ \beta_1 = \gamma_1\beta_1' \end{matrix} \quad \text{et} \quad \begin{matrix} \alpha_2 = \gamma_2\alpha_2' \\ \beta_2 = \gamma_2\beta_2' \end{matrix} \tag{1}$$

En considérant maintenant, au cas des ensembles homologues, les oppositions finales proportionnelles $(\alpha_1, \alpha_2)_f$ et $(\beta_1, \beta_2)_f$ on a $\alpha_1' = \alpha_2'$ et $\beta_1' = \beta_2'$. Alors, on obtient, des relations (1), la représentation suivante pour ces oppositions finales proportionnelles:

$$\begin{matrix} \alpha_1 = \gamma_1\alpha_1' \\ \alpha_2 = \gamma_2\alpha_1' \end{matrix} \quad \text{et} \quad \begin{matrix} \beta_1 = \gamma_1\beta_1' \\ \beta_2 = \gamma_2\beta_1' \end{matrix} \tag{2}$$

Proposition 1. Si les oppositions initiales proportionnelles $(\alpha_1, \beta_1)_i$ et $(\alpha_2, \beta_2)_i$ sont des oppositions zéro, les oppositions finales proportionnelles $(\alpha_1, \alpha_2)_f$ et $(\beta_1, \beta_2)_f$ coïncident.

Démonstration. Les oppositions $(\alpha_1, \beta_1)_i$ et $(\alpha_2, \beta_2)_i$ étant des oppositions zéro il résulte, des relations (1), que $\alpha_1' = \beta_1' = 0$ et $\alpha_2' = \beta_2' = 0$,

$$\text{donc } \alpha_1 = \gamma_1 = \beta_1 \text{ et } \alpha_2 = \gamma_2 = \beta_2.$$

Observation 1. Dans les conditions de la proposition 1, les oppositions finales $(\alpha_1, \alpha_2)_f$ et $(\beta_1, \beta_2)_f$ ne peuvent pas être des oppositions zéro.

En effet, si ces oppositions étaient des oppositions zéro, on aurait $\gamma_1 = \gamma_2$. Mais comme les oppositions finales considérées coïncident il faudrait que les oppositions

initiales $(\alpha_1, \beta_1)_i$ et $(\alpha_2, \beta_2)_i$ soient homogènes, ce qui vient en contradiction avec l'hypothèse.

Proposition 2. Si les oppositions initiales proportionnelles $(\alpha_1, \beta_1)_i$ et $(\alpha_2, \beta_2)_i$ sont privatives à droite, les oppositions finales proportionnelles $(\alpha_1, \alpha_2)_f$ et $(\beta_1, \beta_2)_f$ ne pourront être des oppositions zéro, disjonctives et impropres.

Démonstration. Parce que $\beta_1' = 0$, $\alpha_1' \neq 0$, $\beta_2' = 0$, $\alpha_2' \neq 0$ on obtient, des relations (2), que $\beta_1 = \gamma_1$ et $\beta_2 = \gamma_2$. Mais alors, il s'ensuit que les termes de l'opposition finale $(\beta_1, \beta_2)_f$ sont les sous-chaînes différentielles de l'opposition $(\alpha_1, \alpha_2)_f$ et donc: a) Si l'opposition $(\alpha_1, \alpha_2)_f$ était une opposition zéro, c'est-à-dire $\beta_1 = \beta_2 = 0$, on aurait immédiatement $\gamma_1 = \gamma_2 = 0$ et donc les oppositions initiales $(\alpha_1, \beta_1)_i$ et $(\alpha_2, \beta_2)_i$ seraient disjonctives ce qui vient en contradiction avec l'hypothèse; b) Si l'opposition finale $(\alpha_1, \alpha_2)_f$ était disjonctive alors $\alpha_1' = 0$ ce qui, de nouveau, est en contradiction avec l'hypothèse; c) Si l'opposition finale $(\alpha_1, \alpha_2)_f$ est impropre on obtient aussi les conclusions inscrites sous a) et b).

Proposition 3. Si les oppositions initiales $(\alpha_1, \beta_1)_i$ et $(\alpha_2, \beta_2)_i$ sont privatives à gauche, les oppositions $(\alpha_1, \alpha_2)_f$ et $(\beta_1, \beta_2)_f$ ne pourront être des oppositions zéro, disjonctives et impropres.

La démonstration est ressemblante à celle de la proposition 2.

Corollaire. Des propositions 2 et 3 il s'ensuit qu'au cas où les oppositions initiales $(\alpha_1, \beta_1)_i$ et $(\alpha_2, \beta_2)_i$ sont privatives, les oppositions finales $(\alpha_1, \alpha_2)_f$ et $(\beta_1, \beta_2)_f$ ne peuvent pas être des oppositions zéro, disjonctives et impropres pouvant être, en échange, des oppositions privatives ou équipollentes.

Proposition 4. Étant donnés les oppositions initiales proportionnelles $(\alpha_1, \beta_1)_i$ et $(\alpha_2, \beta_2)_i$ si l'opposition $(\alpha_1, \beta_1)_i$ est équipollente et l'opposition $(\alpha_2, \beta_2)_i$ est équipollente ou disjonctive, les oppositions finales proportionnelles $(\alpha_1, \alpha_2)_f$ et $(\beta_1, \beta_2)_f$ seront équipollentes ou privatives à droite.

Démonstration. a) Si l'opposition $(\alpha_2, \beta_2)_i$ est équipollente alors $\gamma_2 \neq 0$ ce qui entraîne que les sous-chaînes différentielles des oppositions finales sont nonvides. Il s'ensuit donc, des relations (2), que ces oppositions sont équipollentes. b) Si l'opposition $(\alpha_2, \beta_2)_i$ est disjonctive, on aura $\gamma_2 = 0$. Mais en ce cas on obtient $\alpha_2 = \alpha_1'$ et $\beta_2 = \beta_1'$ et par suite les oppositions $(\alpha_1, \alpha_2)_f$ et $(\beta_1, \beta_2)_f$ sont privatives à droite.

Observation 2. On obtient les mêmes résultats si l'opposition $(\alpha_1, \beta_1)_i$ est propre et l'opposition $(\alpha_2, \beta_2)_i$ est équipollente ou disjonctive.

Proposition 5. Si les oppositions initiales $(\alpha_1, \beta_1)_i$ et $(\alpha_2, \beta_2)_i$ sont proportionnelles, l'opposition $(\alpha_1, \beta_1)_i$ étant disjonctive tandis que l'opposition $(\alpha_2, \beta_2)_i$ est équipollente ou disjonctive, alors les oppositions finales proportionnelles $(\alpha_1, \alpha_2)_f$ et $(\beta_1, \beta_2)_f$ seront privatives à gauche ou des oppositions zéro.

Démonstration. On a donné $\gamma_1 = 0$, $\alpha_1' \neq 0$, $\beta_1' \neq 0$. a) Si $(\alpha_2, \beta_2)_i$ est une opposition équipollente, alors $\gamma_2 \neq 0$ et parce qu'aussi $\gamma_1 = 0$ on obtient de (2) que $\alpha_1 = \alpha_1'$ et $\beta_1 = \beta_1'$. Il résulte que $(\alpha_1, \alpha_2)_f$ et $(\beta_1, \beta_2)_f$ sont des oppositions privatives à gauche. b) Si l'opposition $(\alpha_2, \beta_2)_i$ est disjonctive, alors $\gamma_2 = 0$. Mais comme on a aussi

$\gamma_1 = 0$ il s'ensuit que $\alpha_1 = \alpha_1' = \alpha_2$ et $\beta_1 = \beta_1' = \beta_2$. Par conséquent les oppositions $(\alpha_1, \alpha_2)_f$ et $(\beta_1, \beta_2)_f$ sont des oppositions zéro.

Proposition 6. Soit les oppositions initiales proportionnelles $(\alpha_1, \beta_1)_i$ et $(\alpha_2, \beta_2)_i$, où l'opposition $(\alpha_1, \beta_1)_i$ est impropre. Alors: a) Si $\alpha_1 = 0$ et l'opposition $(\alpha_2, \beta_2)_i$ est privative à gauche, ou $\beta_1 = 0$ et l'opposition $(\alpha_2, \beta_2)_i$ est privative à droite, alors les oppositions finales proportionnelles $(\alpha_1, \alpha_2)_f$ et $(\beta_1, \beta_2)_f$ sont privatives à gauche. b) Si $\alpha_1 = \beta_1$ et $\alpha_2 = \beta_2$, les oppositions finales $(\alpha_1, \alpha_2)_f$ et $(\beta_1, \beta_2)_f$ coïncident.

Démonstration. Soit $\alpha_1 = 0$ et l'opposition $(\alpha_2, \beta_2)_i$ privative à gauche. On obtient alors $\alpha_1' = 0$, $\gamma_1 = 0$ et $\alpha_2' = 0$ ce qui entraîne $\alpha_2 = \gamma_2$ et $\beta_1 = \beta_1'$. Donc les oppositions $(\alpha_1, \alpha_2)_f$ et $(\beta_1, \beta_2)_f$ sont privatives à gauche. Si nous allons prendre, maintenant, $\beta_1 = 0$ tandis que l'opposition $(\alpha_2, \beta_2)_i$ est privative à droite, on obtiendra le même résultat. Le deuxième cas a été étudié dans la proposition 1.

Comme une conclusion finale, on peut dire que les oppositions proportionnelles $(\alpha_1, \alpha_2)_f$ et $(\beta_1, \beta_2)_f$ sont du même type avec les oppositions proportionnelles $(\alpha_1, \beta_1)_i$ et $(\alpha_2, \beta_2)_i$ seulement au cas où les dernières sont équipollentes ou privatives. Dans ce dernier cas si les oppositions $(\alpha_1, \beta_1)_i$ et $(\alpha_2, \beta_2)_i$ sont privatives à droite, ayant $\beta_1 = 0$, les oppositions $(\alpha_1, \alpha_2)_f$ et $(\beta_1, \beta_2)_f$ seront privatives à gauche.

BIBLIOGRAPHIE

[1] CANTINEAU, J.: Le classement logique des oppositions. Word, vol. 11, Nr. 1, 1955.

[2] DIACONESCU, P.: Pe marginea unor lucrări despre morfem. Studii și cercetări lingvistice, vol. 13, Nr. 4, 1962.

[4] MARCUS, S.: Lingvistica Matematică. II-ème éd. E. D. P., București 1966.

[5] MARCUS, S.: Introduction mathématique à la linguistique structurale. Paris, Dunod 1967.

[6] ORMAN, G.: Remarques sur la théorie logique des oppositions linguistiques. Cahiers de linguistique théorique et appliqée, 1966, Nr. 3.

[7] ORMAN, G.: Sur les classes d'homologie des adjectifs et des substantifs français. Communication presentée au XII[e] Congrès International de Linguistique et Philologie Romanes, Bucarest, 15—20 avril 1968.

On the Definition of the Word-forming Paradigm

KLÁRA BUZÁSSYOVÁ, BRATISLAVA

The description of the system of word formation may be approached either from the point of view of the form and then we may proceed to the meaning, or, in the reversed direction, from the meaning to the form of its expression. I shall proceed from the meaning to the form in this contribution. The object of this contribution is the investigation of differential meanings of the deverbative derivatives, that is the investigation of structural meanings on the paradigmatic axis in contrary with the structural meanings on the syntagmatic axis. The most general classification frame in this description are deverbative word-forming paradigms; within this frame, the differential word-forming meanings are described by means of distinctive features. We distinguish between two hierarchically different word-forming and semantic features, which mark the word-forming meaning: categorial features, that is such as constitute particular classes, and concomitant features, that is such as differentiate members of a particular class form each other.[1]

Categorial meanings derived from a given word are decisive for the inclusion of the word into a word-forming paradigm. We introduce the concept of the word-forming paradigm into our description as one analogical to the concept of the semantic paradigm. In semantics, paradigms are made use of in the description of elements which have some common semantic features which allow their inclusion into the same semantic universe. Within such a universe, specified, for instance, as names of relationship, names of colours, etc., that is within closed lexematic strings, the significative meaning of the lexemes is described by means of semantic components.[2]

---

[1] The inventory of word-forming meanings in the relation to lexical meanings and the categorial and concomitant word-forming features have been limited in our study Über die distinktiven Merkmale bei den deverbativen Ableitungen. Recueil linguistique de Bratislava III. Publishing House of the Slovak Academy of Sciences, Bratislava 1972, pp. 85—98. Paradigmatical classification of word-families of 1600 verbs is given in our disertation Sémantická štruktúra slovenských deverbatív (in print).

[2] Cf. W. H. GOODENOUGH, Componential Analysis and the Study of Meaning. Language, *1*, 1956, pp. 195—216; Anthony F. C. WALLACE and J. ATKINS, The Meaning of Kinship Terms. American Anthropologist, vol. 62, February 1960, pp. 58—80.

The concept of the paradigm may also be used in the description of the word-forming level, especially when ascertaining the derivative capacity of words (verbs in our case), though it should be kept in mind that word-forming paradigms will be open classes here. And it is namely here that the description by means of features, that is by means of properties, is advantageous. Since it is impossible to define a system of open classes extensionally, that is by enumerating all the members of the class, but it is possible to do so intentionally, that is by giving the properties of the members of the given class.

Verbal derivatives are characterized by one common semantic feature, namely, their relation towards the action; that feature allows of their inclusion into the same semantic space, and hence it is a condition for describing the meaning of the derived words by means of features. A juxtaposition of word families (of verbs in our case) and the study of their members will show that the abundance of a verbal word-family, and hence the derivative capacity of the verbs in question are conditioned in a considerable degree by the intention of verbal action of the deriving verbal lexeme which reflects, in the most abstract form, the semantics of the verb. All the verbs under discussion can be thus grouped according to the meaning of their possible derivatives. It could be put metaphorically that the derivatives of a given word (head word or basic word) form the "word-forming declension" of the same head word. The word-forming paradigm then includes words which have the same "word-forming declension".[3] For instance, the word-forming declension of the Slovak verb *tkať* (weave) will include the following derivatives: *tkáč* (weaver — the name of the agent), *tkáčstvo* (weaver's trade, weaving—action), *tkanivo* (warp, weft — the material used in the realization of the action), *tkanica*, *tkanivo*, *tkanina* (lace, web, fabric — the results of the action), *tkáčovňa* (the weaver's workshop — the place of the action), *tkací*, *tkaný* (weaving — adjectives denoting the properties of the action).

As it may be seen, it is necessary to distinguish between the two particular derived lexemes and the word-forming meanings of these lexemes, since the meaning and the form need not be in strict one-to-one correspondence

$$\left(\text{e.g. } tkanivo \left\langle \begin{array}{l} \text{material used} \\ \text{result of the action} \end{array} \right.\right).$$

[3] The label derivational paradigm for the word-family (i.e. head word and its derivatives) had been used by Dean S. WORTH in his study The Role of Transformations in the Definition of Syntagmas in Russian and other Slavic Languages, American Contributions to the Fifth International Congress of Slavists as early in Sofia 1963 and it is understood as a term from morphological level. In our view the word-forming paradigm does not principially contradict with the views of Worth's derivational paradigms, our starting point as well as his are also the word-families, but we are searching for which word (verbs in our case) have equal and different word-families. We are getting arrived at the limitation of word-forming paradigms after having confronted different word-families. The confrontation is going on the lexical not morphological level, an that is why our primary starting point is the aspect of word-forming meaning.

The deverbative word-forming paradigm is defined as a class of verbs with derivatives of the same categorial meaning. Within the paradigm, the derivatives of the same categorial meaning may differ from each other due to their concomitant features and formally due formants. Word-forming paradigms differ from those current in inflexional morphology in that they have not only a vertical dimension; they are characterized both vertically and horizontally (it follows from the fact that the semantic space is modelled by means of them). In tabular notation, on the horizontal axis at the head of the table are given all the categorial meaning which are derived from the given verb, in the vertical direction are given verbs from which the particular categorial meanings are derived. The derivative expressing the given meaning is entered in the intersection of the verb and the categorial meaning. In the description of the deverbative word-forming system by means of paradigms, we in fact face a confrontation of word families of verbs, however, with certain limitations. In the horizontal direction, the deverbative word-forming paradigm is delineated in such an extent that only the substantives and adjectives are taken into consideration (the adjectives are investigated with the purpose of making possible the determination of the occurrence of the secondary deverbative substantives expressing the bearer of the quality of the action and of the substantivized quality of the action, that is of the abstract nouns in *-osť*). The derivation of verbs from verbs could have been excluded because we are considering only the derivation by means of sufixation and conversion and not the derivation by means of prefixation to which the derivation of verbs from verbs predominantly belongs.

The categorial word-forming meanings and hence categorial features of deverbative nouns are determined according to the intention of verbal action. The investigation of the intention of verbal action has become a traditonal field of interest of Slovak linguistics; up to now, however, this quality of the verb has been investigated from the point of view of syntax only.[4] The investigation into the deverbative derivation gave convincing proofs that the intention of verbal action, besides reflecting the semantics of the verb in a most abstract form, is projected in a certain manner not only into the valency qualities of the verb and on the syntactic level, where the members of the intentional system, that is the agent, the action and the patient, are in certain correspondence with sentence elements and with types of sentences, but also into the word-forming system in such a way that certain verbs can express the agent and the patient by word-forming means, whereas other verbs cannot. Of course, the situation in the word-forming system is a more complex one, since in addition to the agent, action and patient acting as word-forming meanings, the patient (that is

[4] Cf. E. PAULINY, Štruktúra slovenského slovesa, Bratislava 1943; Morfológia slovenského jazyka, Bratislava 1966; J. ORAVEC, Väzba slovies v slovenčine, Bratislava 1967; J. RUŽIČKA, Valencia slovies a intencia slovesného deja, Jazykovedný časopis, *19*, 1968, pp. 50—56.

the substance affected) is modified here in various ways, and, moreover, subsidiary intentions, such as the intention towards instrument and place, also come to act as word-forming meanings. The number of categorial deverbative meanings is stated as the sum of all the possible participants on the action present in all the intentional types of verbs. They are

1. the agent of the action (with the bearer of the action as a distributionally conditioned variant of agent in verbs of state),
2. the instrument of the action (the substance with which the action is exerted on its object),
3. the action,
4. the material used in the realization of the action,
5. the result of the action,
6. the residue of the action,
7. the object of the action,
8. the place of the action,
9. the quality of the action,
10. the substantivization of the quality of the action,
11. the bearer of the quality of the action.

This is then the delimitation of the paradigm in the inventory of categorial word-forming meanings, that is in the horizontal direction.

The vertical delimitation reveals which verbs belong to the same paradigm. The identity of the derived categorial meanings which has been mentioned in the definition of the paradigm, should not be taken too literally, since the lexis is not characterized by such consistency as the morphological paradigm, where the members of the same paradigm have an identical sum of inflected forms. In the lexis, it is not obligatory that certain meanings should be expressed by means of word-formation, only some tendencies of such expressing means can be observed. These tendencies can be revealed in the following way: in constructing the deverbative word-forming paradigm, we depart from an intentional type of the derivative bases (verbs), this type implying a potentional occurrence of certain categorial meanings. Where the potentially possible word-forming meaning is not realized because it is prevented by a concrete verbal meaning, the paradigm will show blanks which are current in semantic paradigms. The decisive fact is that the sum of more regularly derived categorial meanings is characteristic of a given intentional type and of a given paradigm.

Within the paradigms, the particular derivatives are arranged in oppositions according to their categorial meanings (if various categorial meanings are expressed by means of the same formant), and according to concomitant features (if various non-synonymic derivatives of the same categorial meaning are derived from the same word-forming base by means of different formants, for instance the Slovak words *písateľ* — *písár* are in an opposition according to the features of "occasional-

ness—constancy", which are realization of the opposition of "limitedness—non-limitedness" in nouns of agents).

Now, let us turn our attention to an outline classification of the deverbatives according to their paradigms. (The intentional types are presented according to *Morfológia slovenského jazyka*.) In addition to the criterion of the intentional type, some more concrete semantic criteria will also come in play.

In the first intentional type, the starting point substance — that is the agent — and the aim substance — that is the patient — are explicitly pronounced. This type includes object verbs expressing the external action of the agent. In the concrete meanings of these verbs, the subject or the object component of their intentional field can be stressed. Accordingly, these verbs can be divided into three word-forming paradigms.

The first paradigm includes verbs in whose meaning the affected component of the field of intention is stressed, which is reflected in word-formation in such a way that besides deriving the categorial agent of the action and the action from them, we can also derive those derivatives with categorial meanings whose concomitant feature is passivity, such as the result of the action, the material used, the residue of the action. The emphasis on the patient can be seen in the tendency to derive actional adjectives with passive meaning and passive potentiality by means of the formants *-teľný*, *-ný*. This paradigm includes verbs denoting various types of concrete production activities; these verbs act frequently as bases for the derivatives with the meaning of instrument and place. The object of these activities itself may vary, and therefore it is not expressed by word-forming means, but by lexical ones.

The second paradigm includes verbs acting as bases for the derivation of denominations with the categorial meaning of the agent of the action and the object of the action, but not other meanings connected with the external action of the agent.

The third paradigm includes verbs in whose meaning the subject component of the field of intention is emphasized, which, in word-formation, is reflected as the tendency of forming derivatives with the categorial meaning of the agent of the action, the action, and, in adjectives, of actional quality with the active meaning (with formants *-avý*, *-ivý*, *-ný*, with the polysemantic formant *-n-* used in active meaning here). There are no derivatives connected with the external action of the agent in this group.

The second intentional type 4th and 5th paradigm contains verbs in which the agent and the action are expressed separately, the patient is expressed implicitly with the action. The object of the action is the noun in the word-forming base (for instance, *dymiť* — to give off smoke), and therefore the derivative object of the action and the categorial meanings connected with it are not derived, since they are logically excluded. Denominations with the categorial meaning of the agent, the action, and the property of the action with potential intensification are derived from this type. Although the expression of the agent is possible from the point of view of intention, it is expressed derivatively only with verbs denoting such activity which can be

performed by a living subject, and such names of agents are expressive in character. Thus, they differ from the names of agents derived from the verbs of the first intentional type in that they have a concomitant feature of intensification and in their formants.

The third intentional type consists of verbs whose patient is not specified for being understood as identical with the agent, it merges with it. These verbs express the internal action of the agent. They include two basic semantic groups: verbs of motion (the 6th paradigm) and the verbs of location (the 7th paradigm). In word-formation, the difference between the verbs of the 3rd and of the lst intentional type (the internal and the external action) is relevant with more categorial meanings. The names of instruments derived from the verbs of the first intentional type have the concomitant feature of activity, they denote the instruments proper, which are an inevitable condition of the action (*rezačka* — cutter), whereas the names of instruments derived from the verbs of the 6th paradigm are not an inevitable condition of the action, they denote accesory instruments with a concomitant feature of passivity (*plavky* — swimming-costume). The names of instruments derived by means of the formant *-dlo* have a concomitant feature of place when derived from the verbs of the 7th paradigm (*sedadlo* — seat, *kľačadlo* — kneeling desk), but not so when derived from the verbs of the 1st paradigm. The names with the categorial meaning of the result of the action have, in the first paradigm, concomitant feature of concreteness and passivity (*rezanka* — chopped straw, *odliatok* — cast); the names with the categorial meaning of the result of the action in the 6th paradigm have a concomitant feature of abstractness and activity (*odchýlka* — deviation, *ústupok* — concession), and are correspondent with the names of action. It is characteristic for the verbs of the third intentional type that the following types of names are usually derived from them: names of agents (from verbs of location the names of the bearers of the action), names of action, partly, names of instruments, names of the quality of the action and abstract names in *-osť*. Intention to the place is their most prominent feature.

The fourth intentional type is represented by the verbs of state. The agent and the patient merge here and they are neutralized as regards the action. The neutralization of the agent and the patient to the subject which is the bearer of state eliminates the possibility of deriving such derivatives from the verbs of state that have the categorial meaning of the result and the object. Names of place can be derived from the verbs denoting the change of state and the constant physical state. These names of places bear a concomitant feature of resultness (*opuchlina* — swelling, *prasklina* — crack), thus they differ both in the meaning and the form from the names of place derived from actional verbs (these have the formants of *-áreň*, *-ovňa*, *-isko*/*-ište*). In states (conceived as static actions) the meaning of the action may merge with the meaning of the substantivized quality of the action (with the meaning of abstract nouns in *-osť*, *-stvo*).

The derivative capacity of non-personal verbs is rather limited. Only the verbs

denoting atmospheric phenomena admit of deriving the meaning of the action, with a facultative concomitant feature of intensification (*blýskavica* — lightning, *kĺzavica* — slide).

We depart from the word-forming bases in the description of the word-forming system by means of word-forming paradigms. This procedure is complementary to those starting from the formant, and it may be considered a contribution to them, especially because it shows by what the meaning of a polysemantic formant and its distribution are conditioned. Thus, for instance, the formant *-dlo* forms, when used with a verb of the first intentional type, the meaning of the instrument, and when used with the verbs of the fourth intentional type, the meaning of the place. The formant *-ok* with the verbs of the first intentional type forms the meaning of the result of the action, with the verbs of the third intentional type, the meaning of the action (*ústupok* — concession), with the verbs of the fourth intentional type — the meaning of the bearer of the action (*výrastok* — stripling, youngster), etc.

The results of such paradigmatic classification may be used in the generative description of word-forming systems, in constructing the rules of forming derivatives with a certain meaning from word-forming bases with a certain meaning.

Die semantische Struktur der primären Präpositionen in der slowakischen Sprache

JÁN ORAVEC, BRATISLAVA

I

Weder den Grammatiken, noch den Wörterbüchern einiger Sprachen gelang es bisher, die abstrakte Struktur der Beziehungen zu finden — die alle, das heißt sowohl die primären, wie auch die sekundären Bedeutungen der primären Präpositionen verbinden. Die Grammatiken sowie auch die speziellen Studien bleiben bisher bei der Suche nach der Struktur der Bedeutungen nur bei der konkretesten Bedeutung der Vorwörter — bei der räumlichen (spazialen) Bedeutung. Diese bestimmen sie ausführlich den geometrischen Koordinaten nach — nach der horizontalen und vertikalen Achse — und die anderen Bedeutungen lassen sie außer Acht. Darum halten wir es für nützlich, auf die einigenden Prinzipien hinzuweisen, auf denen der semantische Bau aller Bedeutungen einer primären Präposition beruht und sich auch bei anderen primären Vorwörtern wiederholt.

Als primäre werden vom synchronischen Standpunkte aus die Präpositionen angesehen, welche heute in keiner anderen als der präpositionalen Funktion verwendet werden, und wenn doch, dann nur in einer Ellipse, z. B. in Wendungen: *Kto je za a kto proti?* (Wer ist dafür und wer dagegen?). Auf Grund dieses Kriteriums gehören im Slowakischen zu den primären auch die Präpositionen: *okrem* (außer), *proti* (gegen), *medzi* (zwischen), so daß es im Slowakischen 22 nichtzusammengesetzte primäre Präpositionen gibt. Ungefähr 20 primäre Präpositionen haben auch andere slawische und indoeuropäische Sprachen.

Noch vor der Feststellung der Bedeutungsstruktur müssen wir die Frage beantworten: ob es eine Hierarchie der Bedeutungen gibt, das heißt, ob man berechtigt ist, von primären und sekundären Bedeutungen des Vorwortes zu sprechen, und weiter — ob auch die sekundären Bedeutungen hierarchisch geordnet werden. Beide Fragen beantworten wir positiv. Man hätte der Sprachwirklichkeit ein wichtiges Element entzogen, wenn man die Hierarchie der Bedeutungen nicht gesehen hätte. Wir sind der Ansicht, daß es notwendig ist, die primären und sekundären Bedeutungen zu unterscheiden. Und weiter sollten im Rahmen der sekundären Bedeutungen die

paradigmatischen und syntaktischen Bedeutungen unterschieden werden. Beide Fragen werden wir ausführlicher erklären.

Weniger problematisch ist die Frage, welche der Bedeutungen die Grundbedeutung der Präposition ist. In den älteren Studien hielt man für Grundbedeutungen auch mehrere Bedeutungen, z. B. die Bedeutung der Symmetrie, der Transitivität, der Konnexität usw. Heute überwiegt die Ansicht, daß die räumliche Bedeutung die Grundbedeutung der Präposition ist. Manche Autoren halten sie für die Grundbedeutung via facti, andere schreiben wohl darüber, aber am häufigsten ohne ihren Standpunkt zu begründen, oder sie begründen ihn sprachwissenschaftlich nicht angemessen. Wir stellen fest, daß die räumliche Bedeutung die Grundbedeutung ist,

1. weil man sie zum Unterschied von den anderen auch ohne den Kontext begreift,

2. weil sich aus dieser Grundbedeutung die meisten sekundären Bedeutungen entwickelten,

3. weil diese Bedeutung alle primäre Präpositionen mit Ausnahme der Vorwörter *s* (mit), *bez* (ohne), *okrem* (außer) haben. Das Primat der spazialen Bedeutung beweisen auch weitere wichtige, wenn auch weniger evidente Fakten, die eine ausführlichere Erklärung erfordern. Schon der Terminus *lokale* Bedeutung, den man in vielen Grammatiken und Wörterbüchern benutzt, braucht eine Erklärung und Verbesserung, weil er nur einen Aspekt der Grundbedeutung betont, das heißt, den statischen Aspekt. Deshalb ersetzen wir ihn durch einen breiteren Terminus — *räumliche* (spaziale) Bedeutung, weil dieser Ausdruck beide antithetischen Aspekte umfaßt: a) die statische (lokale) and auch b) die richtungweisende (dynamische) Hinsicht. Die beiden Aspekte kann man durch Interrogativpronomina charakterisieren: *kde* (wo)? — die Lokation; *kam* (wohin)? — die Direktion (den Richtungshinweis). Die Opposition Lokation—Direktion ist die wichtigste Opposition der präpositionalen Bedeutungen. Sie bezieht sich auf die absolute Mehrzahl der Vorwörter und ihrer Bedeutungen. Leider unterschätzt man in Wörterbüchern und in Grammatiken oft diesen Gegensatz, z. B. werden in Wörterbüchern die Bedeutungen der Präposition mit Akkusativ und Lokal, mit Akkusativ und Instrumental gemeinsam verarbeitet. So wird die richtungweisende Hinsicht mit der statischen gemischt und so wird das Wesen der semantischen Struktur, der Bereich und die Anordnung einiger Bedeutungen verdunkelt.

Im Slowakischen wird der Gegensatz Direktion—Lokation auch mit Hilfe der grammatischen Form, durch den Kasus ausgedrückt. Den statischen Aspekt der spazialen Bedeutung drücken nur jene Präpositionen aus, welche ursprünglich die lokalen Kasus, das heißt den Lokal und Instrumental regierten. Den richtungweisenden Aspekt drücken hingegen die Präpositionen aus, welche den Akkusativ, Genitiv und Dativ regieren.

Jeder der Bestandteile der Grundopposition wird weiter auf Grund einer niederen Opposition geteilt: mit Kontakt und ohne Kontakt. Der lokale Bestandteil zerfällt so in die Vorwörter, die den Lokal regieren — das ist die kontakte Lokation (Präpositionen *v* — in; *na* — an, auf; *o* — von; *po* — nach) und in Vorwörter, die den Instrumental regieren — das ist die distante Lokation, die noch genauer situiert wird (Präpositionen *nad* — über; *pod* — unter; *za* — hinter; *medzi* — zwischen). Der Richtungsbestandteil zerfällt in Vorwörter, die eine Richtung mit Kontakt ausdrücken (den Genitiv regierende Präpositionen *do* — in; *od* — von; *z* — aus) und Präpositionen, die eine Richtung ohne Kontakt ausdrücken (die den Dativ regierenden Präpositionen *k* — zu; *proti* — gegen). Der Akkusativ als der allgemeinste Richtungskasus enthält die beiden Bestandteile dieser niederen Opposition. Er drückt die kontakte Funktion (Kontaktivität) (Vorwörter *o* — von; *po* — nach; *v* — in) und auch die distante Funktion aus (*nad* — über; *pred* — vor; *za* — hinter; *medzi* — zwischen).

Von diesen drei richtungweisenden Kasus ist der Dativ am wenigsten richtungweisend, denn bei den den Dativ regierenden Präpositionen mischt sich zu den richtungweisenden Bedeutungen auch eine Nuance der statischen Bedeutung. Das Vorwort *k* (zu) hat die Bedeutung „zur Stelle in der Nähe von etwas (hin)" und die Präposition *proti* (entgegen) steht nicht nur mit richtungweisenden Zeitwörtern (z. B. *ísť proti niekomu* — jemandem entgegenkommen), sondern auch mit statischen Zeitwörtern (z. B. *stáť* — stehen, *ležať* — liegen, *proti niekomu* — jemandem gegenüber).

Auch der Genitiv ist nicht so sehr Richtungsfall wie der Akkusativ. In seinem Randbereich stehen auch solche Präpositionen, die weder Richtungscharakter noch irgendeine spezielle Bedeutung haben. Die Präposition *za* drückt nur die Zeit, die Präposition *bez* die Bedingung und die Präposition *u* die zueignende Einbeziehung aus; letztere wird nur bei Personennamen gebraucht.

Wenn wir diese Überlegungen zusammenfassen, zeigen sich uns für die Charakteristik der Grundbedeutung zwei weitere Zeichen und zwar:

4. Die Grundbedeutung enthält nicht nur die meisten Bestandteile, sondern auch die größte Zahl der Oppositionen zwischen den Bestandteilen, besonders:

a) die Grundopposition Richtungweisung — statische Bestimmung,

b) die niedere Opposition kontakte Funktion (Kontaktivität) — distante Funktion (Distantivität).

5. Die erwähnten Oppositionen werden auch formel ausgedrückt.

Auch die temporale Bedeutung (die manchmal mit der räumlichen verknüpft wird) hat nicht so viele Bestandteile und Oppositionen wie die räumliche Bedeutung. Die Oppositionen verschwinden hier, weil ihre Bestandteile neue Nuancen bekommen. Das hängt mit der außerlinguistischen Wirklichkeit zusammen: die Zeit fließt in eindimensionaler Richtung, darum verengt sich auch ihre Ausdrucksweise auf Mittel, die räumlich auf einer Achse liegen und nicht auf drei Achsen, wie die räumlichen Beziehungen. Aber trotzdem wird auch die temporale Bedeutung nicht ganz von der Grundbedeutung getrennt, weil:

1. dabei wesentlich beide Bestandteile der Grundopposition erhalten bleiben, wobei modifiziert wird als Dynamik — Statik (z. B. *do tejto chvíle* — bis zu diesem

Augenblick; *v tejto chvíli* — in diesem Augenblick) und die Bestandteile der niederen Opposition Kontaktivität — Distantivität (z. B. *pred 12. hodinou* — vor 12 Uhr, *o 12-ej* — um 12 Uhr).

2. Auch aus der temporalen Bedeutung werden einige sekundäre Bedeutungen (wie z. B. die kausale) — wenn auch nur selten — gebildet.

II

Beim Entstehen der sekundären Bedeutungen der Präpositionen betont man gewöhnlich den Einfluß des Kontextes. Anderseits ist es aber nicht richtig, wenn man den Einfluß der Bedeutung des Vorwortes selbst auf das Konstituieren (Entstehen) der sekundären Bedeutungen nicht sieht, weil auch die Präpositionen eine oder mehrere Bedeutungen hat. Die Präposition ist zwar ein synsemantisches Wort, aber kein leeres Morphem. Anders wäre es nicht zu erklären, warum in einem bestimmten Kontakt ein Vorwort eine bestimmte Bedeutung hat, z. B. kausale, während ein anderes diese Bedeutung nicht hat.

Einen entscheidenden Einfluß bei der Entstehung der Mehrheit der sekundären Bedeutungen hat die Grundbedeutung der Präposition selbst. Sie ist richtungweisend oder nicht richtungweisend, also statisch. Die sekundären Bedeutungen sind zum Unterschied von der Grundbedeutung wesentlich begrenzt. Der Grundgegensatz Richtungweisung — Statik widerspiegelt sich in den sekundären Bedeutungen derart, daß einige sekundäre Bedeutungen nur auf einen der Bestandteile begrenzt sind: die einen werden auf dem richtungweisenden Bestandteil dieses Gegensatzes aufgebaut, die anderen wieder nur auf dem statischen Bestandteil. Die Übergriffe sind vom strukturellen Blickpunkt aus nur von geringer Wichtigkeit. Auf dem richtungweisenden Bestandteil wurden im Slowakischen folgende Bedeutungen aufgebaut: die Bedeutung des Maßes, der Wirkung, des Zieles, der Ursache und die Funktion des Objektes. Auf den statischen Bestandteil gründen sich: die Bedeutung der Hinsicht, der Bedingung, der Einräumung (Konzessivität), die Funktion des sekundären Prädikates und im Grunde auch die modale Bedeutung. Der Anschaulichkeit halber zitieren wir die typischen Repräsentanten der ersten und der zweiten Gruppe.

I. Gruppe — die sekundären Bedeutungen, die ihren Grund in dem richtungweisenden Bestandteil haben.

a) Die Bedeutung des Maßes und der Wirkung drücken die den Akkusativ regierenden Vorwörter aus: *na* (auf), *nad* (ober, über), *po* (nach), *ponad* (über, herüber), *cez* (durch), *pod* (unter) und die Genitiv richtende Präposition *do* (in). Es handelt sich größtenteils um Vorwörter, die eine Richtung mit Kontakt im Ziele ausdrücken.

b) Die Bedeutung des Zieles drücken besonders die den Akkusativ regierenden Präpositionen aus: *na* (auf), *po* (für); ferner: das den Genitiv regierende Vorwort *do* (in) und die Dativ regierende Präposition *k* (zu).

c) Die kausale Bedeutung wurde größtenteils auf der richtungweisenden Achse konstituiert und zwar bei den den Akkusativ regierenden Präpositionen *pre*, *za* (für), bei den den Genitiv regierenden Vorwörtern *od* (von), *z* (aus), die größtenteils den Kontakt im Ausgangspunkte und eine konsekutive (zeitliche) Nachfolge ausdrücken. Es handelt sich um folgende Typen von Verbindungen: *vyhorieť za niečo* (für etwas üble Vergeltung erhalten), *urobiť niečo zo strachu* (etwas aus Furcht machen), *zomierať od hladu, smädu* (vor Hunger, Durst sterben).

d) Die Bedeutung des Objekts, das heißt die abstrakteste Bedeutung, erlangten alle richtungweisenden Vorwörtern, die den Akkusativ, Genitiv und Dativ regierten, und von den statischen Präpositionen jene, die sekundär eine richtungweisende Nuance bekamen, z. B. durch eine temporale Bedeutung der Folge: *bežať za niekým* (nach jemandem rennen). So entstanden auch die Objektverbindungen, z. B. nach den Verben: *pozerať za dakým/po dakom* (nach jemandem schauen); *túžiť za dakým/po niekom* (sich nach jemandem sehnen). Von der finalen und kausalen Bedeutung unterscheidet sich syntaktisch die Objekt-Bedeutung durch totale Unterordnung der Rektion des übergeordneten Verbs.

II. Gruppe — Bedeutungen, die auf dem nichtrichtunggebenden (statischen) Bestandteil entstanden.

a) Der räumlichen Grundbedeutung am nächsten steht die Bedeutung der Hinsicht. Es ist eine Bedeutung, welche die Gültigkeit der Aussage (des Ausdrucks) auf einen bestimmten Kreis der Erscheinungen, möglicherweise auch nur auf eine einzige Erscheinung begrenzt. Sie wird durch den Lokal regierende Präpositionen und zwar *pri* (bei), *na* (an, auf) ausgedrückt, beschränkt (bei Personennamen) auch durch das seltene den Genitiv regierende Vorwort *u* (bei).

b) Die semantischen Faktoren kommen am meisten bei der modalen Bedeutung zur Geltung. Aber auch dies kommt am häufigsten bei den nichtrichtungweisenden Präpositionen *s* (mit), *bez* (ohne) vor.

c) Es ist schwer von der modalen Bedeutung die Funktion des sekundären Prädikates zu unterscheiden — dies bei den nichtrichtungweisenden Präpositionen *s* (mit), *bez* (ohne) und bei den den Lokal regierenden Vorwörtern *v* (in), *na* (an, auf). Die Funktion des Prädikativums ist hier in dem erweiterten Ausdruck evident, wo der eine Teil das Prädikativum (also sekundäres Prädikat) ist, z. B. *vybehol bosý, s holou hlavou, bez kabáta a v gatiach* (er lief barfuß hinaus, mit unbedecktem Kopf, ohne Mantel, nur in Unterhosen).

d) Die Bedeutung der Bedingung drücken auch nur nichtrichtungweisende Präpositionen aus: die den Lokal regierenden *pri* (bei), *v* (in), *o* (von), *na* (an, auf), die den Instrumental regierenden *s* (mit), *pod* (unter), die den Genitiv regierenden *bez* (ohne), z. B. *pri teplote 100°* (bei hundert Grad Hitze, Wärme), *v núdzi* (in der Not), *s podporou* (mit Hilfe, mit Unterstützung), *pod tlakom* (unter Druck), *bez pomoci* (ohne Hilfe). Auf die Bedeutung der Bedingung knüpft sich die Bedingung der Einräumung, weil diese nur eine Art der Bedingung ist — eine nicht gültige Bedingung.

Darum drücken die Einräumung alle Vorwörter der Bedingung aus, aber gewöhnlich in Verbindung mit den betonenden Partikeln *aj*, *i* (ja, auch) — also in solchen Verbindungen wie: *aj pri teplote 100°* (auch bei 100 Grad Wärme), *i v núdzi* (ja, auch in der Not).

III

Beim Konstituieren der sekundären Bedeutungen unterschätzen wir auch den Kontext nicht. Die erwähnten Beispiele (Richtungweisung — Statik; Kontakt — Distantivität) und ihre grammatische Form (die richtungweisenden Kasus — die nicht richtungweisenden Kasus) bilden zwar den Grund des Systems der Präposition, aber in konkreten Bedeutungen modifizieren diesen Grund die lexikalischen Bedeutungen der den Präposition übergeordneten Wörter, manchmal wieder die Bedeutungen der Wörter, die auf die Präposition folgen, oder die Bedeutungen dieser beiden zusammen.

Bei den der Präposition übergeordneten Wörtern spielt die Wortart eine gewisse Rolle: ob es sich um ein Verb, ein Adjektiv, oder nur ein Substantiv handelt. Bei den Verben muß man unterscheiden, ob es sich um ein Verb mit Objekt oder um ein Verb ohne Objekt handelt. Den syntaktischen Charakter des mit einem Vorwort verbundenen Kasus beeinflussen auch kleinere semantische Gruppen von Verben, z. B. Zeitwörter, die eine Bewegung benennen — Zeitwörter, die keine Bewegung beschreiben, und ähnliches.

Die lexikale Bedeutung eines Wortes (des Substantivs), das auf eine Präposition folgt, verändert die syntaktische Geltung des Vorwortes bei den adverbialen Bedeutungen. Es ist maßgebend dabei, ob es sich um ein Konkretum ader Abstraktum handelt, weiter im Rahmen der Abstrakta, ob es sich um ein Deverbativum oder Deadjektivum handelt, um eine Benennung der Quantität oder um Benennung eines Zeitbegriffs. Den Einfluß des Kontextes können wir hier nicht umständlich beschreiben.

IV

Anstatt eines zusammenfassenden Abschlusses werden wir zeigen, wie die erwähnten theoretischen Ergebnisse die Beschreibung der konkreten Bedeutung präzisieren und zwar die Beschreibung der sekundären, finalen Bedeutung.

Im Slowakischen kann das Ziel durch 22 primäre und 12 sekundäre Präpositionen ausgedrückt werden. Auf den ersten Blick könnte dies den Anschein erwecken als ob es hier keine Ordnung gäbe, und daß die These von dem richtungweisenden Grund der Zielbedeutung nicht haltbar ist. Aber schon beim Betrachten der Frequenz der Zielbedeutung stellt man fest, daß alle nicht richtungweisenden Präpositionen das

Ziel nur vereinzelt — als eine Nuance anderer Bedeutungen ausdrücken. Es handelt sich um eine okkasionelle Verwendung.

Das reine Ziel drücken nur sechs (6) Vorwörter aus und diese sind alle richtungweisende: die den Akkusativ regierenden Präpositionen *na* (an, auf); *po* (nach), die den Genitiv regierende Präposition *do* (in), das den Dativ regierende Vorwort *k* (zu), das den Lokal regierende Vorwort *po* (nach) und das den Instrumental regierende Vorwort *za* (hinter). Die den Lokal regierende Präposition *po* (nach) fällt ab, weil diese hier nur deshalb gebraucht wird, weil sie — zum Unterschied von anderen den Zweck bezeichnenden Vorwörtern — den Zweck unter eine breitere distributive Bedeutung reiht; ähnlich, wie sie dies mit anderen adverbialen Bedeutungen und anderen syntaktischen Funktionen (z. B. mit dem Objekt und dem Subjekt) tut. Vergleichen wir den Unterschied zwischen den finalen Ausdrücken *ísť na návštevu* (auf Besuch gehen) — *chodiť po návštevách* (Besuche machen). So bleiben nur fünf Präpositionen, die den Zweck als paradigmatische Bedeutung ausdrücken, während alle anderen es nur als syntaktische Bedeutung ausdrücken. Aber nicht alle diese fünf Vorwörter drücken denselben Zweck aus. Die den Instrumental regierende Präposition *za* (nach, für) drückt ein entferntes oder unsicheres Ziel aus; vergleichen wir den Unterschied zwischen: *ísť po vodu* (Wasser (zu) holen gehen) — *ísť za vodou* (Wasser (zu) suchen gehen). Ähnlich ist es mit der den Dativ regierenden Präposition *k* (zu), während den anderen, das heißt den den Akkusativ und Genitiv regierenden Präpositionen diese Eigenschaft fehlt. Diese werden dann weiter geteilt — je nach dem, wie sie vom Kontext begrenzt sind. In den erwähnten Erklärungen haben wir erwähnt, daß der Kontext die sekundären Bedeutungen in konkreter Form modifiziert. Der Kontext beteiligt sich auch an der Konstituierung der Zweckbedeutung. Die Zweckbedeutung drückt gewöhnlich Abstrakta (der Handlung) aus, ferner ihren Ersatz aus dem Gebiete der Konkreta, z. B. *pozvať na pohostenie* (*hostinu*) (zum Festessen /Gastmahl/ einladen) — *na pohár vína* (auf ein Glas Wein). Vom Standpunkte des übergeordneten Verbs ist die Zweckbedeutung so begrenzt, daß sie vorkommt:

a) nach Verben, die eine Bewegung ausdrücken, und zwar die Bewegung des Subjekts der Handlung oder die Bewegung des Objekts der Handlung (z. B. *poslať* — senden; *pozvať* — einladen),

b) nach Verben, die eine Vorbereitung und ihre Umstände ausdrücken (*chystať sa* — sich vorbereiten; *pomáhať* — helfen; *potrebovať* — brauchen).

Diese Kontextbedingungen beschränken am wenigsten die den Akkusativ regierende Präposition *na* (an, auf), darum halten wir sie für ein Grundvorwort des Zwecks. Die den Akkusativ regierende Präposition *po* (nach) drückt das Ziel nur bei Konkreta, das den Genitiv regierende Vorwort *do* (in) nur bei den Abstrakta aus. Vom Standpunkte des Verbs sind beide Präpositionen begrenzt, die das entfernte Ziel ausdrücken: *na* — nach, *k* — zu.

Die wichtigsten zielweisenden Präpositionen sind also durch zwei Merkmale unterschieden:

1. durch ein semantisches Merkmal: das entfernte Ziel — das spezifizierte Ziel;
2. durch syntaktische Begrenzung und zwar:

a) durch eine Begrenzung von Seiten des Substantivs (Abstrakta — Konkreta);

b) durch eine Begrenzung von Seiten des Verbs.

Ihr System stellt dieses Schema dar

Präpositionen	*na* an, auf	*po* nach, für	*do* in	*k* zu	*za* hinter
1. das entfernte Ziel +	—	—	—	+	+
2. a) Beschränkung von Seiten des Nomens +	—	+	+	—	—
b) Beschränkung von Seiten des Verbs +	—	—	—	+	+

So kann man alle Bedeutungen der Präpositionen beschreiben. Manche haben wir schon so bearbeitet.

Bei der Beschreibung der Bedeutungsstruktur der primären Vorwörter gingen wir von den Eigenschaften aus, die alle Präpositionen und alle Bedeutungen haben. Wir gingen bis zu der höchsten Abstraktion, das heißt: bis zu den binären Gegensätzen. Aus diesem Grunde gaben wir die primäre — spaziale Bedeutung auf und wir erklärten ihre strukturelle Verbundenheit mit den sekundären Bedeutungen. Bei der spazialen Bedeutung schreiben wir die größte Wichtigkeit der Wirklichkeit zu, daß sich diese Bedeutung in einen richtungweisenden und einen nicht-richtungweisenden Bestandteil teilt. Diese beiden Bestandteile werden formel (grammatisch) durch die Kasus fixiert. Die Opposition Richtungweisung — Statik bildet die Grundachse auch bei der Teilung der sekundären Bedeutungen. Ein Teil der sekundären Bedeutungen ist auf dem richtungweisenden (dynamischen) Bestandteil dieser Grundopposition aufgebaut, andere auf dem statischen Bestandteil.

A Model of Morphemic Description of Russian Words

ZDENĚK F. OLIVERIUS, PRAHA

Introduction

Frames of reference within which analysts approach language systems differ considerably in minor points while retaining a few almost identical characteristics which allow to classify descriptive models into a comparatively small number of relatively distinct ones.

Even the three types around which most grammatical descriptions seem to cluster about according to Charles F. Hockett's *Two models of Grammatical Description* [9] have some characteristic features in common. Item and Arrangement, Item and Process and Word and Paradigm can all be reduced to two principal components: 1. an inventory of basic units and 2. rules governing their arrangement into given structures. Transformational grammar, in fact, also involves two basic components: a set of given elements and a set of rules.

The usefulness of an *Elements and Rules* model of morphemic description does not consist only in the very considerable simplification of analysis and description, but also in close affinity to the discrete character of language. The words are thus viewed as strings of discrete elements, morphemes, which in their turn are composed of discrete units, phonemes (on the expression plane) and distinctive semantic components (on the plane of content).

The purpose of this paper is not to offer a totally new concept for morphemic description, but rather to systematize notions serving as basis for establishment of principal elements and rules. What may be considered as new in this approach especially if compared with other papers on morphemic analysis of Contemporary Standard Russian is the deliberate emphasis placed on the plane of content in operations leading to constituting morphemes as classes of allomorphs having identical sets of distinctive semantic components.

Morphemic analysis has been applied to the lexicon of Slavonic languages with increasing frequency in recent years, especially in the U.S.S.R., the United States and Czechoslovakia.

Pioneering work has already been done in compiling dictionaries of morphemes

[22], [23], [32], in exploring the possibility of analysing main rules of phonemic constitution and alteration of morphemes [11], [20], [25], [28], [29], [30], and in approaching the main problems of a full description of morphemic valency and alterations [10], [21], [28], [31].

But despite many notable achievements, morphemic analysis of Contemporary Standard Russian arrives at an impasse at several crucial points. There is at present an acute need for a serious attempt to introduce adequate methods of semantic analysis of Russian morphemes, allowing to solve otherwise insoluble problems of interfixes, enlarged suffixes and assigning allomorphs to morphemes.

1. Elements and rules

Word-formational analysis examines the relation of foundation as a relation of two words (founding and founded) [4]. For the word-formational analysis each word is constituted by two elements only: basis and formant.

Morphemic analysis goes further and examines the morphemic constituency of words from a broader point of view (irrespectively of immediate foundation). For morphemic analysis words are constituted by a chain of morphemes [19].

The purpose of morphemic analysis starting from the text as its datum is to establish the inventory of morphemes and their semantic components as well as their phonemic constitution and to reveil the hierarchy and valency of morphemes together with semantic and phonemic alterations.

Contemporary Standard Russian with its increasing tendency to agglutination [21] and with its comparatively transparent morphemic structure seems to be a rather suitable material for this kind of morphemic analysis. However, objections might be raised and counter-examples could be found to disprove the validity of the assumption that agglutination prevails in morphemic structure of Russian words. A feeling of dissatisfaction can be avoided if it is born in mind from the very beginning that apart from endocentric word structures there are also exocentric ones in Contemporary Standard Russian, and that there is and undoubtedly always will be a certain amount of Russian words which cannot be semantically deciphered on the morphemic but only on the lexemic level. There is no question about the technique of synthesis of Russian words, which is for the most part agglutinative: the solely questionable matter might be the relevancy of the extremely limited number of counter-examples. Whether different types of morphemic description of Contemporary Standard Russian that have been developed up to now are adequate and effectively formalized is, of course, another matter. It seems to me that it is necessary to devise a formalized version of a morphemic description model, incorporating a certain amount of formalized semantic analysis.

The present characterization of an Element and Rules Model of morphemic

description of Contemporary Standard Russian will include a general outline of some problems connected with segmentation of Russian words and assignment of allomorphs to morphemes, a survey of some of the problems implicit in rules governing not only valency and distribution of Russian morphemes but also alteration of semantic components and phonemic constitution of morphemes in broad enough terms to cover some general questions.

2. Elements

The requirements of self-consistent, exhaustive and simple description can best be satisfied if unnecessary elements or notions are not introduced. In rather numerous and contradictory descriptions of Contemporary Standard Russian morphemics notions like *empty morphemes*, *interfixes*, *portmanteau morphs* etc. [21], [24], [34], are sometimes used. What is the nature of these elements? It is of primary importance for us to investigate the status of empty morphemes, interfixes, portmanteau morphs, morphonemes etc. from the point of view of language as a sign system. The terms *discrete* and *continuous* have become in our day expressions to conjure with; but even if they are mentioned and implied rather often in general linguistics and phonology, they are, nevertheless, seldom referred to and applied in morphology. In his interesting paper [14], B. Mandelbrot has shown the necessity for discrete units on all levels of language systems. Discrete units, unlike continuous elements are liable to binary decisions: they either *are* or *are not* identical, tertium non datur. They cannot be more identical or less identical. Interfixes and morphonemes as accepted in some morphemic descriptions of Russian are intermediate cases between discrete and continuous phenomena which contradicts the postulate of necessity for discrete units.

Empty morphemes (empty morphs, pustyje morfemy, prázdne morfémy) [24], "have no meaning and belong to no morpheme" [8]. Apart from phonemes, which have the status of figurae in L. Hjelmslev's terminology [7], all discrete elements in natural languages have the status of signs, which are bilateral having expression and content. The information conveyed by a morpheme can converge to zero, i.e. can approximate towards zero, but cannot reach it — by definition — then the morpheme loses its morphemic status and becomes a part of another morpheme or morphemes.

In essence it seems desirable to follow more precise techniques to identify morphemes avoiding any confusion of non-discrete and discrete elements.

Suffice it to say at this point that to keep the analysis of the plane of expression and the plane of content in balance may prove essential for solving many of the above-mentioned problems.

2.1 Semantic components

There is general agreement among linguists that morphemes are minimal, bilateral units. Every sequence of phonemes which has meaning and cannot be segmented into smaller units having meaning is assigned status of morpheme.

There are sequences of phonemes in Russian words as e.g. /*al,*/, /*en,*/, /*j*/, /*vl,*/ in words: *t,eatral,nij*, *vr,em,en,i*, *l,ist,ja*, *davl,u*, which can be described as having certain meaning. Nevertheless for some reason or another these cases are regarded dubious and therefore these morphemes are labelled as interfixes, empty morphemes, morphonemes, etc.

The segment /*j*/ in the Russian word *l,ist,ja* is sometimes supposed to convey the meaning of "plural", which seems to some linguists sufficient reason for assigning morphemic status to it. This assumption is based on contamination of distinctive and non-distinctive semantic components. To draw a line between distinctive, relevant, criterial, essential etc. components and non-distinctive, irrelevant, non-criterial, accidental etc. ones seems essential to some linguists (Bendix, Lounsbury). Distinguishing between distinctive and non-distinctive semantic components is pertinent to the establishment of morphemic status of segments of words, otherwise one ends with as many morphemes as allomorphs: each allomorph inevitably conveys a slightly different information than all other allomorphs, cf. e.g. *knife* versus *knive-*, the former excluding plural, the latter singular. The segment /*l,ist,-*/ has the non-distinctive semantic component "singular" conveyed by the segment /*l,ist*/ in words *l,ist*, *l,istom* etc. while desinences *-a-*, -0-, *-om-* have distinctive semantic components "plural" or "singular" respectively. The same applies to similar segments mentioned above as examples.

Semantic componential analysis revealing distinctive values of semantic variables (as e.g. "genus", "number", "countability" etc.) allows for a description of morphemes as particular combinations of these values and can substantiate also the existence of the so-called *zero morphemes*. Two segments /*škol*$_1$/ and /*škol*$_2$/ taken from words *škola* and *škol* have identical sets of distinctive semantic components with the only exception of values of two variables "case" and "number": /*škol*$_2$/ has distinctive components corresponding to values "genitive" and "plural" while /*škol*$_1$/ does not have any values of variables "case" and "number". This may lead either to accepting /*škol*$_2$/ as a portmanteau morpheme or — more logically — to assuming a zero allomorph characterized in the plane of content by the above-mentioned values as distinctive semantic components and in the plane of expression by a zero sequence of phonemes as its primary designator and its position as its secondary designator.

Introducing the notion of zero into mathematical history was one of the greatest achievements of human thinking. Redundancy of a natural language is a necessary condition for the presence of zero elements in that language. In an ideal code, without

redundancy, signs have only one type of designator, in natural languages, they have two types of designators.

There is a general tendency among lexicographers, specialists engaged in word-formational analysis and authors of grammars of Contemporary Standard Russian to give rather narrow and unduly concrete semantic definitions of Russian morphemes. In Russian explanatory dictionaries we find following types of definitions: "*l,d,ina*" — "*gliba l,da*"; "*č,ern,ič,ina*" — "*jagoda č,ernik,i*"; "*k,ievl,an,in*" — "*žitel goroda K,ieva*" etc., which leads to splitting the segment *-in-* into a series of homophonemic morphemes with different meanings. Another striking example of the way Russian suffixal morphemes are usually treated is grouping the suffix {*ic*} in words "*car,ica*", "*imp,eratr,ica*" with suffixes denoting persons, but the same suffix in words "*l,v,ica*", "*volč,ica*" with suffixes denoting animals (cf. [28]). "Person" or "animal" are not distinctive semantic components of the morpheme {*ic*}, and simlarly values as "block", "berry" or "resident" are not distinctive components of the morpheme {*in*}. The semantic testing obviously should not stop at one or two concrete examples, at one or two possible interpretations or simple transformations, but must somehow encompass all potential contexts, interpretations or simple transformations, i.e. take all possible morphemic variables (environments) into question. Following such a broad procedure the investigator excludes all non-distinctive semantic components one by one and ends with a few relevant, distinctive ones, allowing for a rather abstract semantic interpretation, covering the whole range of distribution of the given morpheme.

2.2 Basic form of morphemes

In the expression plane allomorphs are constitued by sequences of phonemes. Archiphonemes or morphonemes as elements with a limited distinctive power contradict the postulate of discreetness of basic linguistic elements. In essence archiphonemes are a statement about linguistic vagueness, about resoluble or irresoluble syncretisms. Many instances of syncretism on phonemic level seem to be irresoluble simply because of artificial limitations of analysis to inherent acoustic features of given segments without taking recourse to broader environment, which would enable to resolve phonemic syncretisms and use phonemes instead of archiphonemes or morphonemes (in the narrow sense of that word) as elements constituting allomorphs.

The point of departure of this reasoning is the obvious fact of greater than zero redundancy in actual communicative systems (with the necessary implication of a) absolute limitations on sequence of segments and b) variations in relative frequency of different sequences).

All Russian allomorphs can thus be shown to be constituted by sequences of phonemes. Semantically equivalent allomorphs can be grouped into classes usually called morphemes. For the purpose of simplifying certain operations within the

Element and Rules Model it is convenient to find a sequence of phonemes which may, but need not, coincide with one of the allomorphs of the given morpheme and treat this sequence as the basic underlying form from which all other forms can be derived by means of a set of rules.

The idea of a basic form or structure from which other forms or structures can be deduced or derived is far from being new in linguistics: word and paradigm, vocables in dictionary articles and derived words, founding and founded words in word-formation are only a few examples. An example which was recently rather popular might be added: kernel sentence structures and their transforms. "Theoretical basic form" or "artificial underlying form" mentioned in 1933 by L. Bloomfield [2] was applied to morphological analysis of Russian verbs by R. Jakobson in 1948 [11]. Only recently linguists started to make fuller use of this rewarding notion [28]. The basic form, perhaps the most explicit of existing ones, allows a derivation of all other forms by means of a set of rules with clearly defined domains of applicability. The choice of the most explicit form as the basic form of a morpheme has its justification in the fact that truncation and alteration rules are more easily formulated than completion or addition rules.

3. Rules

In different positions morphemes undergo certain modifications which might be explained as the result of interaction between the given element and its position. It seems generally accepted that the element in phonological analysis is a segment characterized by articulatory and acoustic features and that the position includes neighbouring segmental and suprasegmental features, while in morphology two very important facts are usually disregarded: firstly, morpheme alternants in linguistic analysis are normally understood to be alternants of phonemic constitution of morphemes, alteration of content is ignored, secondly, alteration of phonemic constitution is not defined in terms of morphemes as elements but in terms of phonemes constituting them. This approach may reflect adequately certain instances of alteration based entirely on diachronic phonemic changes but fails to account for other alterations which are rather numerous in Russian and certainly fails to indicate fully the domains of applicability of rules formulated in terms of components of elements instead of elements. Consequently the amount of exceptions is unnecessarily high.

To cover the morphotactical pattern of Contemporary Standard Russian it is necessary to have three sets of statements:

1. rules governing valency and distribution of morphemes,
2. semantic alteration rules,
3. morphonemic rules.

3.1 Valency and distribution

For the purpose of the present discussion the term valency will be used in a sense corresponding with competence while distribution will be used as corresponding to performance. The distribution of a morpheme will be understood as a sum of all its environments (i.e. all other morphemes, each in its particular position, with which the given morpheme occurs to yield a word of Contemporary Standard Russian). Valency is to be understood as a set of potential environments with which a given morpheme can (under certain circumstances) yield a new Russian word. To proceed from the stage of distribution (i.e. from the stage of a sum of environments) to the stage of valency we have to formulate an abstract definition covering the whole domain of potential environments including both words already existing in the Russian language and potential words which have not yet emerged into existence and perhaps never will.

In natural languages, with a certain degree of redundancy, only some combinations of elements are possible, while others are excluded. It is practical to distinguish and keep apart realizable and unrealizable combinations of elements on different levels.

On morphemic level the assumption of unrealized combinations of elements, empty slots in the morphemic system, may allow for a more detailed segmentation ending in a lower amount of elements. Many of the so-called enlarged morphemes (suffixes etc.) [28] which cannot be simplified by application of procedures of word-formational analysis, based on distribution and not valency, may be split into two or more morphemes: {*n,ik*}, {*č,ik*}, {*ovšč,ik*} "*pomošč,n,ik*", "*putn,ik*", "*v,estn,ik*"; "*gruzč,ik*", "*sč,otč,ik*", "*pul,em,etč,ik*"; "*lampovšč,ik*", "*č,asovšč,ik*" etc.

Until recently there have been very few attempts at a complete description of distribution and valency of Russian morphemes. Attention should be drawn to a notable attempt to express word structure by means of a theory of graphs. J. Horecký [10] shows how a set of Slovak morphemes may be understood as a graph, the crossing points of which are separate morphemes constituting Slovak words.

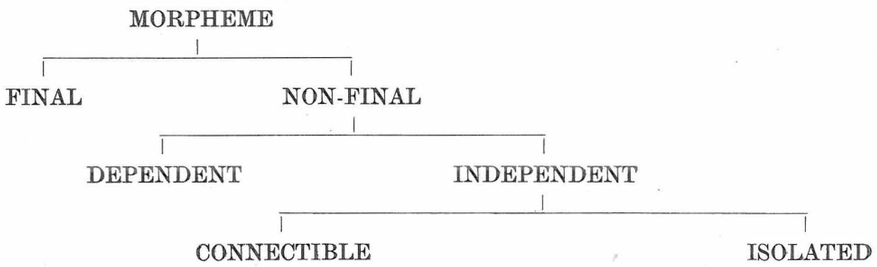
3.1 Types of morphemes in ČSSR

Positions inside word boundaries are essential for establishment of morpheme types. Word-final position as opposed to all other positions in a word determines a type of morphemes endowed with purely relational content. Some questions pertaining to problems of classification of parts of speech may be raised in connection with listing final morphemes — it is needless to mention that there is no general agreement among Russianists in listing desinences, certain morphemes are sometimes classed as desinences, sometimes as suffixes, cf. {*l*}, {*t*}, {*vši*} etc. Can words differing in

pre-final as well as in final morphemes be grouped together as one lexeme (which is the usual practice) or shall a lexeme be defined as a group of words differing only in desinences? The latter possible solution is in my opinion a more logical one, drawing a distinct line between inflection and derivation.

Non-final morphemes are clearly divided into independent and dependent ones. The class of independent morphemes unites both roots, which may occur as single morphemes within two word boundaries (final morphemes do not count), and uninflected parts of speech (prefixes are united with prepositions on the basis of identical sets of distinctive semantic components). The former type of independent morphemes may be called connectible, the latter — isolated. Dependent morphemes do not occur as single morphemes within word boundaries, they only accompany independent morphemes, cf. {*ak*} in words +*k-ak*-0+, +*k-ak-oj*, {*k*} is obviously an independent morpheme, cf. +*k–ogo*+. Test words for morphemes are those with the smallest possible number of morphemes within word boundaries: the Russian word *kakoj* cannot be accepted as test word for {*k*}, while the word *kogo* consisting (apart from a final morpheme) of only one morpheme is evidently an acceptable test word.

We obtain thus a hierarchy of morpheme types departing in a few points from the usually accepted classification of morphemes in Russian:



Grouping prefixes together with prepositions, conjunctions etc. under the heading of independent isolated morphemes clearly shows a distinct difference between suffixes and prefixes in traditional terminology.

Prefixes are less numerous both as regards the average number per word and as regards the upper limit of occurrence in one single word, than suffixes. In addition to that it is worth noticing that the average number of prefixes per word coincide to a large extent with the average number of roots preceding the final root in compounds and that the character of junctures between prefixes and roots and two roots respectively is practically the same.

Classes of morphemes are based on semantic components and distribution and valency of morphemes.

3.2 Semantic rules

Classifying morphemes like {*ic*} with suffixes denoting persons and at the same time with suffixes denoting animals depending on its occurrence after morpheme strings denoting either persons or animals is evidently a contamination of semantic interpretation of the given morpheme and semantic interpretation of the whole word of which the morpheme is a part. The content of Russian morphemes, especially the dependent ones can be shown to contain only a few semantic components making the semantic load of these elements rather abstract. More concrete meaning is arrived at in an interplay of semantic components of all morphemes constituting the given word.

(I deliberately do not touch further concretization of meaning in the process of interplay of contents of words within sentence boundaries.)

The process of establishing word content on the basis of interaction of morpheme contents within word boundaries may be well illustrated by words containing the Russian singulative suffix {*in*}.

The only distinctive semantic component characterizing the Russian morpheme {*in*} can be interpreted as "countability", which also accounts for its distribution and valency only after stems of mass words, adjectival stems and other strings of morphemes comparable with them on the basis of uncountability. The Russian morpheme {*in*} occurring with morpheme strings as {*č,ern,ik-*, *brusn,ik-*, *z,eml,an,ik-*} give rise to a semantic component to be interpreted as "berry". It can be argued that this semantic component is originally present in the morpheme {*in*} as an alternative non-distinctive component and emerges as distinctive only in a clearly defined position. Much the same process can be seen in words "*lis,ina, vpad,ina*" bringing out the semantic component "*spot*", "*place*"; "*l,d,ina, b,is,er,ina, solom,ina*" bringing out the semantic component "*piece*" etc.

A full morphemic description of Contemporary Standard Russian should include a dictionary or a list of morphemes supplying not only their distinctive semantic components but every possible non-distinctive semantic component as well, together with projection rules which would select the appropriate non-distinctive semantic components emerging in given environments.

3.3 Morphonemic rules

Morphemic description of Contemporary Standard Russian seems to be very well suited for what has generally been known as the morphonemic approach aiming at evolving correct phonemic forms of given morphemes from one basic form. Morphonemic rules for Russian have usually very close relation to diachronic phonemic laws governing historical changes leading to contemporary morphemic alterations. This

approach dealing with morphemic alterations in much the same way as phonetic changes are dealt with in diachronic phonology does not seem to be theoretically justified. To maintain this approach from a purely practical point of view is also rather difficult, because it leaves many alterations outside limited domains usually treated in Russian grammars unaccounted for.

Accentuation, alteration and truncation rules in a suitable reformulation can become part of a broader system of alteration rules based on morphemes as elements and morphemic environment as position [20]. After listing allomorphs of different morphemes selection rules can be formulated in the following form: allomorph A' of morpheme A occurs before morphemes X, Y and Z. After listing all apparent similarities of selections covering a group of morphemes, accentuation, alteration and truncation rules can be formulated together with a clear indication of the domain of their applicability. Cases which are not liable to this kind of generalization remain covered by selection rules without further specification.

REFERENCES

[1] BENDIX, E. H.: Componential analysis of general vocabulary. The semantic structure of a set of verbs in English, Hindi and Japanese. Indiana University, Bloomington. The Hague, The Netherlands, Mouton & Co., pp. 190.

[2] BLOOMFIELD, L.: Language, 1933. New York, Holt Rinehart and Winston 1962.

[3] COLBY, B. N.: Ethnographic semantics: A preliminary survey. Current Anthropology, *7*, 1966, No. 1, pp. 3—32.

[4] DOKULIL, M.: Tvoření slov v češtině. I. Teorie odvozování slov. Praha 1962.

[5] GOODENOUGH, W. H.: Componential analysis and the study of meaning. Language, *32*, 1956, pp. 195—216.

[6] HARRIS, Z. S.: Morpheme alternants in linguistic analysis. Language, *18*, 1942, pp. 169—18.

[7] HJELMSLEV, L.: Prolegomena to a theory of language. The University of Wisconsin Press, Madison 1961.

[8] HOCKETT, Ch. F.: Problems of morphemic analysis. Language, *23*, 1947, pp. 321—343.

[9] HOCKETT, Ch. F.: Two models of grammatical description. Word, *10*, 1954, pp. 210—234.

[10] HORECKÝ, J.: Morfematická štruktúra slovenčiny. Bratislava 1964.

[11] JAKOBSON, R.: Russian conjugation. Word, *4*, 1948, No. 3, pp. 155—167.

[12] LAMB, S.: The semantic approach to structural semantics. Transcultural studies in cognition. Ed. A. K. Romney and R. G. D'Andrade. American Anthropologist, *66* (3), 1964, Part 2.

[13] LOUNSBURY, F. G.: The method of descriptive morphology. Anthropology, *48*, 1953.

[14] MANDELBROT, B.: Structure formelle des textes et communication. Word, *10*, 1954, pp. 1—27.

[15] MORRIS, E. W.: Signs, language and behaviour. New York, George Braziller 1955.

[16] NIDA, E. A.: Analysis of meaning and dictionary making. International Journal of American Lingustics, 24: 236, 1958, pp. 279—292.

[17] ОЖЕГОВ, С. И.: Словарь русского языка. Москва 1960.

[18] OLIVERIUS, Z. F.: K distribuci alomorfů současné ruštiny. Československá rusistika XI, 1966, pp. 207—214.

[19] ОЛИВЕРИУС, З. Ф.: Морфемный анализ современного русского языка. Проблемы современной лингвистики. Монографии, посвященные VI международному съезду славистов (Прага 1968). Philologica 1967, Praha 1967.

[20] ОЛИВЕРИУС, З. Ф.: Основы описания морфемных альтернаций в современном русском языке. Československá rusistika, *15*, 1970, pp. 49—55.

[21] ПАНОВ, М. В. (ред.): Словообразование современного русского литературного языка. Русский язык и советское общество. Социолого-лингвистическое исследование. Москва 1968.

[22] PATRICK, G. Z.: Roots of the Russian language. New York 1938.

[23] ПОТИХА, З. А.: Школьный словообразовательный словарь. Изд. второе. Москва 1964.

[24] RUŽIČKA, J.: Prázdna morféma. Jazykovedný časopis, *15*, 1963, pp. 3—7.

[25] SHEVELOV, G. I.: The structure of the root in modern Russian. Slavic and East-European Journal, XV, 2, 1957, pp. 106—124.

[26] Russian-English dictionary. Ed. A. I. Smirnitsky. Moscow 1962.

[27] STANKIEWICZ, E.: Declension and gradation of Russian contemporary standard Russian. Description and analysis of contemporary standard Russian. Ed. R. Jakobson and C. H. van Schooneveld. The Hague—Paris, Mouton 1968, 173 pp.

[28] TOWNSEND, Ch. E.: Russian word-formation. New York, McGraw-Hill Book Company 1968.

[29] TRUBETZKOY, N. S.: Sur la "morphonologie". TCLP I—Mélanges linguistiques dédiés au premier Congrès des philologues slaves, Prague 1929, pp. 85—88.

[30] TRUBETZKOY, N. S.: Gedanken über Morphonologie. TCLP IV—Réunion phonologique internationale tenue à Prague 18—21/XII 1930, Prague 1931, pp. 160—163.

[31] TRUBETZKOY, N. S.: Das morphonologische System der russischen Sprache. TCLP V. — Description phonologique du russe moderne. Deuxième partie. Prague 1934, p. 94.

[32] WOLKONSKY, C.—POLTORATZKY, M.: Handbook of Russian roots. New York, Columbia University Press 1961, XXVI, 414 pp.

[33] Грамматика русского языка. 1. Фонетика и Морфология. Москва 1952.

[34] Основы построения описательной грамматики современного русского литературного языка. Москва 1966.

Topic-comment in Child Language and in Diachronic Typology

LÁSZLÓ DEZSŐ, BUDAPEST

1. The importance of child language for the study of universal properties of language has been demonstrated by R. Jakobson's brillant studies.[1] The early stages of child language, its subsequent development is of particular interest also for the study of word order in the framework of a topic-comment analysis.[2] In its early stages of one-, two- and three-member sentences, the child language has no — or scarcely developed — morphological structure and only some supra-segmental elements are present. Since the special grammatical means of surface structure are lacking, they do not differentiate the language of Hungarian children from that of Serbian or Swahili children.

The language of a child is closely connected with the speech situation. The analysis of utterances is often possible only if we know this situation and the child's behavior. This circumstance has to be accounted for in topic-comment analysis. In this paper we shall compare the data of topic-comment analysis of child language with those of word order typology. A more detailed treatment of the subject is to be found elswhere.[3] The whole problem needs further investigations; at present our paper may be regarded only as a first approximation to the question.

2. **Topic-comment in child language.** The most important element of the sentence is the comment. It is placed in the focus of attention, all elements of the sentence may be eliminated except the comment. We shall examine sentences containing a grammatical object. These kinds of sentences are very frequent also in child language. In (adult) language proper the object is usually the most relevant element

---

[1] JAKOBSON, R.: Child language, aphasia and phonological universals. The Hague 1968.

[2] For bibliography on word order of early child language see DAN I. SLOBIN, Early grammatical development in several languages with special attention to Soviet research, Working papers, N° 11. Language-Behavior Research Laboratory. Berkeley, Calif. July 1968. For brevity's sake we shall refer only to Slobin's paper for further data.

[3] For details see DEZSŐ L.: A gyermeknyelv mondattani vizsgálatának elméleti-módszertani kérdései (Some theoretical and methodological questions of syntax in child language), Általános nyelvészeti tanulmányok VII, Budapest 1970, 77—99.

of the sentence from the point of view of communication, and the verb is of secondary importance, it is a transitory element in word order typology. The two main types of word order are to be distinguished according to the place of the object: main type *A* has the basic order *S* # *O* (i.e. subject # object), main type *B* has *O* # *S*. The types of main type *A* have the verb in different position: *SOV*, *SVO*, *VSO*.[4] The category of definiteness is marked on the object if it is expressed at all (e.g. by marked object in Altaic, by the so-called objective conjugation in Ugric languages etc.).[5]

In child language the object is also of primary importance, but when speaking of early child language we had better to use " object" in more general sense, as it is used by Ch. J. Fillmore,[6] and to bear in mind the difference between the object in child language and the different kinds of object in typology. For our present analysis this difference is of secondary importance, when it becomes relevant we shall refer to it. In child language the object is usually "given" in the speech situation, it would be definite in adult language if this category is expressed; e.g. Hung. *Adide labda* (2; 0, 02) 'Give the ball'.[7]

In early child language the surface structure restrictions on word order are not relevant as well as the morphological means of surface structure are lacking. An English child may use also word order (*S*)*OV*: *Coat button* (i.e. 'Button coat') and *VOS*: *Watch, bake cake mama*,[8] but the dominant order will be (*S*)*VO* as in adult speech. Since the child language is closely bound to situation, a Hungarian child

[4] Division of word order types into main types *A* and *B* is motivated by topic-comment. J. H. GREENBERG examined only main type *A*, main type *B* (*VOS*) is only mentioned by him (Some universals of grammar with particular reference to the order of meaningful elements. Universals of language. Cambridge, Mass. 1963, pp. 58—90). For psycholinguistic interpretation of Greenberg's typology see MILLER, G. A. and McNEILL, D: Psycholinguistics. Handbook of psychology[2]. (Ed. by E. G. Lindsey and E. Aronson), Reading, Mass. 1969, pp. 743—747. The main type *B* (*VOS*) is described in L. DEZSŐ (1967), Typological questions of the Swahili word order (to appear in Proceedings of 2nd International Congress of Africanists. Dakar 1967). The topic-comment analysis of different types of main type *A* is to be found in L. DEZSŐ, GY. SZÉPE, Adalékok a topic-comment problémához (Notes on the problem of topic and comment), Nyelvtudományi Közlemények, *69*, 1967, pp. 365—88.

[5] MORAVCSIK, E.: Determination. Working papers on language universals 1. Stanford, Calif. 1969, pp. 64—98. For data on Uralic and Altaic see L. BESE, L. DEZSŐ, J. GULYA: On syntactic typology of Uralic and Altaic languages. Theoretical problems of typology and the Northern Eurasian languages (Ed. L. Dezső and P. Hajdú), Budapest 1970, pp. 113—128.

[6] FILLMORE, Ch. J.: Case for case. Universals in linguistic theory (Ed. by E. Bach, R. T. Harms), New York 1968, pp. 1—90. Fillmore's case theory is considered relevant also by DAN I. SLOBIN (Universals of grammatical development in children, Working paper No. 22. Language-Behavior Research Laboratory. Berkeley, Calif. 1969 September, 9).

[7] DEZSŐ, L.: (1970).

[8] LEOPOLD, W. F.: Speech development of a bilingual child. A linguistic record. Vol. 2. Grammar and general problems in the first two years. Evanston, Ill. 1949, pp. 70—84.

will use mainly word order *SVO* that is used in adult speech with "individual" usually "definite" object: e.g. *A fiú írja a levelet* 'The boy is writing *the* letter'. Word order *SOV* is used if the object is not "individual", not concrete: *A fiú levelet ír* 'The boy is letter-writing', and this sort of sentence supposes some kind of unity between verb and object, and it is less familiar to a child. In Russian where this distinction is not relevant variant *SOV* is also frequent in early child language. Both *SVO* and *SOV* are typologically possible non-contextual variants. The emphasis proper to *SOV* in (adult) Russian becomes relevant at a later stage.[9] Contextual variants with *O* in initial position (*OVS*, *OSV*) are impossible both in Russian and Hungarian early child language.[10]

Since child language has close connection with speech situation, it prefers the basic word order: *SOV* in type (1), *SVO* in type (2) and *VSO* in type (3); the secondary word order: *OSV* in type (1), *OVS* in type (2) presupposes some previous context and it is rare, and appears later in child language. Both in child language and in topic-comment typology the basic word order is to be considered as primary because it is based only on speech situation and it is universal. Besides, many languages may have another variant of word order used when the object is mentioned in the previous context. Some languages need also special emphasis for the use of a secondary word order.[11] (The use of remaining variants will not be considered now.) In Hungarian child language the secondary variant is used mainly under emphasis: *Ezt addide* (2; 1,10) 'This give (me)!',[12] *Adide párnát! Párnát adide mátkámasszonynak!* (1;11,15)[13] 'Give (me) the pillow! The pillow give to Márta!'. These examples are recorded from later stages of development, but they are rare even then.[14]

3. Topic-comment analysis in word order typology. — As we have seen in child language, the basic word order is established first, the secondary word order variant supposing context is only a possible result of later development and often supported by emphasis. The surface structure constraints on word order become relevant only parallelly to the development of morphological devices shaping the surface structure.

We have spoken of two main types of word order from the point of view of topic — comment. The main type *A* was described by J. H. Greenberg in his fundamental

[9] For data of *SOV* in Russian see SLOBIN (1968), p. 23.

[10] See SLOBIN (1968), p.22 where he refers to Jakobson.

[11] This is the case in Vakh Ostyak where *OSV* is used in emphasis [for details see J. GULYA'S Vakh Ostyak syntax (in manuscript)] and in Buryat Mongolian where *VSO* is recorded in emphatic sentences: G. A. BERTEGAJEV, C. B. CYDENDAMBAJEV, Grammatika buryatskogo yazyka. Sintaksis. Moskva 1962, p. 100.

[12] See DEZSŐ (1970).

[13] See MEGGYES, K.: Egy kétéves gyermek nyelvi rendszere (The language of a child of two), Budapest 1971, p. 70.

[14] Both Leopold and Stern consider the object as a topic under emphasis in similar examples: SLOBIN (1968), p. 26.

study,[15] the subject is posited before the object in basic order, and the different types are established depending on the place of verb:

(1) *SOV* (2) *SVO* (3) *VSO*.

In languages with fixed word order this is the only possible variant. In languages having limited word order a second variant is admitted, if the object is mentioned in the context:

(1) *OSV* (2) *OVS* (3) *SVO* or *VOS*.

As a third variant *VSO* is to be observed in languages of type (1) and (2). The word order types are connected with types of sentence intonation determined by the place of pause and sentence stress. Main type *B* has word order *VOS*, where the object is placed before the subject. This main type needs detailed description,[16] we restrict ourselves to types (1) and (2) of main type *A*. Further complements may be added to the object, and the sentence is open to the addition of new elements to the left in type (1) and to the right in types (2) and (3):

(1) *S* — — *OV* (2) *SVO* — — (3) *VSO* — —

We shall discuss the details of adding new complements later. So far we have examined pure types, but in many languages the word order type is in process of changing and we may have double or/and mixed type of word order. Both deviation from pure type is to be observed in Hungarian where the change from type *SOV* to type *SVO* takes place. This is a very common phenomenon in different language families, change of type *SVO* to type *VSO* is also to be found.[17]

When looking at the process of word order change from the point of view of topic—comment, we can see that the basic word order changes first and it is followed by the change of the secondary word order. In Hungarian the sentences containing object with article have changed their basic order from *SOV* to *SVO*: *A fiú írja a levelet* 'The boy is writing the letter', but the corresponding secondary word order *OSV* has preserved its function of contextual variant: *Péter írt egy levelet. A levelet a fiú föladta.* 'Peter wrote a letter. The letter was posted by the boy.' The first sentence has word order *SVO*, the second *OSV*. These two variants may be the most frequent ones, and we have word order of mixed type: the basic order is of type (2); the secondary variant is of type (1).[18]

[15] GREENBERG, J.: (1963).

[16] DEZSŐ, L.: (1967).

[17] DEŽE, L.: K voprosu ob istoričesko tipologii slavyanskogo poryadka slov (On the historical typology of Slavonic word order). Studia Slavica, *14*, 1967, pp. 1—87.

[18] This has been recorded in Mamvu language where *OSV* is used in emphasis: A. N. TUCKER, M. A. BRYAN, Linguistic analysis of non-Bantu languages of North-Eastern Africa. London 1966, p. 19.

In neutral order the secondary object, usually the *experiens* of deep structure is placed after the primary object, if the latter one has an article:

Péter írja a levelet Jánosnak (*SVOE*) 'Peter is writing the letter to János',

but the *experiens* may be placed also before the verb if the focus of attention is on the indirect object:

Péter Jánosnak írja a levelet (*SEVO*) 'Péter is writing the letter to János'.

The direct object may be located before the verb only in case of emphasis:

Péter a levelet írja Jánosnak (*SOVE*) 'It is the letter that is written by Péter to János'.

The local modifier may be placed both before and after the verb and other complements:

Péter írja a levelet (*Jánosnak*) *az asztalnál* (*SVO*(*E*)*L*) 'Péter is writing the letter (to János) at the table'

Péter az asztalnál írja a levelet (*Jánosnak*) (*SLVOE*) 'the same'.

The second variant is preferred stylistically: we may better arrange the complements before and after the verb, but we have also typological arguments for it: the direct object has fixed position after the verb, the indirect object also prefers this place, but the local modifier may stand both before and after the verb. The following change of type (1) to type (2) has taken place in Hungarian:

$$SLEOV \rightarrow SLEVO \rightarrow SLVOE \rightarrow SVOEL,$$

but the last step is not finished yet. The direct object has changed its position first, this was followed by indirect object and by local modifier. But the typological change may depend on the features of object: if the object is not concrete, it has no article, the object preserves the place before the verb because of its close contact with the verb:

Péter levelet ír Jánosnak (*SOVE*) 'Péter is letter-writing to János'.

The direct object may preserve its place also in Mande languages where it preceeds the verb, and the other complements follow the verb.[19]

We have to sum up the order of diachronical changes in word order typology. The basic neutral word order changes first, and it is followed by the change of contextual word order. This corresponds to the acquisition of word order in child language where the basic order is established first and the contextual word order is secondary, often connected with emphasis. In word order typology the object stands closer to verb and the other complements are placed further to the left or to the right. But the features of the object and of other complements influence both synchronic order and diachronic change. The place of direct object changes first, it is followed by

[19] TOKARSKAJA, V. P.: Jazyk malinke (mandingo). Moskva 1964, p. 41; CASTHELAIN, R. P. J.: La langue guerzé. Dakar 1952, p. 405.

indirect object and local modifier, but depending on its features the object may be the most conservative as well.

The synchronic order of complements, their diachronic change may be paralleled with the acquisition of the word order of sentences containing object and other complements in child language. We are particularly interested in the process of building up sentences with direct and indirect object and direct object and local modifier.

4. **The acquisition of word order in child language.** — The acquisition of word order of more complex sentences will be examplified in Hungarian. As we have seen, the Hungarian child uses mainly word order of type *SVO*. The transitive verb has an object as its primary complement, placed after the verb. The place of the subject may be both before and after the verb, the latter is non dominant:

SVO: *Málti fölhúzza bécsicipőt* (2; 0, 21) 'Márti is putting on shoes'.[20]

VSO: *Kigombujja | Márti | gombát* (1; 11, 17) 'Márti is unbuttoning (her) button'.[21]

The indirect object, usually an *experiens*, is of later origine and follows the direct object:

VOE: *Adoda teluzát a | Mátikának!* (1; 10, 21) 'Give the pencil to Mártika!'[22]

The local modifier appears as early as the direct object, but it is used with intransitive verbs. Later it may refer also to a transitive verb and then it follows the object:

(*S*)*VOL*: *Elviszem a kabát Másik csobába* (1; 10, 13) '(I) am taking the coat to other room.'[23]

If we have sentences long enough, the local modifier would follow also the *experiens*: (*S*)*VOEL*, as in adult language, but we are more interested in sentences with word order of type *SOV*. Unfortunately they are rare in Hungarian early child language:

(*S*)*OVL*: *Puszit kélsz. Hátul* (2; 1, 3) '(I) want a kiss. Behind.'[24]

The local modifier is placed after the verb. So the variant (*S*)*OVL* is of mixed type. In Hungarian the local modifier may be posited also before the object: *SLOV*, but this has not been recorded from the early child language. The latter is the normal way of expanding sentences in pure *SOV* languages.

According to our data, the Hungarian child expands the sentence from the verb and object onto the right:

$$SVO — —$$

[20] MEGGYES, K.: (1971), p. 79.

[21] MEGGYES, K.: (1971), p. 69.

[22] MEGGYES, K.: (1971), p. 70.

[23] MEGGYES, K.: (1971), p. 74.

[24] MEGGYES, K.: (1971), p. 88. In Russian child language expansion "to the right" is also observed, the "left hand" variant *SOV* is recorded as well [see SLOBIN (1968), p. 24]. In Russian at later stages the use of *SOV* is limited, in Hungarian it will be more frequent because sentences with *SOV* and *SVO* forme an important opposition (see L. DEZSŐ, Einige typologische Besonderheiten der ungarischen Wortfolge. Acta Linguistica Hungarica, *18*, 1968, pp. 356—67).

The linear order of adult language is the order of expansion of sentence in child language as it is the way of the word order change in diachrony. The *SOV* languages has the opposite direction of expanding sentences:

$$S — — OV$$

So the ordering of topic-comment rules into 2 groups: (1) basic, (2) contextually motivated is universal; it regulates also the acquisition of word order in child language and the diachronic change of the types. The linear order of basic rule is relevant both for building up word order in child language and rebuilding it in diachrony, but the diachronic change is more influenced by the features of complements and it is more complex.

Our analysis may be considered only as the first approach to the universal system of word order rules. Evidently it lacks further data both of child language and diachronic typology. We should like to complement our data also with those of aphasy in order to compare the phenomena of word order with the phonology examined by R. Jakobson.[25]

[25] JAKOBSON, R.: op. cit.

List of Participants

Ján Bosák, Jazykovedný ústav Ľud. Štúra SAV, Nálepkova 26, Bratislava
Klára Buzássyová, Jazykovedný ústav Ľud. Štúra SAV, Nálepkova 26, Bratislava
Alexandru Cărăuşu, str. Dr. Babeş, Iaşi 6
Cornel Caşlaru, Institutul politehnic, Bucureşti
Constantin V. Crăciun, Facultatea de Matematica-Mehanica, str. Academiei 14, Bucureşti 22
Karel Čulík, Matematický ústav ČSAV, Žitná 25, Praha 1
László Dezső, MTA Nyelvtudományi Intézete, Szalay utca 10—14, Budapest V
Mihai Dinu, Institutul de studii si cercetări hidrotehnice, Splaiul Independenţei 294, Bucureşti 17
Toma Djukanov, Institut za matematika, bul. A. Ivanov, Sofia 26
Josef Filipec, Ústav pro jazyk český ČSAV, Letenská 4, Praha 1
Michal Frank, Katedra anglistiky Filoz. fakulty UPJŠ, Grešova 3, Prešov
Yves Gentilhomme, Faculté des Lettres et Sciences Humaines, rue Mégavand 47, Besançon
Eva Hajičová, Laboratoř algebraické lingvistiky UK, Malostranské nám. 25, Praha 1
György Hell, MTA Neylvtudományi Intézete, Szalay utca 10—14, Budapest V
Zdeněk Hlavsa, Ústav pro jazyk český ČSAV, Letenská 4, Praha 1
Victoria Hopârteanu, Facultatea de Filologie, Universitatea Babeş-Bolyai, str. Horia 31, Cluj
Ján Horecký, Jazykovedný ústav Ľud. Štúra SAV, Nálepkova 26, Bratislava
Gunnar Jacobsson, Glasmästaregatan 26, Göteborg
László Kalmár, MTA Matematikai, Logikai és Automataelméleti Tanszéki Kutató Csoportja, Somogyi Béla u. 7, Szeged
Sándor Károly, MTA Neylvtudománi Intézete, Szalay utca 10—14, Budapest V
Emese Kis, Facultatea de Filologie, Universitatea Babeş-Bolyai, str. Horia 31, Cluj
Gerda Klimonow, Arbeitstelle für Mathematische und Angewandte Linguistik und

Automatische Übersetzung, Leipzigerstr. 3—4, 108 Berlin
Jan Kořenský, Ústav pro jazyk český ČSAV, Letenská 4, Praha 1
Viktor Krupa, Kabinet orientalistiky SAV, Klemensova 27, Bratislava
Bohumila Kyliánová, Laboratoř počítacích strojů VUT, Obránců míru 21, Brno
Ileana Lascu, Facultatea de Filologie, Universitatea Babeş-Bolyai, str. Horia 31, Cluj
Aleksander Ludskanov, Institut za Matematika, bul. A. Ivanov, Sofia 26
Marie Ludvíková, Ústav pro jazyk český ČSAV, Letenská 4, Praha 1
Vítězslav Maixner, Ústředí vědeckých, technických a ekonomických informací, Konviktská 5, Praha 1
Eleonora T. Marcu, Institutul de Matematica, Calea Griviţei 21, Bucureşti 12
Solomon Marcus, Institutul de Matematica, Calea Griviţei 21, Bucureşti 12
Dan Mârza, Facultatea de Filologie, Universitatea Babeş-Bolyai, str. Horia 31, Cluj
M. Nagao, CETA, Cedex 53, Grenoble
Miroslav Novotný, Janáčkovo nám. 2a, Brno
Zdeněk F. Oliverius, Zahradní město, Práčská 2672, Praha 10
Ján Oravec, Jazykovedný ústav Ľud. Štúra SAV, Nálepkova 26, Bratislava
Gabriele Orman, Institutul Pedagogic, Facultatea de Matematica, str. Karl Marx 50, Braşov
Karel Pala, Katedra českého jazyka UJEP, Arne Nováka 1, Brno
Jarmila Panevová, Laboratoř algebraické lingvistiky UK, Malostranské nám. 25, Praha 1
Ilpo Piirainen, Kirkkosalmentie 5 B 16, Helsinki 84
Walter Priess, Arbeitstelle für Mathematische und Angewandte Linguistik und Automatische Übersetzung, Leipzigerstr. 3—4, 108 Berlin
Jan Průcha, Pedagogický ústav J. A. Komenského ČSAV, Mikulandská 5, Praha 1
Jozef Ružička, Jazykovedný ústav Ľud. Štúra, Nálepkova 26, Bratislava
Rudolf Růžička, Institut für Sprachwissenschaft der Karl-Marx-Universität, 705 Leipzig
Otto Sechser, Ústav vědeckých, technických a ekonomických informací, Konviktská 5, Praha 1
Liana Schwartz, Institutul de Matematica, Calea Griviţei 21, Bucureşti 12
Petr Sgall, Laboratoř algebraické lingvistiky UK, Malostranské nám. 25, Praha 1
Viliam Schwanzer, Katedra germánskej filológie UK, Gondova 2, Bratislava
Eleonora Slavíčková, Na Pankráci 55, Praha 4
György Szépe, MTA Nyelvtudománi Intézete, Szalay utca 10—14, Budapest V
Marie Těšitelová, Ústav pro jazyk český ČSAV, Letenská 4, Praha 1
Oldřich Uličný, Katedra českého jazyka a literatury Ped. fakulty, Hradec Králové
Bernard Vauqois, Centre d'études pour la traduction automatique, 38 St Martin-d'Hères

Recueil linguistique de Bratislava IV

Prebal a väzbu navrhol *Rastislav Majdlen*
Redaktorky publikácie *Klára Moravcová* a *Eva Zikmundová*
Technický redaktor *Jozef Szabó*

Prvé vydanie. Vydalo Vydavateľstvo Slovenskej akadémie vied v Bratislave r. 1973 ako svoju 1681. publikáciu. Strán 292. Vytlačil Tisk, knižní výroba, n. p., Brno, závod 1.
AH 20,09 (text 18,39, ilustr. 1,70), VH 20,68. Náklad 400 výtlačkov.
Číslo povolenia 1438/I-OR-1972

71 — 070 — 73
12/01 509/58

Kčs 40,— I